CENTER FOR EVALUATION AND DEVELOPMENT (C4ED)

ENDLINE EVALUATION OF ASSESSMENT FOR LEARNING PROGRAMME

FINAL REPORT

FOR

UNICEF ETHIOPIA

Center for Evaluation and Development

AUGUST 20

ACRONYMS AND ABBREVIATIONS

AfL	Assessment for Learning
C4ED	Center for Evaluation and Development
CCA	Continuous Classroom Assessment
CfBT	Centre for British Teachers
CPD	Continuous Professional Development
CRC	Convention on the Rights of the Child
CTE	College of Teachers Education
EGRA	Early Grade Reading Assessment
EQ	Evaluation Question
ESDP	Education Sector Development Programme
ESSSWA	Ethiopian Society of Sociologists, Social Workers and Anthropologists
FGD	Focus Group Discussion
FTT	Fully Trained Teacher
GDP	Gross Domestic Product
GEQIP-E	General Education Quality Improvement Program for Equity
IDI	In-depth interviews
KII	Key Informant Interviews
MLC	Minimum Learning Competency
MoE	Ministry of Education
NEAEA	National Educational Assessment and Examinations Agency
NER	Net Enrolment Rate
NLA	National Learning Assessment
ODI	Overseas Development Institute
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PPP	Purchasing Power Parity
RCT	Randomised Controlled Trial
REB	Regional Educational Bureau
SEQ	Sub-evaluation question
SD	Standard Deviation
SNNPR	Southern Nations, Nationalities, and People's Region
ToC	Theory of Change
ToT	Training of Trainers
TTS	Teachers Trained at School
UIS	UNESCO Institute for Statistics
UNESCO	United Nations Educational, Scientific and Cultural Organisation
UNICEF	United Nations Children's Fund
USD	United States Dollar
WEO	Woreda Education Office

GLOSSARY OF TERMS

Term	Description
Cluster School	School directly targeted by the AfL intervention. At Cluster Schools one teacher per subject (Mother Tongue, English, Mathematics, Environmental Science) should obtain in-depth training in AfL techniques (see Training of Trainers).
AfL fully trained teacher (FTT)	Trained teachers from cluster schools. They are supposed to pass training onto teachers within their school and to teachers in nearby schools (satellite schools) in cooperation with cluster supervisors. At each cluster school there should be up to four AfL fully trained teachers.
Satellite School	School indirectly targeted by the intervention. At satellite schools, all teachers in the relevant subjects receive training by cluster supervisors and AfL fully trained teachers.
AfL teachers trained at school (TTS)	All teachers who obtained training by AfL fully trained teachers and/or by cluster supervisors.
Cluster supervisor	An education official who obtained the same in-depth training as AfL fully trained teachers. Cluster supervisors are responsible for supervising the implementation of AfL techniques in cluster and satellite schools, to support trained teachers and to train teachers in satellite schools together with AfL fully trained teachers.
Training-of-Trainers (ToT)	The training of selected teachers from cluster schools. Teachers who have attended the ToT are referred to as AfL fully trained teachers.
ToT trainers	Qualified staff – typically from CTEs – who conduct the ToT.
CTE graduate	Teachers who obtained AfL training as part of their education at Colleges of Teacher Education (CTE).
School Training	Training of teachers working at cluster school who did not attend the ToT
Cluster training	Training of teachers working at satellite schools and trained at cluster school
Pre-deployment training	Training provided to graduates of CTE prior to deployment within schools

CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

This is the endline evaluation of the 'Assessment for Learning' (AfL) programme in Ethiopia. The evaluation was conducted between September 2019 and July 2020.

Background

The AfL programme aims to develop the capacity of teachers for implementing continuous assessment in their classrooms and has the overall goal of improving the quality of first cycle primary education (Grades 1-4) and student learning outcomes. For this purpose, the programme provides teachers with training and materials on various techniques that can help them implement continuous assessment in their own classroom and improve their teaching practices by utilising the information gained from regular assessment. One key specific aim is to align continuous assessment practices with Minimum Learning Competencies (MLCs) for subjects and implement learner-focussed pedagogy with assessments that not only measure achievement but will also be used as an input for instructional planning and student-tailored instruction. The AfL programme focusses on the subjects of Literacy (Mother Tongue and English), Mathematics and Environmental Science. The inputs for the programme initially included on two areas; firstly, developing a comprehensive in-service teacher training programme (e.g. guidelines, materials), and secondly, training-of-trainers to prepare a core group of trained local personnel from CTEs. During the growth of the programme, the inputs now include the integration of the AfL programme into the curriculum of CTEs.

The programme began in 2013 and was piloted in three regions (Addis Ababa, Harari and Amhara) and has since been expanded to of Oromia, Tigray and Somali (2016) and Afar, SNNPR, Gambella and Benishangul-Gumuz (2017). Since 2018, the AfL programme has been integrated into pre-service training at CTEs in Oromia and is in the process of integration in further regions. Key stakeholders of the programme include high-level policy makers in UNICEF, MoE, Regional Education Bureaus (REBs) and actors at the local level including Woreda Education Offices, Cluster Supervisors, School Directors, teachers, parents and students.

Purpose and Objectives of the Evaluation

In September 2019, the Center for Evaluation and Development (C4ED) was contracted by UNICEF Ethiopia country office as an independent evaluator to evaluate the AfL programme in Ethiopia. In particular, the evaluation assessed whether the Assessment for Learning (AfL) programme designed by UNICEF in collaboration with the MoE, REBs had the desired outcomes of improving the quality of education. In light of the ongoing overall national education goals in Ethiopia detailed in the ESDP-IV and accompanying the Education Quality Improvement Package II (2012-2016) the AFL has been designed and further developed by different stakeholders.

The evaluation will address the main objectives, which are to:

- Assess the impact of the AfL programme on teacher practice, learner participation and on assessment of learning and to
- Evaluate the extent to which the AfL intervention has improved teacher practice and learning outcomes, and
- Assess how key stakeholders (students, teachers/school administrators, parents, communities) view the relevance and effectiveness of the AfL training, AfL materials/package.

Evaluation Methodology

Based on the discussion with the UNICEF, MoE, the REBs & ELIXIR, the finalised evaluation strategy was based on a mixed-methods approach involving both, qualitative and quantitative data collection. For the quantitative approach, we conducted an assessment in Literacy (English and Mother Tongue), Mathematics and Environmental Sciences for students in the third and fourth grade in the Oromia region. In addition, we have administered questionnaires for teachers and school principals, as well as classroom observations of teaching practices and student / teacher behaviour in the classroom.

For our qualitative approach, we have conducted In-depth Interviews (IDI) with teachers, trainers and school principals and Key Informant Interviews (KII) with cluster supervisors, CTE instructors, staff from UNICEF and the MoE and its regional dependencies. Furthermore, we have hosted Focus Group Discussions (FGD) with students' parents in Oromia, Tigray and Benishangul-Gumuz.

The focus of the quantitative analysis was to measure the impact of the AfL intervention on (i) teacher practice and (ii) on student outcomes. For any causal interpretation on the impact of a programme on its desired outcomes, a good impact evaluation essentially relies on the quality of the chosen counterfactual situation. The counterfactual situation corresponds to what would have happened if the AfL Programme had not been implemented. Given that the AfL intervention of schools & teachers were not assigned randomly, a randomised controlled trial was unfortunately not feasible in this case. Hence, the impact evaluation was designed based on quasi-experimental methods, namely on matching methods. For this, we selected a group of control schools, not affected by the AfL programme. The impact was then assessed by comparing the performance of the pupils and teachers in AfL schools to the ones in non-AfL schools. With this design we were be able to quantify the AfL programme effect relative to schools that received no treatment and could observe if the cascading model was working within clusters.

In the main analysis, we used multivariate regression models to control for heterogeneous student backgrounds, such as different school, teacher and zone characteristics. Throughout the evaluation process, we made significant efforts to generate high-quality data and credible evidence considering the time, budget and the prevailing circumstance in the field. Yet, since the data collection was only carried out in the Oromia region, in addition to the qualitative approach in the Tigray and Benishangul-Gumuz regions, by design, the findings of this study cannot be generalized to the whole country. However, many lessons learnt from these three regions might still be relevant for other Regional Education Bureaus (REBs).

Main findings and conclusion

Effectiveness

The evaluation showed some promising evidence that the AfL programme led to an improvement in the level of teachers' knowledge on various areas of continuous assessment. As per the programme theory of change, having the required knowledge on the components of continuous assessment and understanding best practice should be considered as a prerequisite for a teacher to be able to effectively implement the techniques in their classroom. Sampled teachers were asked a set of questions that related to the various components of continuous assessment including question development and the provision of feedback to students. An aggregated composite score of correct answers was used as an indicator for knowledge of continuous assessment. Using the multivariate regression models controlling for teacher and school characteristics, the AfL programme improved teachers' knowledge scores in various knowledge tests relating to continuous assessment techniques

by 0.36 standard deviations compared with teachers in schools not connected with the AfL programme at all.

Moving on from teacher knowledge and actual implementation of continuous assessment within teachers' classrooms, there is also some evidence that AfL was effective in encouraging teachers to adapt their teaching to continuous assessment. Teachers in AfL cluster schools were observed to have increase, on average, of 19% in the time they spent in class actively assessing their students (as opposed to more traditional lecturing activities or time spent on classroom management or time spent off-task not related to learning) compared with teachers in non-AfL schools. The results of the multivariate regression models show that teachers in schools that received AfL also spent 30% more of their time interacting directly in class with students when teaching compared to teachers in control schools. This is supplemented by the qualitative study that found teachers were able to shift to a more student-centred learning style after receiving training on AfL. There was, however, no effect on the level of overall student engagement within the classrooms. This indicates that although teachers adapt their teaching the environment to effectively assess students continuously may still be hampered by student discipline or their ability to engage all the students within their classroom.

Sampled teachers were asked to report their use of continuous assessment in the classroom, through the rate of assessments used and the range of assessment methods, such as role-play, group work or homework. When comparing AfL cluster schools' teachers with control schools, there was no difference found. Similarly, no effect was found on the reported structuring of the practice of continuous assessment – through lesson plans and student progress records by teachers. However, during classroom observations, Teachers in AfL schools demonstrated greater structuring of continuous assessment in their classrooms, through the quality of lesson planning, learning introductions directly linked to MLCs and coaching if individual students. The AfL programme at a school was estimated to increase the composite score for observed structuring of continuous assessment, computed through principal component analysis, by 0.86 standard deviations compared with control schools. During the qualitative study, teachers regularly hold up the AfL programme as having provided them with the training to prepare develop their own questions linked to MLCs which would not happen prior to the training.

Though not explicitly part of continuous assessment, parental engagement in their child's education was a key component that was identified as part of the AfL programme. Parental engagement also overlaps substantially with continuous assessment as it requires regular assessment of a student's learning and development of feedback to involve parents throughout a student's education. There was some evidence that AfL had a positive impact on improving the regularity which there was direct communication and physical meetings between teachers and parents. The estimated effect of the AfL programme on the composite score for parental engagement was an increase of 1.4 standard deviations.

Finally, through testing of 1,161 children randomly selected across sampled schools in the core subjects relating to AfL (Mother tongue, English, Mathematics and Environmental Science) there was some evidence that AfL led to improved test scores in two out of the four focal subjects. Through the multivariate regression model, controlling for student and school characteristics, there was an estimated positive impact on mother tongue (Afan Oromo in the case of this study) and Mathematics of 0.3 standard deviations each. There was no observed effect for English or Environmental Science, however. From semi-structured interviews with teachers and school principals, on the whole provided support for the findings above and attributed AfL with positive changes in their students'

outcomes when compared to previous cohorts or to colleagues that rarely used continuous assessment.

There were no positive effects of the AfL programme found on satellite schools, that is the schools within an education cluster where the AfL programme was implemented, in any of the outcome indicators. Given the findings of the quantitative and qualitative study into the success of the cascading knowledge model, a lack of effect is unsurprising given that in the majority of cases, these schools and their teachers had not received any formal training in AfL through cluster trainings. This means that the only spillovers potentially to the satellite schools would have come through informal knowledge sharing or teacher turnover and are likely to be minimal.

Whilst the results in the effect on teacher practice and student learning outcomes are promising, they should be interpreted with caution due to the limitations of the study design, particularly that we are unable to account for unobservable differences between treatment and control schools due to a lack of both randomisation and baseline data. In addition, as Oromia was selected for the quantitative study as an exemplar of the rollout of the AfL programme, the findings may not extrapolate these results to all the regions.

Relevance, Efficiency and Sustainability

A key input of the AfL programme was the development of an AfL package that included teaching reference materials for each of the core subjects and field tested assessment tools. The study found that the programme was successful in developing materials that were specific to the various local contexts across Ethiopia including translation into local languages. Stakeholders reported high levels of satisfaction with the materials and believed them to be relevant to their day-to-day education experiences. One concern that was raised regularly related to the availability of these materials and felt that this constrained their schools receiving the full benefit of the programme as they struggled for reference materials once returning the schools.

The AfL package supplemented the cascading knowledge model for disseminating the training and knowledge. The initial stage of the cascading knowledge model used trainers from College of Teachers Education (CTE) institutions leading training programmes, referred to as Training of Trainers (ToT) trainings, for selected teachers and school directors from cluster schools alongside local education officials such as cluster supervisor. The study found that participants in the training had high levels of satisfaction in the quality of the training they received in various areas relating to continuous assessment and the abilities of the trainers. There was, however, a feeling that the length of the trainings was too short (two-thirds of participants felt the length of the training was insufficient) and an issue with adherence to the planned model for attendance. Most participants in ToT training sessions (86%) reported that they had not attended all three trainings required for graduation, reasons for this include staff turnover and schools rotating which teachers attended. This may limit their ability to fully use continuous assessment and cascade their knowledge to other colleagues.

Upon completing the training course, the training was aimed to be cascaded down both within the cluster school (to teachers that were not selected to attend the initial training) and to staff from other schools within the cluster. Both the quantitative and qualitative data suggest that there were significant challenges to the training cascading to other schools within the cluster, with the main issue identified being financial support required to cover the costs of the training. As the cascading knowledge model appears to have significant bottlenecks in its current form, there is a promising

avenue for scale up through integration of AfL modules into pre-service training in CTE programmes. This is currently officially implemented in Oromia and due to start in Tigray and presents a strong learning opportunity for other REBs wishing to pursue this approach. The pre-service element also avoids issues that affected the cascading knowledge model such as teacher turnover and the cost of cluster trainings.

There also appears to be some contextual challenges to teachers using AfL techniques and applying continuous assessment within the classroom. These issues include large class sizes that affect a teachers' ability to satisfactorily assess individual students' learning progress and employ interactive assessment and teaching methods throughout the class. The lack of funding and resources available at schools, even basic materials such as paper and stationery, cause challenges to teachers both within the classroom and in keeping effective student progress records. Other highlighted issues include student absenteeism and time constraints due to workload which constrain the quality of preparation and recording of assessments in class.

Finally, in terms of influencing education policy and practice in Ethiopia, there is clear success in obtaining buy-in for the AfL programme at all levels of policy making. This can be seen in the various avenues of scaling up for the AfL programme throughout the country. As of 2020, four REBs have, or are in the process of, integrating the AfL programme into their teacher pre-service programmes. In addition, the AfL programme has spread into the General Education Quality Improvement Program for Equity (GEQIP-E) programme's continuous classroom assessment module, which covers half of all schools in Ethiopia. The MoE have also integrated the AfL into their continuous professional development programme available for all teachers in Ethiopia. Stakeholders, from REB officials to individual school teachers feel positively about the programme and feel it improves the level of education quality in Ethiopia, despite some noted challenges.

Implications

The study provides several implications for users of the evaluation. These are addressed to the Ministry of Education, UNICEF Ethiopia Country Office and stakeholders at national and local levels, to engage all major stakeholders in an ongoing effort to improve AfL and improve the level of education quality in Ethiopia:

Firstly, while the AfL programme in its current form has been shown to increase teacher knowledge in continuous assessment, the length of the training has been criticised as insufficient by many ToT participants. Hence, it should be best practice to make sure all teachers receive the full intended amount of training. This finding also extends to potential scale-ups and integration into CTE courses to ensure that sufficient time is provided to fully learn the techniques and the components of continuous assessment included in AfL.

Secondly, if the cascading knowledge model is to be continued in certain regions, substantial improvements need to be made in the organisation and monitoring of the cluster and school training. Further budget to cover basic costs and incentives should be considered to ensure that the training is implemented as intended (especially with regards to training length and provision of materials) and that teachers are motivated to attend. Experience sharing workshops at CTE, regional or woreda level may be a cost-effective way of improving local implementation and maintaining or increasing the engagement of school staff.

Thirdly, Oromia has included the AfL programme in its CTE pre-service training. It is recommended that other regions consider doing the same. The pre-service training pathway to scale-up appears to be a cost-effective way of extending the reach of the programme to all newly trained teachers.

Finally, efforts should be dedicated to strengthening the supervision and monitoring of the AfL programme at the cluster level. This includes conducting frequent supervisory support, and capacity strengthening for cluster supervisors and woreda education officers.

CHAPTER 1: INTRODUCTION

In 2019, the Center for Evaluation and Development (C4ED) was contracted by UNICEF Ethiopia to conduct an endline evaluation of the 'Assessment for Learning' (AfL) programme in Ethiopia. This endline report presents our findings from the evaluation. In this report, we first provide an overview of the AfL programme in the rest of Chapter 1, before describing the details of evaluation and the methods in Chapters 2 and 3, respectively. Our findings are presented in Chapter 4. Chapter 5 closes with our conclusions and recommendations.

1.1 BACKGROUND AND CONTEXT

Ethiopia is the second most populated country in Africa after Nigeria, with a population of 105 million people (UNICEF, 2019) made up of over 90 ethnic and linguistic groups (Ministry of Education [MoE], 2015). Growing at an annual rate of 2.6% the population is young, with 41% aged below 15 (UNESCO Institute for Statistics [UIS], 2019). Providing education to the population remains a major challenge in Ethiopia given that around 80% of the population live in remote and rural areas, making Ethiopia one of the least urbanised countries in the world (MoE, 2015). Ethiopia is a landlocked country with an area of approximately 1.14 million square kilometres and shares borders with the Sudan, South Sudan, Somalia, Djibouti, Eritrea and Kenya. The country's administration is sub-divided into 10 regional and ethno-language states: Tigray, Afar, Amhara, Oromia, Somali, Benishangul-Gumuz, Southern Nations, Nationalities, and People's Region (SNNPR), Gambella and Harari, Sidama and two administrative councils: Addis Ababa and Dire Dawa. The regions are further divided into zones, woredas, and kebeles, the latter of which represent the smallest administration unit.

Over the past two decades, Ethiopia has seen encouraging improvements in the economic realm. From the financial year 2008 until the financial year 2018 the country's gross domestic product (GDP) grew at an annual average rate of 9.6% (World Bank, 2019). Per capita GDP reached USD 768 in 2017 (UIS, 2019); in terms of purchasing power parity (PPP) this amounts to USD 1,903. The most important economic sector in the country is services, accounting for 39% of GDP, closely followed by the agricultural sector which accounts for 35% of GDP as of 2018 (World Bank, 2019). However, more than three-quarters of the economically active population remain employed in agriculture (Central Statistical Agency of Ethiopia [CSA], 2016). The share of people living below the poverty line has decreased dramatically from 44% in 2000 (MoE, 2015) to 26% in 2016 (World Bank, 2019). Given the income level in Ethiopia, poverty is relatively low, however, as the World Bank (2019) points out, transmission between economic growth and poverty reduction is also rather low. The vision of the country is to become a middle-income country by 2025 (Moller, 2015).

EDUCATION IN ETHIOPIA

Over the past two decades there has been remarkable progress in education with a dramatic increase in access to education. In particular, access to primary education has increased substantially. The net enrolment rate (NER) – the proportion of enrolled children to all children in an age group that ought to be enrolled – in primary schools increased from 30% in 1995/96 (MoE, 1998) to 85.8% in 2013/14 (MoE, 2015). This increase followed the abolition of school fees in 1995/96 (Overseas Development Institute [ODI], 2011) and a large school-building campaign that tripled the number of primary schools over the past 20 years (MoE, 2015). However, as the government predicted in its first Education Sector Development Plan (ESDP-I) in 1997, a quick increase in access to education inevitably came at the cost of quality (Orkin, 2013). In fact, the National Learning Assessment (NLA)

2007 showed that the massive increase in enrolment had led to a decline in student scores (Overseas Development Institute [ODI], 2011). In 2011, an Early Grade Reading Assessment (EGRA) in Grades 2 and 3 found that 34% of students enrolled in Grade 2 were unable to read a single word of a grade-relevant story and 48% of second grade students failed to answer any comprehension questions on a reading comprehension test. Furthermore, only 5% of enrolled second-grade students reached a 60 word per minute reading fluency, which was the expected standard for that age group in 2010. While promoting access to primary education for girls – a major focus of the government – helped narrow the gender gap in education (ODI, 2011), it remains a major challenge for Ethiopia to further narrow the gap, especially in rural areas. In the EGRA girls generally outperformed boys on early reading tasks. However, this pattern was exclusive to urban areas; in rural areas boys significantly outperformed girls in all reading tasks (United States Agency for International Development [USAID], 2011). The latest numbers show promising progress with respect to the gender gap in literacy. While the literacy gender gap in the adult population is substantial (15 percentage points), it is only 2 percentage points among young adults aged 15 to 24 years. Overall, 44% of women and 59% of men in Ethiopia can read and write, while among young adults 72% of women and 74% of men are literate (UIS, 2019). Further challenges in the education sector are high dropout rates at about 25% in the first grade and 10% in subsequent grades as well as the low share of pupils (62.1%) completing Grade 8 (MoE, 2019). In general, the overall low literacy rate of only 52% in the adult population presents a barrier to achieving development goals in Ethiopia. Specifically, it presents an obstacle to the vision of becoming a middle-income economy by 2025 (MoE, 2015).

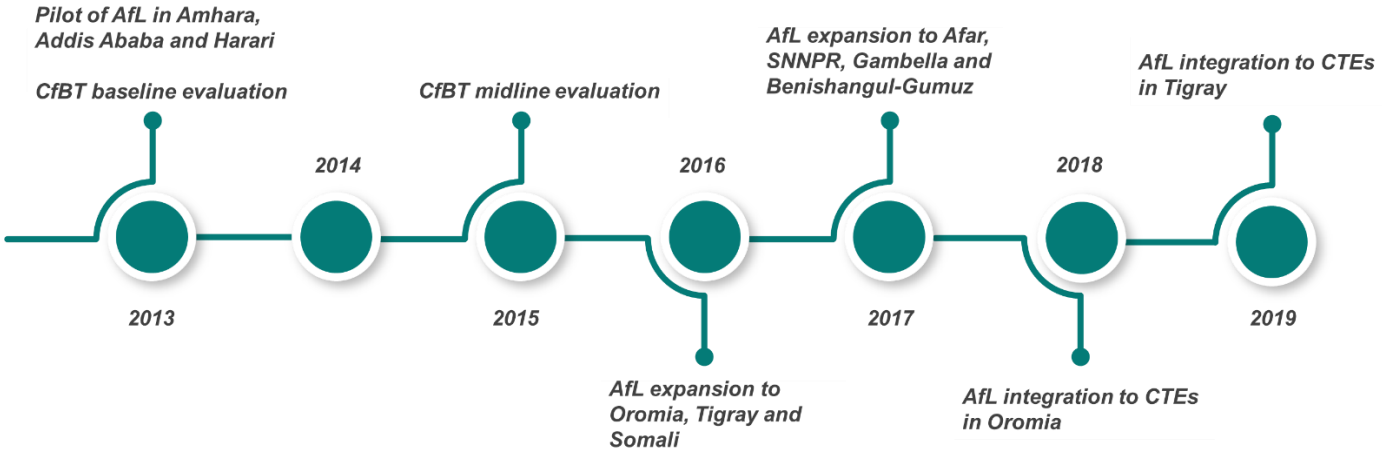
BACKGROUND OF THE PROGRAMME

Aiming to grant every child and adolescent the fundamental human right of quality education (UNESCO, 2019) and in response to the findings in the NLA 2007 and the results of the EGRA in 2011, the MoE has put substantial effort into increasing the quality of education (MoE, 2015). Supporting the overall national education goals in Ethiopia detailed in the ESDP-IV and accompanying the Education Quality Improvement Package II (2012-2016), UNICEF designed the Assessment for Learning (AfL) programme in collaboration with the MoE, the Regional Education Bureaus (REBs) and with initial consultancy support from the American Institute for Research in 2012. From 2015 onwards, the programme was significantly further developed and implemented by ELIXIR Research and Consultancy, a consultant firm based in Addis Ababa.

The AfL programme aims to develop the capacity of teachers for implementing continuous assessment in their classrooms and has the overall goal of improving the quality of first cycle primary education (Grades 1-4) and student learning outcomes. For this purpose, the programme provides teachers with training and materials on various techniques that can help them implement continuous assessment in their own classroom and improve their teaching practices by utilising the information gained from regular assessment. One key specific aim is to align continuous assessment practices with Minimum Learning Competencies (MLCs) for subjects and implement learner-focussed pedagogy with assessments that not only measure achievement but will also be used as an input for instructional planning and student-tailored instruction. The AfL programme focusses on the subjects of Literacy (Mother Tongue and English), Mathematics and Environmental Science. It focusses on two different areas; firstly, developing a comprehensive in-service teacher training programme (e.g. guidelines, materials), and secondly, training-of-trainers to prepare a core group of trained local personnel.

In 2013, the AfL programme was piloted in the regions of Addis Ababa, Harari and Amhara, with in-service teachers at selected primary schools. It was extended to the regions of Oromia, Tigray and Somali in 2016 and recently to Afar, SNNPR, Gambella and Benishangul-Gumuz. In order to reach a greater number of schools and teachers and to make sure all future teachers know how to apply AfL in theory and practice, UNICEF has been collaborating with the government to integrate AfL in the College of Teachers Education (CTE) curricula. Oromia was the first region to integrate the AfL into CTE curricula in 2018. From 2019, the integration of AfL into the curriculum for Oromia was officially implemented. CTEs in Afar and Tigray are to follow in 2020. Figure 1 below presents an overview of the programme timeline.

Figure 1. Overview of AfL timeline



SIGNIFICANCE OF THE PROGRAMME

The following description presents the significance of the programme for children, Ministry of Education Ethiopia and UNICEF:

Programme's significance for children

The AfL programme aims to increase the capacity of education professionals and teachers to implement continuous assessment within their classrooms and ultimately increase the quality of primary education in Ethiopia. This is significant for children in Ethiopia when viewed through both a human-rights based approach and human capital lens.

Education has been formally recognized as a human right since the adoption of the Universal Declaration of Human Rights in 1948. This has been affirmed by numerous treaties since including the Convention on the Rights of the Child (CRC) which broadened the concept to include that education should be empowering for children and promote development. Beyond, international treaties, a number of education conferences, such as the World Education Forum in 2000, committed to the improvement of education commits nations to the provision of primary education of good quality and to improving all aspects of educational quality and adopted the Dakar Framework for Action.

In addition to a human-rights based approach to education, access to high quality schooling and educational attainment is important for children as it's a powerful tool in supporting child development, well-being and breaking the cycle of poverty. One such benefit is that receiving a quality education can lead to accumulation of human capital and productive skills that will lead to improved labour market outcomes and increased wages. Improved levels of education can also lead to improved health outcomes through greater capacity to follow healthier behaviours with indicators such as life expectancy and child mortality rates improving with education level. A high-quality education is also important for girls as it is an important tool in engendering female empowerment. For example, better educated women tend to be healthier, have greater participation in the formal labour market, have higher incomes, marry later and have fewer children. Education also has benefits for for future generations of girls as households with higher levels of education are more likely to value girls' education and send girls to school.

Programme's significance for MoE

To overcome the challenges faced in the Ethiopian education system, the MoE in their latest programme for the development of the education sector (ESDP-V) focused on the quality of education with a specific emphasis on the need for increased need for continuous assessment. It also includes the implementation of the GEQIP programme that was introduced by the Government of Ethiopia. The AfL programme is linked with both of these priorities. Firstly, as it represents an avenue for providing teachers with the skills to conduct continuous assessment and translate the policy into action in classrooms and UNICEF have been working with the MoE since 2012 in implementing the programme. Secondly, as an effort to scale up the AfL programme, UNICEF and ELIXIR have been working closely together with the World Bank and the Ministry of Education to integrate a module on Continuous Classroom Assessment (CCA) in the GEQIP-E programme.

Programme's significance of UNICEF

The programme holds high significance for UNICEF Ethiopia and is considered a flagship initiative within the country and contributes to the aim of ensuring equitable and improved delivery of quality primary education. Through this programme, UNICEF emphasized to improve teaching practice through training programmes on the utilization of continuous assessment practices that support them in responding to the individual learning needs and progress of their students. In addition, UNICEF prioritization the provision of educational services for displaced (refugee and internally displaced) children and young people is supported through the expansion of the AfL programme into Benishangul-Gumuz Region which included training primary teachers from refugee camps. These priorities were in line with UNICEF global mandate and programming priorities for realization of child rights and child well-being.

1.2 OVERVIEW OF CONTINUOUS ASSESSMENT

CONTINUOUS ASSESSMENT

As described in the previous section, the AfL programme aimed to improve the quality of primary education and student learning outcomes, through developing the capacity of teachers for implementing continuous assessment. This section provides a brief literature review on continuous assessment.

As part of the World Education Forum that took place in May 2015, education leaders throughout the world agreed on the Education 2030 Incheon Declaration that outlined a set of objectives and approaches for 'inclusive and equitable quality education and lifelong learning for all' (United Nations Educational, Scientific and Cultural Organisation [UNESCO], 2015). As part of the declaration, participants identified the importance of measurement of progress at all levels. There is a clear trend of countries across the world increasing the number of large-scale national learning assessments (Benavot & Tanner, 2008) to measure and appraise the efficacy of education systems at a national level and make comparisons at an international level. In addition to this, education systems will often mandate yearly summative examinations of all students to inform and aid decision-making at all levels, including at the level of schools, individual students and their parents. Whilst such assessments have the benefit of incentivising study and focusing teacher efforts on key areas of the curriculum (Heyneman, 1987; Hill, 2013), the use of such single, high-stakes assessments, particularly at crucial points of transition for students in their academic life-cycle, are often criticised as open to malpractice (Kellaghan & Greaney, 2019), narrowing the scope of learning in a classroom (Dundar et al., 2014; Kellaghan & Greaney, 1992) and of the assessments themselves being of poor quality (Burdett, 2017). In addition, for students themselves, education systems focusing on high stakes summative assessments can be detrimental as students do not receive information on their own learning and do not receive signals on how they perform on these examinations (Kapambwe, 2010) and subsequent poor performance can lead to reduced self-esteem (Harlen & Deakin, 2002).

The other category of using assessments to measure student learning and influence decision-making can be described as classroom-based, continuous assessment. Continuous assessment can take a summative form, meaning assessment conducted at the end of a learning block to understand the accumulated knowledge of a student. Continuous assessment may also be formative, which is when students are assessed throughout a learning block. Formative assessment is designed to measure the achievement of specific learning aims and the results can be used to inform and, if required, change action within a classroom.

There is a significant challenge in describing continuous assessment as a single defined concept and definitions often vary across the literature. It can however be operationally considered as a continuum between fully structured assessment and spontaneous, unstructured assessment, rather than one or the other (UNESCO, 2015). Across various frameworks of continuous assessment within the literature (Nitko, 1995; Black & William, 1998a; Le Grange & Reddy 1998; Carlson et al., 2003), there are common and consistent features:

- Developing formal and informal assessment tools that are aligned with the curriculum
- Regular assessment of students' learning progress within the classroom
- Using a mix of formative and summative assessment methods
- Utilising a range of assessment techniques in the classroom (oral questioning, group work, role-play etc.)
- Recording student performance both formally and informally (i.e. teacher's own personal records and official school records)
- Using information from assessments to identifying students' strengths, weaknesses and challenges
- Providing timely, and often immediate, feedback to students on their progress
- Evaluation of student performance to modify teachers' and students' work to make teaching and learning more effective

As a note, the use of formative assessment to help guide teaching and learning in classrooms is a concept often referred to in literature as 'Assessment for Learning'. For clarity, in this report any reference to Assessment for Learning or AfL refers to the name of the specific programme that is the subject of this report.

USE OF CONTINUOUS ASSESSMENT IN THE DEVELOPMENT CONTEXT

In many developing countries, teachers stand in front of classes and transmit information to students who are passive in the learning process – often referred to as 'chalk and talk' teaching (Chisholm & Leyendecker, 2008). As a result, Banerjee and Duflo (2011) identified that a major source of poor educational outcomes in developing countries is related to teachers regularly progressing through teaching the curriculum, regardless of the understanding of all students within their class. Whilst high performing students can keep up, poorer performing students fall behind, and these gaps are maintained – leading to students leaving primary education with no basic skills. The implementation of continuous assessment is identified as a potential remedy for this problem, by providing teachers with adequate and timely information on students' performance. Continuous assessment programmes have been implemented across a number of countries, including Ethiopia (UNESCO, 2008) with the aim of improving educational standards.

While there is a wealth of literature on the benefits to student learning in using continuous assessment as a methodology (Black & William, 1998b; Kingston & Nash, 2011), evidence on the effectiveness of continuous assessment programmes on student outcomes in developing countries is relatively scarce despite their widespread use. In Zambia and Malawi, Kapambwe (2010) and Kamangira (2003), respectively, found that students in schools where a continuous assessment programme was piloted improved their learning outcomes. On the other hand, a recent randomised controlled trial (RCT) of the Continuous and Comprehensive Evaluation programme in India, a programme containing teacher training on continuous assessment techniques, found no impact on student test scores (Berry et al., 2020). In terms of impact on teacher behaviour, the evidence is mixed on whether continuous assessment programmes can effectively elicit change in the classroom through teaching practices. Although Banerjee et al. (2016) found that teachers can, even with light training, improve teaching methods to judge student learning progress and adapt accordingly, other studies did not find any impact on teaching practices, with teachers ignoring the crucial formative assessment element (De Lisle, 2016).

CHALLENGES TO IMPLEMENTING CONTINUOUS ASSESSMENT

Poor teacher quality and lack of training is a common reason for continuous assessment failing to be implemented effectively (Kellaghan & Greaney, 2003). This is sometimes due to policy makers implementing a top-down reform with little provision accompanying teacher training or resource materials, leading to an incoherent roll-out with individual schools and teachers implementing it differently based on their understanding (Iipinge & Kasanda, 2013; Modupe & Sunday, 2015). Teachers can often lack the basic fundamentals of continuous assessment to construct assessments, administer them and record results (Uiseb, 2009). The lack of adequate training can often be compounded by a lack of support from school and education officials in monitoring and supervising teachers attempting to implement continuous assessment within their classrooms (Kapambwe, 2010). It is also common that teachers who do receive adequate training and support and can demonstrate a good understanding of continuous assessment still fail to implement it in the classroom. One reason for this is that teachers may experience inertia to adapting their teaching

styles or actively view continuous assessment as not beneficial to either them or their pupils, with some viewing it simply as a box-ticking exercise (Browne, 2016).

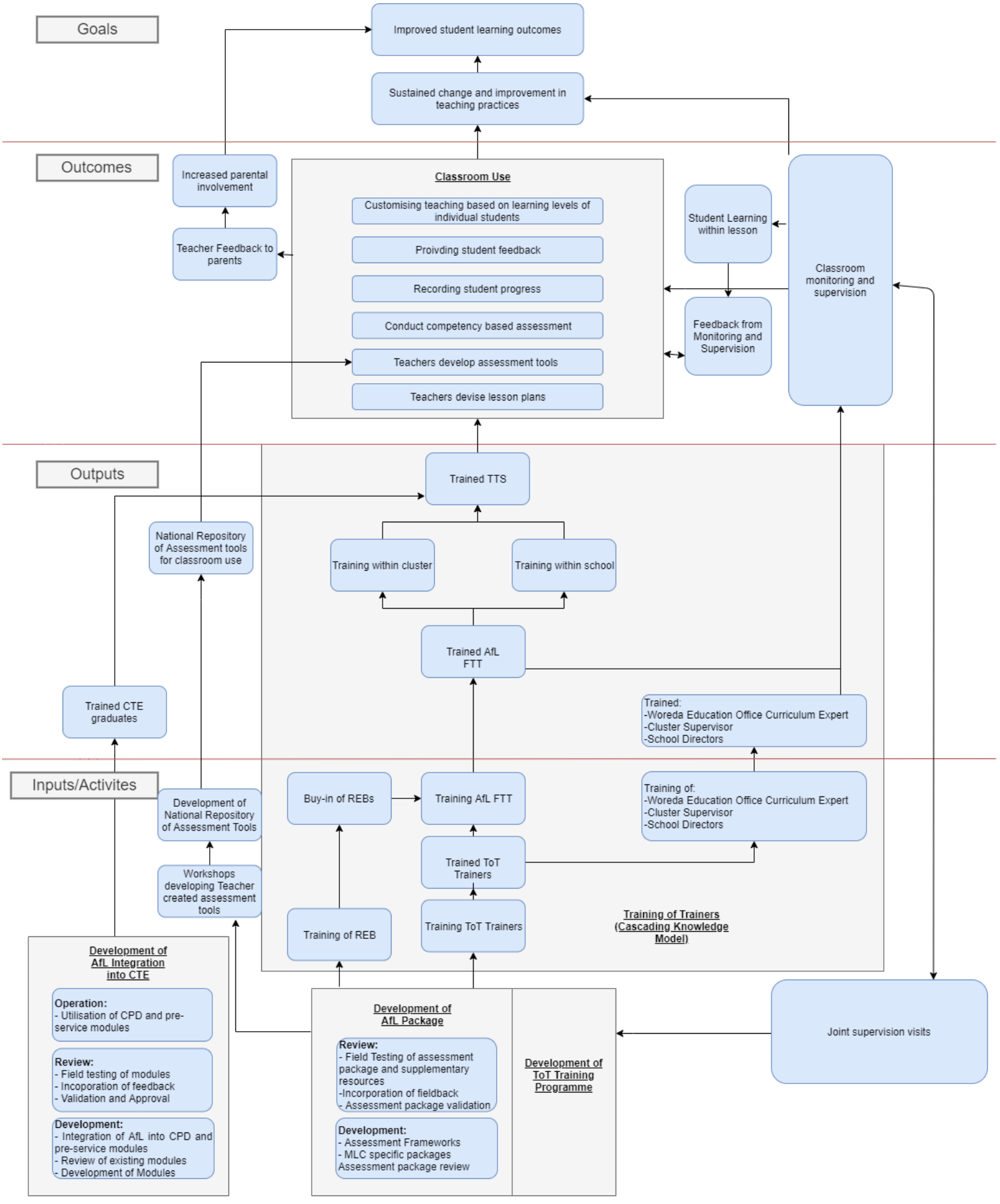
Other challenges to implementing continuous assessment are down to contextual factors, in particular class sizes. Having a large number of students in a class can make providing individualised attention for students impossible (Hayford, 2007). Developing regular testing regimes is also challenging with large class sizes (Kapambwe 2010), and even when teachers manage, the questions are often true-false, multiple choice or other short formats (Muskin, 2017). Teachers, in large classes, often lean on the more capable students for in-class questions at the expense of lower-achieving students (Quansah, 2005). Similarly, regardless of class size, continuous assessment with an ambitious curriculum can place a burden on teachers and increase their workload (potentially at the expense of other activities valuable to student outcomes) (Pritchett & Beatty, 2012) particularly if there are bureaucratic elements imposed such as regularly updating student records and filling in forms. Finally, lack of school infrastructure and supplies can impede the use of continuous assessment as it requires notebooks, stationery and basic templates for teachers to work from and even just storage space for student records (Dowrich, 2008).

1.3 PROGRAMME THEORY OF CHANGE

This section describes the elements of AfL in more detail. As described above, continuous assessment is a multi-faceted intervention with several features. As a framework for the evaluation, a Theory of Change (ToC) was used to map out how the inputs and activities undertaken in AfL should lead to the desired outcomes and impact through causal chains. The AfL programme has previously outlined its ToC at both the design stage and the baseline and midline evaluation stage. For the endline evaluation, we have built upon these ToC models, through discussions with stakeholders, and reviewing programme documents and relevant literature. A key feature of the ToC model we are proposing is explicitly stating the causal pathways and assumptions for the drivers of change to occur. Through outlining the causal chains, the evaluation can move away from a 'black box' strategy (i.e. simply considering the inputs and the outcomes) and test and understand how and why the intervention works or not.

In our simple ToC, represented diagrammatically in Figure 2, we outline the causal pathways from inputs and activities to outcomes and goals. In the following sections we also break the ToC down, focus on key pathways, discuss the various components at play and the assumptions that are required for the pathway to hold.

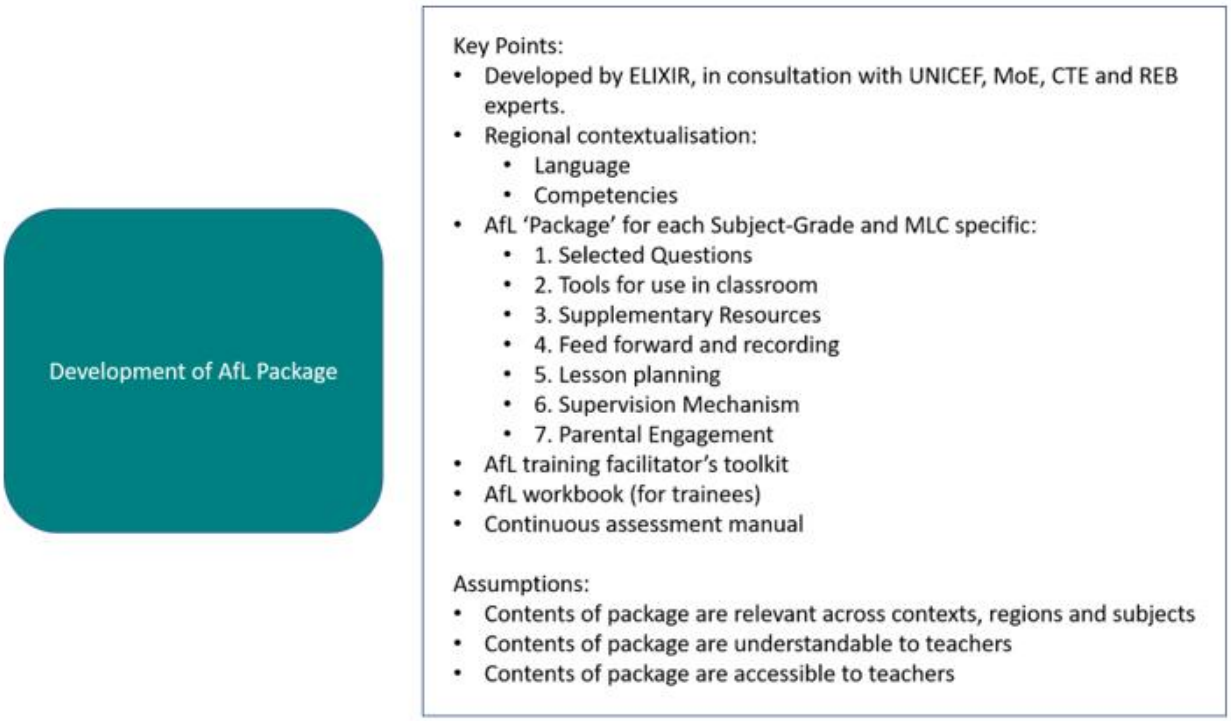
Figure 2. Theory of change



DEVELOPMENT OF THE AFL PACKAGE

The key input/activity underpinning the ToC is the development of the AfL package and training programme for the Training-of-Trainers. Figure 3 outlines the key points of the development of the AfL package, which included the materials for the training and those for use in the classroom by teachers. The tools went through a process of development by ELIXIR, and review by key stakeholders such as MoE, Regional Educational Bureaus (REBs), CTE and UNICEF, before being implemented. To address the assumptions of relevance and teachers' ability to utilise the package, the review process involved field testing of the tools and feedback was incorporated. These assumptions will still also form the basis of evaluation questions to ensure that the review process worked effectively.

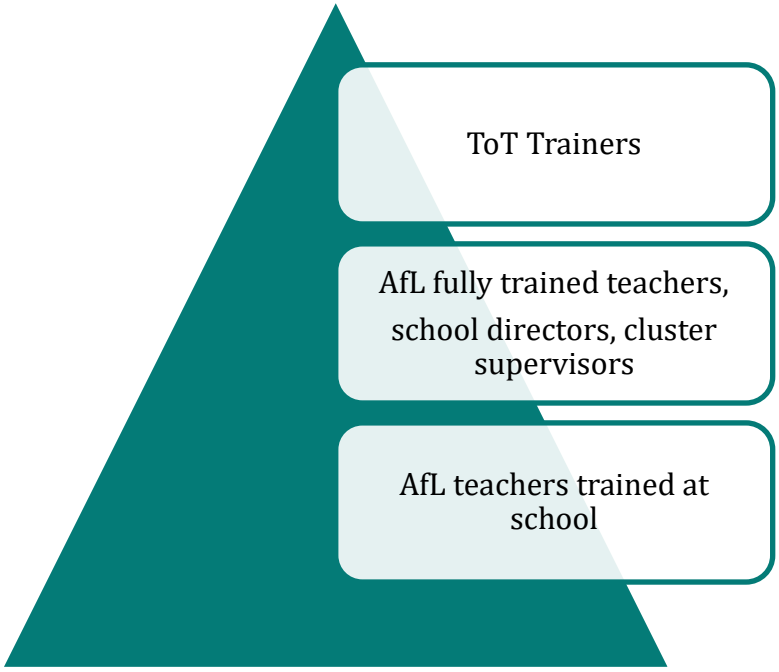
Figure 3. Development of the AfL package



CASCADING KNOWLEDGE MODEL

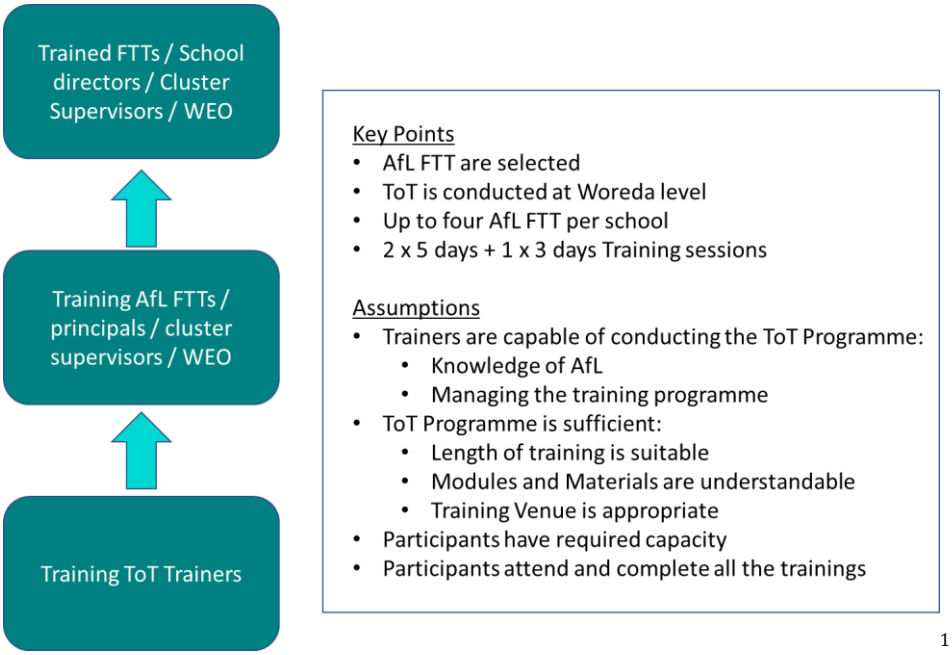
The AfL programme utilises a cascading knowledge model through training-of-trainers (ToT). After developing the training programme and materials, ToT trainers – typically trained staff from CTEs – conduct the training programme for selected teachers – referred to as AfL fully trained teachers (AfL FTTs) at cluster schools. Per cluster school one teacher per subject (Mother Tongue, Mathematics, English and Environmental Science) is invited to participate in the ToT training. The ToT training programme consists of two five-day basic training courses and one three-day refresher course, conducted at the woreda level in the relevant local language. In addition to AfL FTTs, school directors, Woreda Education Office (WEO) Curriculum/Assessment Experts and cluster supervisors are included in the ToT training.

Figure 4. Cascading knowledge model



Upon completion of the training, graduates of the ToT training (AfL FTTs, school directors, WEO curriculum/assessment experts and cluster supervisors) are expected to cascade knowledge to other teachers in the cluster. The cluster supervisors are responsible for organising the cluster training for other teachers in their schools and clusters – who are referred to as AfL Trained Teachers at School (AfL TTSs). In theory, the cascading knowledge model should lead to participating schools and neighbouring schools having fully trained Grade 1-4 teachers in Mother Tongue (e.g. Amharic, Afan Oromo), English (Literacy), Mathematics and Environmental Science at a much lower cost than directly training each teacher. This causal chain is outlined in the ToC and is further examined and broken down into two parts in Figure 5 and Figure 6 below. The first part relates to the training provided directly by the ToT Trainers and the second when the ToT graduates return to their home schools and clusters to conduct training for colleagues.

Figure 5. Cascading knowledge pathway 1

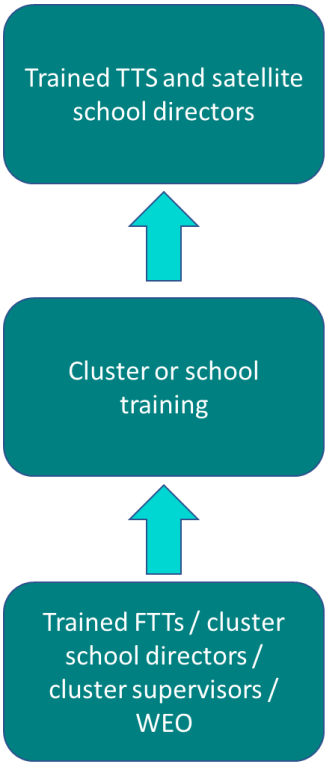


[1]

Assumptions identified include the capacity of trainers and participants, that the training programme developed is appropriate and that participants attend and complete the entirety of the training sessions. The second part of the causal chain behind the cascading model is how the AfL graduates – in particular the cluster supervisors and AfL FTT – begin training and sharing knowledge with colleagues in their schools and cluster, as shown in Figure 6.

[1] In Oromia, ToT training was conducted at a CTE level

Figure 6. Cascading knowledge pathway 2



Key Points
- AfL FTTs, cluster school directors return to their schools and conduct training on AfL for other teachers with support from cluster supervisor and WEO

Assumptions
- AfL FTTs have the capacity to conduct the training
- AfL FTTs have the motivation to conduct the training
- Resources are available to conduct the training:
  - Materials
  - Time
  - Venue
  - Other Costs etc.
- School director buy-in (linked to above)
- TTSs attend and complete all the training
- TTSs achieve learning objectives

Similarly to the causal chain for the ToT training for AfL FTT, cluster supervisors, school directors and WEO curriculum/assessment experts, there are a number of assumptions that must hold for the knowledge and training material to effectively filter down to the AfL teachers trained at school.

CLASSROOM USE OF ASSESSMENT FOR LEARNING

The second key element within the ToC, after the ToT training and cluster training has led to Grade 1 – 4 teachers in the targeted subject areas being trained, is that these teachers put their newly-gained knowledge into practice within the classroom. According to programme materials, the in-classroom use of AfL can be broken down into:

1. Designing lesson plans with the goal of students attaining the relevant MLC linked to the curriculum;
2. Developing assessment tools that will allow them to assess whether students have achieved the MLC;
3. Implementing the lesson plan and conducting the competency-based assessments;
4. Recording students' progress; and
5. Providing effective and timely feedback to students on their progress in the assessments.

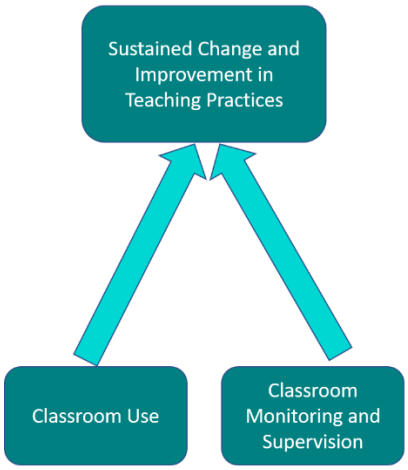
There is an additional overarching element of teachers customising their teaching based on the learning levels of individual students.

Figure 7. AfL Classroom use components

Classroom Use	Description	Assumptions
Customising Teaching Based on Learning Levels of individual students	• Teachers use assessment results to modify teaching.	• Ability to analyse the data of student progress • Use data to identify where adjustments to their teaching needs to be made • Identify and implement appropriate measures to support individual students. •
Providing Feedback	• After the assessment, students receive feedback on their performance. • Useful feedback informs student about their progress and how they can improve.	• Teachers have capacity to provide useful feedback. • Teachers provide feedback in a timely manner. • Teachers have time to provide useful feedback. • All students receive feedback.
Recording Progress	• Once assessed, student performance in competencies should be recorded. • Use records to understand student performance and compare with peers.	• Consistent system of recording all students' performances in assessments. • Updated in a timely manner. • Data kept in a useable format
Conduct Competency Based Assessment	• Assessments outlined in lesson plans are conducted	• Conducted in a timely manner • All students participate in assessment
Develop Assessment Tools	• Learning competencies are used to develop competency based assessments. • Assessments can be Oral, Written Practical. • Can be developed or taken from repository. • Incorporate supplementary materials	• Teachers understand MLC. • Teachers have capacity to develop questions/assessments linked to MLCs. • Access to repository. • Teachers have access to supplementary materials
Devise Lesson Plans	• Develop lesson plans that outline the competencies. • Outline the teaching and student activities and the specific assessments used.	• Teachers have capacity to develop lesson plans • Teachers have time to develop lesson plans • Lesson plans are reviewed by department head.

In Figure 7 we describe each of these components of classroom use of the AfL programme along with assumptions for the components to work effectively. These, taken together, should lead to classroom use of the AfL programme having longer term impacts on teaching practices and student learning. For the programme to have impacts on teaching practices, the changes must be sustained in the longer term with feedback being provided to teachers. In Figure 8 we show the link between classroom use and the AfL's system of monitoring and supervision.

Figure 8. Teacher practice change pathway



Key Points:
- Supervision Types:
  - Peer Supervision by Key Teachers
  - School Supervision by School Directors
  - Cluster Supervision by Cluster Supervisors
  - Join Supervision by ELIXIR/UNICEF/REB
- Focus of Supervision is on student learning within the classroom.
- Based on the level of student learning within classroom, feedback is provided to the teacher to identify strengths and weaknesses in implementing continuous assessment to improve their practices.
- Feedback and good practice is shared throughout the school

Assumptions:
- Teachers feel comfortable conducting and receiving Peer Supervision
- Supervisors have time to conduct Supervision
- Supervisors have the capacity to conduct Supervision
- Supervision is conducted on a timely basis.
- Teachers are provided useful feedback.
- Teachers incorporate feedback into their teaching practices.

The various supervision types are used to firstly monitor that classroom use of AfL is occurring and over a substantial period of time, and secondly, to allow for observation of practice to provide feedback.

CHAPTER 2: EVALUATION PURPOSE, QUESTIONS, STAKEHOLDERS AND TIMELINE

Chapter 2 presents an overview of the purpose of the evaluation, the evaluation questions to be addressed, the stakeholders and possible users of the evaluation and the evaluation timeline.

2.1 EVALUATION PURPOSE

UNICEF Ethiopia requested that C4ED conduct an endline evaluation[2] of the AfL programme to identify the programme's impact and to inform the design and strategic direction of future programmes to improve the quality of learning outcomes in the new country programme. The endline evaluation addressed the following objectives:

- Assess the impact of the AfL programme on teacher practice, learner participation and on assessment of learning; and
- Evaluate the extent to which the AfL intervention has improved teacher practice and learning outcomes; and
- Assess how key stakeholders (students, teachers/school administrators, parents, communities) view the relevance, efficiency and sustainability[3] of the AfL training and the AfL materials/package.

The evaluation provides an opportunity to identify what worked well and what did not, and to disseminate findings of best practice across the regions on successful practices of teachers in conducting effective classroom assessment.

2.2 EVALUATION QUESTIONS

The evaluation has complied with evaluation criteria of the Development Assistance Committee (DAC) of The Organisation for Economic Co-operation and Development's (OECD). The evaluation criteria include Relevance, Effectiveness, Efficiency, Impact, and Sustainably of OECD-DAC criteria. However, since DAC criteria are neutral regarding gender equity we included additional criteria to assess the gender equity of the programme.

As per the Terms of Reference (included within the Annex) issued by UNICEF Ethiopia, the following main evaluation questions were provided:

[2] The AfL programme was subjected to a baseline and a midline evaluation conducted by Centre for British Teacher (CfBT) Education Trust (now Education Development Trust) in 2013 and 2015, respectively. The baseline and midline evaluations referred to AfL as the "Accelerated Development of Literacy, Numeracy and Life Skills among First Cycle Learners in Ethiopia". Due to modifications in the scope and direction of the programme throughout its lifetime, this endline evaluation is not directly related to the baseline and midline.

[3] This evaluation objective was modified from the ToR, initially "Assess how key stakeholders (students, teachers/school administrators, parents, communities) view the relevance and effectiveness of the AfL training and the AfL materials/package". This was to include the DAC criteria of efficiency and sustainability within the objectives.

Evaluation Criteria	Evaluation Question
Relevance and gender equity	1. What is the quality and relevance (including gender equity) of programme interventions (materials, modules, tools and training)? How can they be improved for future use?
Sustainability	2. What can the programme do at both policy level and decentralised structure levels to improve programme interventions and impact and promote sustainability and the scale-up of promising practices? 7. Have any changes been achieved in relation to policy, practice, attitudes of practitioners and policy makers?
Effectiveness	3. To what extent have programme inputs made a difference to teachers' ability to use continuous assessment techniques?
Impact	4. To what extent are changes in teacher practice attributable to the AfL project activities?
Efficiency and effectiveness	5. What were the most efficient and effective approaches used by regions, woredas, schools or teachers to bring about change? What worked, what did not work, and why?
All	6. What overall lessons can be learned from the delivery of the AfL?

The questions were incorporated into an overarching evaluation matrix that broke the main evaluation questions down into sub-evaluation questions (SEQs) with accompanying measures and indicators along with the proposed source of the information. The full evaluation matrix is shown in Appendix A with all SEQs and sources of information.

2.3 EVALUATION STAKEHOLDERS

UNICEF Ethiopia, the MoE and the REBs involved with implementing the programme are expected to be the primary intended users of this study. UNICEF is expected to use the findings to inform the scale-up of education programmes it is involved in. The MoE and REBs are expected to use the evaluation findings to develop and implement policy directives to improve classroom practice of classroom assessments for improved learning outcomes. The evaluation is also be expected to be used by schools and education clusters to guide the development of a functional and user-friendly system of classroom assessment.

2.4 EVALUATION TIMELINE

The evaluation followed a linear approach comprising of four main stages. Each phase included activities contributing directly or indirectly to evaluation deliverables. Figure 9 visually outlines the key phases of the evaluation and timelines with associated deliverables.

Figure 9. Evaluation Phases

Inception Phase	• **September - December 2019** • Inception mission • Tool development • Evaluation design • Deliverable: Inception report
Field Data Collection	• **January - March 2020** • Field visits and data collection
Data Processing and Consolidation	• **March - June 2020** • Data cleaning and analysis • Deliverables: Fieldwork report and Preliminary findings report
Reporting and Dissemination Phase	• **June - July 2020** • Draft report writing • Feedback and review of draft report and revision of final report • Dissemination • Deliverable: Final evaluation report

CHAPTER 3: EVALUATION DESIGN, METHODS, QUALITY ASSURANCE AND ETHICS

3.1 EVALUATION DESIGN

Based on discussions with staff from UNICEF, MoE, the Oromia REB & ELIXIR, the C4ED evaluation team finalised a strategy based on a mixed-methods approach involving both qualitative and quantitative data collection. Such an approach allows for addressing all relevant questions, and not only the ones acquiescent to a certain method. Quantitative impact evaluations can have a strong internal validity, and qualitative methods help explain the context of the intervention and provide an understanding for the settings in which the results may be generalised.

QUANTITATIVE DESIGN

The focus of the quantitative analysis was to measure the impact of the AfL intervention on (i) teacher behaviour and (ii) student outcomes in Mother Tongue (Afan Oromo), English, Mathematics and Environmental Science. For any causal interpretation on the impact of a programme on its desired outcomes, a good impact evaluation relies on the quality of the chosen counterfactual situation. The counterfactual situation corresponds to what would have happened if the AfL programme had not been implemented. Of course, since at any one point in time a specific teacher either benefits or not from an intervention, by definition, the counterfactual situation is just a conceptual view and will never be observed. As a result, the task of the evaluation team is to construct or mimic the counterfactual by selecting a control group that would behave as close as possible to what would have happened in the absence of the AfL programme. A properly implemented RCT, which is considered the most reliable method to assess impact and which has now become the gold standard for impact evaluation, could provide an average value of the effect of the intervention.

Given that AfL was not implemented in schools through random assignment, a randomised controlled trial is not feasible. In addition, while some schools had been chosen as controls in the initial baseline evaluation design in 2013, these schools later received the intervention and were no longer suitable for the counterfactual. Finally, although a baseline and midline data collection had been conducted, no data or instruments could be made available to the evaluation team, impeding our ability to observe changes over time. Given these shortcomings, our endline evaluation is based on quasi-experimental methods, namely on matching methods. Such methods rely on selecting a group of control schools not affected by the AfL programme. The impact is then assessed by comparing the performance of the pupils and teachers in AfL schools to the ones in non-AfL schools.

To minimise the selection problem that AfL schools might be different than non-AfL schools, leading to differences in outcomes not caused by AfL itself, we use a basic matching technique, using available administrative data on school inspection scores, to match schools with similar characteristics. The matching process is discussed in more detail in Section 3.2.2. In this case the unit of comparison is schools, as we expect most of the differences can be observed between schools. While matching designs are not as robust as RCTs or experimental designs, they provide an alternative where such designs are not feasible or possible.

The sample for the quantitative analysis consisted of 60 schools in the region of Oromia. These were divided into three treatment groups, namely cluster (treated) schools, satellite (semi-treated) schools and control schools, as illustrated in Figure 10. The first group of 30 schools are

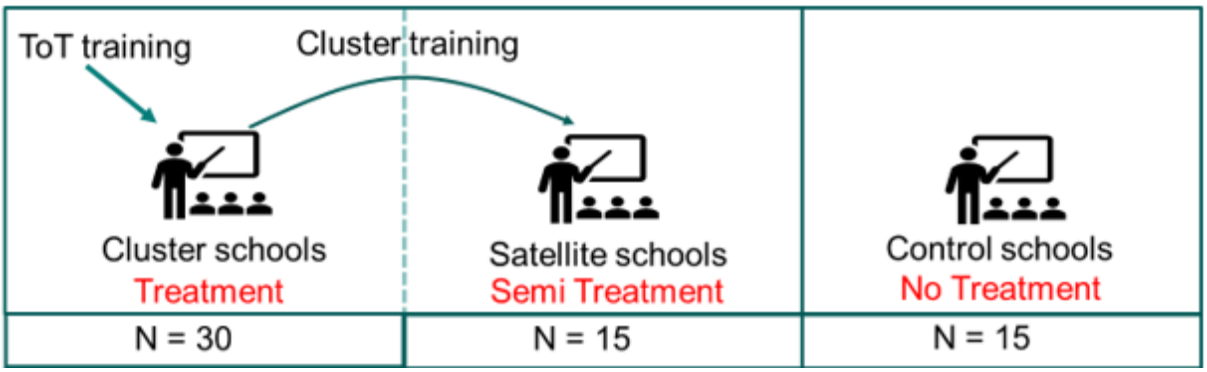
those targeted by UNICEF and its implementation partner ELIXIR to receive the AfL intervention. They are the core group of cluster (treated) schools.

The second group of 15 satellite (semi-treated) schools were not directly targeted but received some training through the cluster supervisors and AfL fully trained teachers in the same cluster. Depending on the assumption we make about the effectiveness of the cluster training, these 15 schools can be seen as control or 'cluster-treated' schools. Given that not only the cluster supervisor and the AfL fully trained teacher but also education officials in the same cluster play a crucial role in adopting AfL, these schools provide an interesting comparison group.

Finally, 15 control schools, which should have received no direct or indirect treatment through in-service training, were selected as the 'woreda control schools'. However, we note that these schools may have benefitted from some contamination effects of teachers moving from cluster schools or satellite schools or having received pre-service training in the Oromia CTEs that integrated AfL into their programmes.

For clarity, hereinafter, in this report we refer to the groups of schools in the study that are differentiated by their treatment status as treatment groups. The three treatment groups are 1) Cluster schools (treatment), 2) Satellite schools (semi-treatment), and 3) Control schools (control).

Figure 10: Sample composition in the Oromia region



With this control group design, we should be able to observe different intensity of treatment effects and compare:

1. AfL cluster schools against non-AfL control schools
2. AfL cluster schools against satellite schools
3. Satellite schools against control schools

By comparing the groups of schools, we can examine if there are differences in the usage of pedagogical tools and mode of instruction. Whether we can observe different learning outcomes will depend on the assumptions we make about the time horizon of any effects. Furthermore, any differences we will observe should be taken with caution since this is a quasi-experimental design. There may be potential unobserved school characteristics, independent from the AfL training, that also influence students' learning outcomes. Since the AfL programme mainly acts through teachers' mode of instruction and teacher behaviour, we can assess these intermediate effects, in line with the AfL programme ToC.

With this design we aimed to quantify the AfL programme effect relative to schools that received no treatment (point 1 above). We also hoped to observe if the cascading model is working within clusters (points 2 and 3).

QUALITATIVE DESIGN

The focus of the qualitative analysis was to get a deeper understanding of the 'how' and 'why' of what happened in the implementation of AfL. The qualitative data collection took place after the quantitative data collection in Oromia, allowing us to tailor the qualitative tools to initial findings from the quantitative survey. In addition to Oromia, the qualitative analysis included two other regions, Tigray and Benishangul-Gumuz (where the AfL programme had been recently implemented). The qualitative analysis therefore enabled us to paint a broader picture of the implementation of the AfL programme across the different regions. Among the numerous regions where AfL has been implemented, Tigray was of particular interest, because it followed Oromia in integrating the AfL programme into the curriculum at CTEs. The intention of collecting data from both regions was to enable us to further explore this avenue for scale-up. In Benishangul-Gumuz the AfL programme has been implemented in refugee schools. While access to refugee schools is restricted, we included schools in the same woreda as refugee schools (i.e. host community schools) in our sample. By visiting these schools, we hoped to gain insights into how the programme has been implemented in and around refugee schools with highly vulnerable and marginalised students.

In total we selected 14 schools for qualitative data collection, 6 in Oromia and 4 each in Tigray and Benishangul-Gumuz. In all schools we conducted in-depth interviews with teachers and school directors. Furthermore, we hosted two focus group discussions in each region with parents from selected schools. In order to get a broader picture of the implementation in the different regions, we visited two woredas in each region. In Tigray and Benishangul-Gumuz we randomly sampled two woredas, taking security and logistical restrictions into consideration. In Oromia we focussed on schools that were already included in the quantitative data collection, allowing us to better triangulate the qualitative findings with the quantitative results. In all regions we aimed to balance the sample equally between cluster and satellite schools. Including both cluster and satellite schools in our qualitative sample allowed us particularly to explore drivers and challenges in relation to the cascading knowledge model. Moreover, we included one cluster and one satellite school from a refugee host community in the sample of Benishangul-Gumuz.

3.2 EVALUATION METHODOLOGY

3.2.1 QUANTITATIVE SURVEY DATA COLLECTION INSTRUMENTS

As the survey instruments from the baseline and midline evaluations were not available, the C4ED evaluation team developed new survey instruments for data collection.

To measure student outcomes in the focal subjects for the AfL programme – Mother Tongue (Afan Oromo), English, Mathematics and Environmental Science – written tests were administered to randomly sampled students from Grade 3 (10 students) and Grade 4 (10 students). The tests contained questions provided by the implementing partner of AfL, ELIXIR Research and Consultancy. These questions were initially developed by teachers and aligned with the appropriate MLCs for each grade. The questions, reviewed by experts and through a series of consultancies with teachers, were assessed against criteria of validity, fairness, objectivity,

acceptability, effectiveness and efficiency. Selected questions were field tested at schools and analysed psychometrically for indicators of quality and the minimum threshold for acceptance. During the data collection pre-test, the questions selected for the test were reviewed and minor adjustments were made, including removing questions that showed little variation to ensure a range of difficulty levels. The tests were conducted in the following order: Afan Oromo, English, Mathematics and lastly Environmental Science and each was allocated 15 minutes. Upon finishing each test, the invigilator would collect the scripts before moving on to the next subject. At the end of each field day, the enumerator would mark the scripts using a mark scheme and input the data into a Survey Solutions questionnaire.

The school director questionnaire was developed to gain information about the school characteristics, leadership and buy-in for the programme. Teacher questionnaires were developed to collect information on teacher characteristics and background, time use, knowledge, use and reflections on AfL and teaching practices.

We complemented the teacher interviews by conducting in-class observations of teaching practices and behaviour in the classroom. In this way, we aimed to identify whether and to what extent the pedagogical strategies, classroom activities and assessment techniques used corresponded to those promoted as part of the intervention. This tool, albeit only taking into account the currently observable situation and features, aimed to serve as an entry-point for in-depth analysis, especially relating the questions of effectiveness, relevance, and sustainability of AfL. We conducted three classroom observations per school. All instruments were translated into the appropriate language, Afan Oromo.

3.2.2 SAMPLING FOR QUANTITATIVE DATA COLLECTION

We used a multi-level sampling approach to select schools. The initial step was to select zones in Oromia from those in which the AfL programme had been implemented. The selection of zones was non-random and constrained by only being able to select zones that were considered secure for field teams to visit and collect data in, which numbered six in total. The sampled zones were: Arsi, West Arsi, East Shoa, West Shoa, South West Shoa and West Hararghe. It should be noted that secure zones tended to be closer to Addis Ababa and may not be representative of Oromia as an entire region. Once the zones were selected, to achieve the target sample size of 60 schools, we used all treated woreda in the zones.

DEVELOPING A SAMPLING FRAME

To sample schools, we developed a sample frame using three databases provided by the Oromia REB including a list of AfL Cluster Schools, a list of inspected schools in 2014-2016 and a list of inspected schools in 2017. Substantial effort was made to merge the lists, remove duplicates, correct discrepancies and clean the data to produce a single sample frame that included information on location, school type, locality (rural/urban) and the school inspection score[4]. As cluster (as in clusters of schools from which to select satellite schools) information was not provided within the database, phone calls were made to WEOs to identify the clusters that schools within treated woredas belonged to. Schools that were in treated woredas but outside of treated clusters were removed from the sample frame.

[4] Miliqaye School was removed from the sampling frame as it could not be matched with any of the school inspection lists and therefore did not have an inspection score.

Four control woredas were randomly selected from each of the following zones (grouped based on proximity): Arsi/Arsi West, West Shoa/South West Shoa, East Shoa and West Hararghe. The selected control woredas were Jibat (West Shoa), Gimbichu (East Shoa), Ancaar (West Hararghe) and Jaju (Arsi).

The sample frame, in total consisted of 651 schools, with 40 cluster schools, 445 satellite schools and 165 control schools.

MATCHING PROCESS

Ideally, we would have a rich set of data on school characteristics that are time-invariant or collected prior to the programme implementation for computing propensity scores – the likelihood that a school would enrol in the programme. However, as the only available data on school baseline characteristics within the sample frame were the school inspection scores and whether the school was urban or rural we aimed to match exactly on these two characteristics. The school inspection score is an indicator of school quality that is provided by an independent inspectorate and based on standardised guidelines across Ethiopia since 2014. The inspection score is a measure (in percentage) that takes into account the following domains: input (25%), process (35%) and output (40%) levels of the school. A description of each criterion is provided in Table 1. These scores are combined for a single indicator of school quality[5].

Table 1. School inspection score criteria (from MoE, 2013)

Criteria	Focus area
Input	**Focus Area 1:** School facility, buildings, human and financial resources **Focus Area 2:** The learning environment
Process	**Focus Area 3:** Learning and teaching **Focus Area 4:** The school's engagement with parents and the community
Output	**Focus Area 5:** Student outcomes and ethics

To match cluster schools with control schools, both types of schools were allocated into four performance quartile bands based on their inspection score. Within each treatment woreda, each cluster school was matched with a similar school considering inspection score and whether it was urban. To be precise, each cluster school was matched with the single nearest neighbour for performance score within their performance quartile band that was the same urban/rural type (with no replacements), and the sample selected the best match (i.e. minimum difference in performance score between cluster and control). Matches were made in rounds, selecting one match from each quartile band before starting again to ensure a distribution of sampled schools across all performance scores. A diagrammatic representation of the matching process is provided in Appendix B.

The process for matching satellite schools with cluster schools was similar – within each treated cluster, the cluster schools were matched to the nearest neighbour within their cluster and the two best matches per woreda were selected, meaning that half the cluster schools had a matched

[5] A detailed description of the school inspection process is outlined in the National General Education Inspection Framework (MoE, 2013).

satellite school. All matching was conducted through an automated do-file, implemented in the statistical software Stata.

During data collection, some schools had to be replaced, with the majority of schools in Fantalle woreda (discussed in more detail in the section 'Quantitative fieldwork challenges' below). Appendix D provides more detail on all the cases where schools were replaced. It should be noted, that although Baha Biftuu was replaced, the team was able to complete part of the data collection (surveys with the school director and teachers) and it was only replaced after it became clear that classes would not resume (and therefore further data collection was not possible). The school director and teacher interviews from Baha Biftuu remain in the data and were included in the analysis. Replacement schools were selected – where possible – from the same woreda/cluster and were matched based on the inspection score and whether the school was in an urban location.

The balance of matching characteristics between the three treatment arms for the final sample is shown in Appendix E. If a sample is balanced it means that the matching was successful in selecting similar schools, at least on the matching characteristics. The only significant difference ($p \leq 0.05$) between the three groups was that there were more urban control than satellite schools – this was due to a shortage of urban satellite schools in the sample frame. Selecting more urban satellite schools would have had an effect on the matching on the school inspection score, and the trade-off for school quality would have been more problematic than for the urban/rural split. Otherwise, the schools in the final sample (including replacements) were well balanced on the matching criteria.

3.2.3 QUALITATIVE SURVEY DATA COLLECTION INSTRUMENTS

IN-DEPTH INTERVIEWS INSTRUMENTS

We planned to conduct semi-structured in-depth interviews (IDIs) with a sub-sample of up to 56 Grade 3-4 primary teachers who have benefited from the AfL programme's in-service teacher training and material provision services. In each of the 14 schools sampled, up to four teachers of Mother Tongue, English, Environmental Science or Mathematics were interviewed. Semi-structured interviewing was used to encourage two-way communication between the interviewer and interviewee. Pre-determined, but open-ended questions set a framework of themes explored, while new ideas could be brought up as the conversation progressed. Interviewers were provided with an interview guide and had received training on how to adequately use it to ensure the reliability and comparability of the data collected.

We also interviewed school directors and selected cluster supervisors to shed light on the intensity of the training and the support that their institution received as part of the AfL programme, and which further services and capacities may be required to help teachers implement their learning.

In addition, we carried out IDIs with one ToT trainer per region. As facilitators of the teacher training programme and mentors to the trained teachers, their input is crucial to evaluating the intervention's long-term viability. During the interviews, we aimed to capture important information on the trainers' coaching capabilities and potential gaps therein. Moreover, we conducted IDIs with CTE instructors in Oromia and Tigray as both regions integrated the AfL programme into their curriculum. Lastly, we interviewed personnel at WEO, who have specialised

knowledge on the AfL programme and Ethiopia's education and training policy and the education system.

KEY INFORMANT INTERVIEWS INSTRUMENTS

To get a full picture of the intervention, we conducted Key Informant Interviews (KIIs) with officials from each of the three REBs, as well as staff from UNICEF and staff from the implementing partner ELIXIR.

FOCUS GROUP DISCUSSIONS

Focus group discussions (FGDs) are a powerful evaluation tool which enable a variety of perceptions on an intervention to be elicited, facilitated by collective stimulation. We conducted two FGDs per region with parents and community members. Planned FGDs were split by gender, therefore, in each region, one FGD was conducted with women and one with men. In total, three FGDs each with women and men were conducted. These FGDs enabled us to collect data on collective and individual experiences and attitudes, but also expectations vis à vis AfL, its modalities, its influence on parental engagement with their children's education and to identify possible bottlenecks of implementation.

QUALITATIVE SURVEY INSTRUMENTS

For the qualitative part of the study the research team developed 11 interview guidelines and one focus group topic guide. The different instruments were adapted specifically to the respondents, as they are involved in the AfL programme in different ways and thus are able to provide insights on different aspects of the programme.

3.2.4 SAMPLING FOR QUALITATIVE DATA COLLECTION

This section describes how the participants for the qualitative data collection were sampled, beginning with interviews in school (namely IDIs with teachers and school directors) followed by non-school-based data collection (i.e. FGDs, IDIs with cluster supervisors, woreda education officials, ToT trainers, and CTE instructors, and KIIs). It should be acknowledged that the aim of qualitative research is to get a deeper understanding of the perceived reasons for and mechanisms of impact, rather than to achieve representativeness. Nevertheless, we aimed to reduce bias within our purposeful sample by randomising selection whenever possible.

SCHOOL-BASED IN-DEPTH INTERVIEWS

In order to select teachers and school directors for IDIs, we first selected schools. For Oromia, we based the selection process on initial findings on having given or received school and/or cluster training, from the quantitative survey data. In Tigray and Benishangul-Gumuz we relied initially on school lists provided by UNICEF. Subsequent information on having given or received training was gleaned from school directors through phone calls.

Based on first findings from the quantitative survey, we surmised that the cascading knowledge model faced challenges in its implementation in Oromia. As explained in the Theory of Change, the AfL programme expects teachers from cluster schools, who have obtained the ToT training, to provide training to fellow teachers and school directors in their school (school training) and in

surrounding satellite schools (cluster training) after having completed the ToT training. However, few schools in the quantitative survey reported having given school and/or cluster training to fellow teachers and school directors. Moreover, preliminary analysis revealed that only a handful of satellite schools included in the quantitative sample reported having received any cluster training. We thus decided to include more cluster schools than satellite schools in the sample, rather than equally splitting the sample between both. In order to gain more insights on the implementation of the cascading knowledge model, and the drivers of and barriers to implementing school and cluster training, we also adapted our sampling strategy for the qualitative data collection to include both cluster schools that had and had not given school and/or cluster training. For satellite schools, we only considered schools that had received cluster training.

Therefore, in Oromia, we included two cluster schools that had given the school and/or the cluster training, two cluster schools that had not provided any training, and two satellite schools that reported having received cluster training. Given security and logistical restrictions, we considered four different woredas across three zones, namely West-Hararghe (2 woredas), West Shewa (1) and South-West Shewa (1). Due to proximity we grouped West-and South-West Shewa, and randomly sampled one zone. The final random selection of woredas were Tole and Ambo in South-West Shewa and West Shewa, respectively.

In order to mimic the same sampling strategy for Tigray and Benishangul-Gumuz we phoned cluster school directors in both regions to gather information on whether they had given school and cluster training and, if so, to which schools. In Tigray, we focussed our sampling on woredas in the central, eastern and southern zone, because of safety and logistical concerns. As a first step we randomly sampled two from the following woredas: Saharti Samre, Raya Azebo, Kelete Awelallo, Gulomekeda, Mereb Leke, Tahtay Maychew. The randomisation resulted in the selection of Kelete Awelallo and Gulomekeda woredas. We then phoned school directors of cluster schools that had taken part in ToT training. Schools that could be reached all reported having provided school and/or cluster training. Therefore, within the selected woredas we randomly sampled one cluster school and one satellite school, respectively.

In Benishangul-Gumuz, in contrast, no cluster school directors reported that cluster or school training had been given. Hence, we further adapted the sampling in Benishangul-Gumuz to only include cluster schools. After considering safety concerns in Benishangul-Gumuz, we were left with two possible woredas for the qualitative data collection, namely Bambasi and Homosha. We randomly sampled two schools in each woreda. While we initially purposively included one refugee school in our sample, due to safety concerns we were not granted access to the school. We therefore replaced the sampled refugee school with a non-refugee school within the host community.

Table 2 gives an overview of the final qualitative sample across the three regions.

Table 2. Final qualitative school sample

School type	Oromia	Tigray	Benishangul-Gumuz	Total
Cluster school that gave school and/or cluster training.	2	2	0	**4**
Cluster school that did not give any training	2	0	4	**6**
Satellite school that received cluster training	2	2	0	**4**

Total	**6**	**4**	**4**	**14**

The second sampling step for IDIs at schools encompassed selecting respondents at schools. In every school we visited, we aimed to interview the school director and up to four teachers of relevant subjects, i.e. Mother Tongue, English, Mathematics and Environmental Science in Grades 1 to 4. In case the school director had not attended a training on AfL techniques, we interviewed the most senior AfL trained staff member instead. In case no senior staff member had taken the AfL training, we interviewed the school director regardless.

For teachers, the sampling strategy differed at cluster and at satellite schools. At cluster schools we aimed to interview up to three ToT-trained teachers and at least one school-trained teacher. In case the cluster school had not provided school training (and therefore did not have any school-trained teachers), we either interviewed an additional ToT-teacher if possible or other non-trained teachers, making sure to interview four teachers in total per school. At satellite schools we interviewed as many cluster-trained teachers as possible and included other non-trained teachers in case there were fewer than four cluster-trained teachers present. We also aimed to interview four teachers at satellite schools.

FOCUS GROUP DISCUSSIONS

For the FGDs, we randomly sampled from which school the FGDs with women and men would be sampled. If possible, we conducted one FGD in a cluster school and one in a satellite school in order to obtain a broad picture of parents' involvement at different types of schools. The FGD participants were recruited in collaboration with kebele officials.

NON-SCHOOL-BASED IN-DEPTH INTERVIEWS AND KEY INFORMANT INTERVIEWS

We also conducted IDIs with cluster supervisors, woreda education officials, and ToT trainers in each region, and in Tigray and Oromia, with CTE instructors. We interviewed one of each per region. Cluster supervisors and woreda education officials were randomly sampled, given that we only visited one cluster per woreda. After taking logistical constraints into consideration, ToT trainers and CTE instructors were randomly sampled from lists provided by ELIXIR or the REB.

Key informants from UNICEF, ELIXIR, and the REBs were purposively selected by the research team in collaboration with UNCIEF.

3.3 DATA COLLECTION AND ANALYSIS

3.3.1 QUANTITATIVE DATA COLLECTION

The quantitative data collection tools were programmed in the Survey Solutions data collection software by the C4ED evaluation team and included translations of the tools from English to Afan Oromo. The data collection was conducted by C4ED in collaboration with Reliance Consultancy PLC (hereinafter referred to as Reliance) based in Addis Ababa, Ethiopia.

In preparation for the data collection, C4ED and Reliance, with the support of UNICEF and the Oromia REB, conducted a pre-test, field staff training and pilot.

For the pre-test, a briefing took place on 23rd January 2020 in Addis Ababa to introduce two staff members from Reliance (later to be trained as supervisors), who would be conducting the pre-

test interviews, to the tools and the overall project. The pre-test took place in Haile Aba Maresa Primary School in Sire woreda on 24th January 2020 with the field coordinator from Reliance and two members of the C4ED evaluation team in attendance. The school was a cluster school provided to us by the Oromia REB and subsequently removed from the sampling frame. As schools were closed due to end of term exams, it was arranged that teachers and the school director would be interviewed while students would complete the assessment. During the pre-test, classroom observations were not possible, however on the whole the pre-tests were useful in fine-tuning the tools.

Prior to the field staff training, the C4ED evaluation team also organised a workshop and invited representatives from UNICEF, ELIXIR, the Oromia REB and two current teachers from an AfL school in Addis Ababa (outside of the sample) to attend. The workshop took place on 30th January 2020 and outlined the evaluation strategy and an overview of the tools. It provided an opportunity for the discussion and clarification of concepts that were important to the evaluation aims. In addition, to further test the questionnaires, a thorough review of the teacher questionnaire was done through mock interviews with the teachers who attended.

Field staff training took place in Addis Ababa and commenced on 5th February 2020, with representatives from C4ED leading the training. The training lasted for five days, including a full field pilot on 12th February 2020. Training included an introduction to the project, an in-depth review of the tools and also provided an opportunity to fine-tune translations. The training also included modules on interviewing skills, research ethics, fieldwork protocols and using the Survey Solutions Collect software. The field pilot took place at Cirao Agamsa (a cluster school), Ibiseta Oda (a satellite school) and Borara Sadeni (a control school) in Sire woreda and all trainees conducted practice interviews, student learning assessments and observations. A debrief for the field pilot was conducted on 13th February 2020 where issues and feedback were discussed amongst the trainees, supervisors, field coordinator and C4ED staff and necessary adjustments were subsequently made to the tools.

The fieldwork launched in on 18th February 2020 after teams travelled to their respective zones. The final data collection concluded on 5th March 2020 as the teams returned to Addis Ababa. One member of the C4ED evaluation team oversaw one week of fieldwork in West Hararghe, while another team member monitored the data collection remotely. Field supervisions and coordination were conducted by a field coordinator from Reliance.

Table 3. Key data collection dates

Dates	Activity
23/01	Pre-test brief with local staff
24/01	Pre-test in Sire
30/01	Workshop with C4ED/UNICEF/ELIXIR/REBs
05/02 – 07/02	Classroom Based Training
12/02	Full Field Pilot in Sire
13/02	Pilot Debrief
18/02 – 05/03	Field Work

In total, 60 schools were visited by the field teams, across 6 zones in Oromia. Table 4 outlines the summary of completed interviews across the four tools (school director, teacher, classroom observations and student assessments). In total, 60 school director interviews, 176 teacher interviews, 177 classroom observations and 1,161 student assessments were completed.

Table 4. Summary of completed interviews

Treatment Group	School director	Teachers	Classroom observations	Student assessments
Cluster School	30	93	87	580
Satellite School	15	43	46	299
Control School	15	40	44	282
Total	**60**	**176**	**177**	**1,161**

Table 5 breaks down the number of interviews by woreda and zone. In East Shoa and West Hararghe, there were three sampled woredas, in West Shoa and Arsi there were two, while in Arsi West and South West Shoa, there was only one sampled woreda.

Table 5. Summary of completed interviews by Zone and Woreda

Zone	Woreda	School director	Teachers	Classroom observations	Student assessments
Arsi	Jaju	4	11	12	74
	Sire	4	16	12	80
Arsi West	Siraaroo	7	20	21	140
East Shoa	Adaamii Tulluu	7	20	21	140
	Fantallee	2	6	6	40
	Gimbichu	4	11	11	68
South West Shoa	Tole	7	20	21	140
West Shoa	Ambo	6	18	18	120
	Jibat	4	11	12	80
West Hararghe	Ancaar	3	7	9	60
	Burqaa Dhimtuu	6	18	15	99
	Daro Labuu	6	18	19	120
Total		**60**	**176**	**177**	**1,161**

QUANTITATIVE FIELDWORK CHALLENGES

- One major challenge was that students do not return to school and classes do not run normally for sometimes weeks into the start of term in many places. Although field work was delayed to account for this, during the initial days in field, teams did encounter cases where classes were not running as normal. If this was the case, the teams were instructed to shift their efforts and return at a later date.
- As data collection was due to start in Fantalle woreda, local tensions and fighting between clans meant that all but two of our sampled schools were closed for the foreseeable future. Therefore, the schools that could not be visited were replaced.
- In Gesala School (cluster school), the school director was on maternity leave and there were no deputies with whom the school director interview could be conducted. As per fieldwork protocol, the team attempted to interview the most senior teacher, however they refused to take part.
- In Araada School (control school), the language of instruction was Amharic after Grade 3, therefore no Grade 4 student learning assessment was conducted.
- In Baha Biftuu School (cluster school), there was strike action taking place and therefore no classes were running. The school director and some teachers were still interviewed. The school was replaced after it became clear that no classes would take place during the field work period.

- Internet connectivity proved to be a challenge with a national blackout during the initial days in field and therefore data could not be submitted by supervisors to the C4ED server in a timely manner.

QUANTITATIVE DATA ANALYSIS

All data processing, cleaning and subsequent quantitative analyses were performed by the C4ED evaluation team using Stata version 16 (Stata Corp., TX, USA). For the quantitative analysis, the unit of analysis is at the school level and treatment effects should be interpreted as average treatment effects across the school as a whole. More specifically, in a cluster school an FTT, TTS or even a teacher that received neither form of training will be considered in the same treatment group under the cluster school. This is in contrast to the qualitative analysis, which also investigated potential heterogeneous effects within schools, depending on what level of training the individual respondent reported receiving.

3.3.2 QUALITATIVE DATA COLLECTION

The qualitative data collection started with a three-day training workshop on 2nd March 2020 and was completed on 25th March 2020. The qualitative data collection was conducted by C4ED in collaboration with Reliance and five regional education experts with extensive experience in qualitative research. While six regional experts were initially recruited and attended the training, one dropped out prior to the start of fieldwork for personal reasons.

In collaboration with Reliance, C4ED led a three-day field researcher training from 2nd to 4th March 2020. The training was given by two members of the C4ED evaluation team. The training for the qualitative data collection included a thorough introduction to the AfL programme and its different facets, an in-depth discussion of all interview guidelines, as well as refreshment sessions on conducting qualitative interviews and ethical considerations in education research. Moreover, the training included a module on field protocols and an introduction to the Survey Solutions software, which was used to collect consent from participants.

A field pilot took place on 5th March 2020 in Metebaber School and Beherawi Betmengist School in Addis Ababa. During the pilot all researchers conducted at least one IDI with either a teacher or a school director. Moreover, in groups of three, all researchers conducted one FGD with parents. Issues and challenges were discussed during a debrief on 6th March 2020 and the necessary changes were made to the qualitative instruments.

Field work was launched on 8th March 2020, when the researchers travelled to their respective regions, and was completed on 25th March 2020 when all researchers returned to Addis Ababa. Two members of the C4ED evaluation team joined the Oromia field researchers for two days to support the team, after which they remotely assisted all researchers in field. The KIIs with UNICEF and ELIXIR were conducted directly by C4ED.

Table 6. Key dates in the qualitative data collection phase

Dates	Activity
02/03 – 04/03	Field Researcher Training
05/03	Field Pilot in Addis Ababa
06/03	Pilot Debrief
08/03 – 25/03	Field Work

In total the qualitative field researchers visited 14 schools across Oromia, Tigray and Benishangul-Gumuz. Additionally, they conducted interviews with non-school based officials involved in the AfL and regional key informants. A full list of respondents is included in Appendix G. Table 7 provides an overview of the qualitative data collection sample by region, zone and respondent.

Table 7. Qualitative data collection overview

Tool	Respondent	Oromia		Tigray		Benishangul-Gumuz		Total
		Ambo	Tole	Gulomekeda	Kelte Awellalo	Bambasi	Homosha	
School-Based IDIs	School Teacher	12	12	6	8	8	8	**54**
	School director	3	3	1	2	2	2	**13**
FGDs	Parents at school	1	1	1	1	1	1	**6**
Non-School-Based IDIs	Cluster Supervisor	1	0	1	0	0	1	**3**
	WEO Officials	1	0	0	1	1	0	**3**
	ToT trainer	1		1		1		**3**
	CTE instructor	1		1				**2**
KIIs	REB	1		1		1		**3**
	UNICEF							**1**
	ELIXIR							**1**
Total								**89**

QUALITATIVE FIELD WORK CHALLENGES

The first challenge faced by the research team in all regions concerned the difficulty of identifying whether cluster schools had given school or cluster training. In Tigray and Oromia, in particular, we received conflicting information that led to re-sampling having to be done while fieldwork was ongoing.

Towards the end of data collection our team also faced severe challenges due to the sudden country-wide school closure on March 17th 2020 because of the COVID-19 pandemic. Even though qualitative data collection in schools was completed in Oromia and Benishangul-Gumuz before the school closure, data collection was still ongoing in Tigray. Following discussions with Reliance and a risk assessment, we decided to cautiously continue with data collection, while continually monitoring the situation. A small number of interviews were conducted over the phone as a result.

The school closure posed a major challenge in reaching teachers and school directors. For this reason, even though the research team visited four schools in Tigray, the team was not able to interview as many teachers and school directors as planned.

QUALITATIVE DATA ANALYSIS

Full transcripts of IDIs, KIIs and FGDs were prepared following data collection by the regional experts, to fully capture what took place in the field without loss of information. Transcripts were subsequently analysed by the C4ED evaluation team by creating analytical categories and

classifying the material and exploring several cases in detail. Specifically, data was analysed at the individual respondent level and similarities and differences between regions and respondent roles were examined. The qualitative data was analysed with the software MAXQDA.

3.4 QUALITY ASSURANCE – FIELD WORK AND DATA COLLECTION PROCEDURES

All data collected was subjected to data quality assurance and checks. For quantitative data collection, the initial step for quality assurance was to program a number of consistency and validity checks into the Survey Solutions tool to avoid errors in the data and flag issues immediately. For both quantitative and qualitative data collection, alongside the training programme, supervisors, enumerators and field researchers were provided with a field manual that outlined all relevant field protocols, including sampling protocols, to follow as well as reference material for each of the instruments.

During quantitative data collection enumerators were instructed to review each completed questionnaire file prior to concluding the interview to ensure completeness and that no errors remained. In the evening, during the team debrief, the team's supervisor reviewed the completed interviews and spot checked the data entry of student assessments before submitting to the main Survey Solutions Headquarters server. The data was downloaded on a daily basis by the C4ED evaluation team and a range of automated and manual secondary data checks were conducted to identify potential errors, outliers and inconsistencies within the data, to check enumerator performance as well as to utilise a rich set of para-data to check for unusual behaviour by field teams. Any issues that arose were communicated back to the field teams for clarification and correction and a log of all changes to raw data from the field was compiled. Finally, the supervisors performed field spot checks on their enumerators and observed interviews and protocol adherence to ensure a high level of performance and consistency between their enumerators.

For the qualitative data collection progress was monitored remotely to ensure that the agreed data collection and sampling protocols were adhered to. Where issues arose, particularly related to the school closures as described above, communication between the regional experts and the C4ED evaluation team enabled issues to be collaboratively resolved. When submitted by the regional experts onto C4ED's secure platform, qualitative data audio files were checked for length and audio quality. Subsequent translations and transcripts by the regional experts were reviewed by the C4ED evaluation team by comparing the length of the transcript to the length of the audio file, and by running automated checks for transcription errors. Any inconsistencies and errors were resolved through discussion with the regional experts.

3.5 ETHICAL CONSIDERATIONS AND CLEARANCE

Throughout the evaluation the evaluation team adhered to the UNICEF's guidelines for conducting Ethical Research Involving Children (Powell et al., 2013) and Ethical Principles and Guidelines for the Protection of Human Subjects of Research as outlined in the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioural Research, 1978) and followed the three principles of: (i) Respect for Persons; (ii) Beneficence; and, (iii) Justice. All supervisors and enumerators received training on both the principles of conducting research and the practical implications of ensuring that these principles are met in practice with clear ethical protocols provided.

For all interviewees, informed consent was obtained. Informed consent includes the ethical components regarding the following: (i) transparency about the objectives and content of the study; (ii) privacy and data security; (iii) voluntary participation; (iv) the right of participants to refuse or skip any questions without incurring any consequence; and (v) a follow-up to inform or receive complaints and supply any further information about the study. For students taking part in the student assessments, it was explained that they did not have to take part in the test if they did not wish to and could leave at any time. Before commencing the assessment, the facilitator ensured that student assent had been obtained before continuing. Parents of all students that took part in the assessments were provided an information sheet explaining the study, outlining the aims and activities taking place. Contact details for the field coordinator were provided to both the parents and school directors in case parents wished to opt-out and withdraw their child's data from the study. As we relied on direct consent from school directors and teachers, and consent from school directors in loco parentis (as locating parents prior to field teams arrival was not practical due to the sampling approach), we collected only the student name, age and gender as personal information.

To conduct the data collection, C4ED obtained ethical clearance from the accredited Institutional Review Board at the Ethiopian Society of Sociologists, Social Workers and Anthropologists (ESSSWA). The approval from the Institutional Review Board at ESSSWA was obtained (Ref: ESSSWA/L/AA/062/20) and hard copies of the certificate were submitted to UNICEF. A copy of the ethical clearance approval letter is included in the Annex.

3.5 EVALUATION LIMITATIONS, CONSTRAINTS AND MITIGATION MEASURES

One limitation of the evaluation relates the quasi-experimental design and lack of randomised selection of treatment for schools. While the matching design proposed by the evaluation can account for observable differences in schools, it rests on the assumption that there are no unobservable characteristics of schools that can influence learning outcomes outside of the AfL. This is a very strong assumption, as there are many unobserved characteristics that may influence learning outcomes that cannot be controlled for by the evaluation team, such as composition and characteristics of the households of the students, including household wealth and parents' education.

Furthermore, as no baseline or midline data was available for the evaluation, we were unable to use matching in conjunction with a difference-in-difference model to account for time-invariant unobservable characteristics and increase the robustness of the matching design. Finally, as matching should be conducted with baseline characteristics, without this data we had to use ex-post covariates for the matching with the assumption that the covariates have been time-invariant since baseline. These assumptions of no unobservable differences between treatment and control schools along with time-invariant matching covariates limit any causal interpretations from the analysis. Given this limitation, we will therefore focus more on descriptive analysis alongside a qualitative data analysis to explore potential mechanisms and causal pathways. This enabled us to gain valuable insights into the programme's implementation.

Linked to limitations in quantitative design and with consideration to the ToC, we put particular emphasis on the analysis on teacher performance outcomes over the student learning outcomes. With this in mind, due to budget limitations, the study uses student assessments conducted by pen-and-paper in a group which may mean that the tests capture reading abilities rather than fully capturing ability in the subject. The rationale behind this is that compared with student learning

outcomes, there are far fewer potential unobservable confounding factors for teacher performance. In addition, we collected a richer set of data on individual teachers to include and control for in our analysis. Therefore, the effect of the AfL on teacher performance was easier to capture and has greater potential for attribution to the programme than student learning outcomes. This focus on teacher performance also follows the ToC which outlines that AfL relies on teachers changing their behaviour, in order to improve the quality of education for students which ultimately improves student learning outcomes.

Another limitation of the design concerns representativeness and external validity. The quantitative data collection was limited to the region of Oromia. Given the strong heterogeneity of regions in Ethiopia and of implementation levels across them, it is difficult to extrapolate results from the quantitative analysis to all regions where the AfL programme was implemented. We selected Oromia for two reasons. First, given the limitations of the quantitative evaluation design, a reduced quantitative sample size was chosen that is not conducive to a representative multi-region survey and has limitations for the internal validity of findings. Secondly, the AfL programme appeared to have been implemented most rigorously in Oromia, so we identified the region as an exemplar with the greatest potential to capture the effects of the programme in the quantitative analysis. Therefore, any findings from the quantitative analysis will have to be interpreted with the caveat that they relate to a high-quality implementation of the programme. To complement the quantitative research design, the qualitative research was designed to cover three regions, including Oromia, so that regional differences and implementation across more than one area could be explored. That said, the qualitative component was conducted with a small sample and was not designed to offer generalisable results that can be extrapolated to all of Ethiopia. Nevertheless, the mixed-methods analysis provides insights into the potential impact of AfL and explores the mechanisms through which the intervention works, as well as implementation bottlenecks and successes across regions.

Finally, there is no quantitative cost-effectiveness analyses that relate to the AfL programme. This was outside the scope of the ToR for the evaluation and unfeasible given resource constraints. In the qualitative study, we do address the sustainability of the programme and alternative scale-up options that may be cost-effective compared with the current design using the cascading knowledge model.

CHAPTER 4: EVALUATION FINDINGS AND ANALYSIS

In this chapter we report the evaluation findings, based on both the quantitative and qualitative analysis. For ease of reading the narrative, we have rearranged the order of the evaluation questions in presenting our findings. The mapping of sections to the main evaluation questions in this chapter is presented in Table 8 below. Evaluation question 6 is addressed in the next chapter on lessons and recommendations.

In the first section, we report on the findings relating to the quality and relevance of the AfL programme interventions and their gender equity. In the second section, we look at the delivery of the programme interventions and how they differed across geographical areas and levels of implementation. In the third section we look the effectiveness of the programme at bringing changes in teacher practice, parental engagement, and student outcomes. The fourth and fifth sections deal with policy, both centralised and decentralised, and the sustainability and future directions of the programme.

Table 8. Mapping of sections to main evaluation questions

Section	Main evaluation question	Evaluation Critera
4.1	EQ1. What is the quality and relevance (including gender equity) of programme interventions (materials, modules, tools and training)? How can they be improved for future use?	Relevance & gender equity
4.2	EQ5. What were the most efficient and effective approaches used by regions, woredas, schools or teachers to bring about change? What worked, what did not work, and why?	Efficiency and effectiveness
4.3	EQ3. To what extent have programme inputs made a difference to teachers' ability to use continuous assessment techniques? EQ4. To what extent are changes in teacher practice attributable to the AfL project activities?	Effectiveness, impact & gender equity
4.4	EQ2. What can the programme do at both policy level and decentralised structure levels to improve programme interventions and impact and promote sustainability and the scale-up of promising practices? EQ7. Have any changes been achieved in relation to policy, practice, attitudes of practitioners and policy makers?	Sustainability

4.1 QUALITY AND RELEVANCE OF PROGRAMME INTERVENTIONS

The AfL programme consists of a package of interventions that included the development of a training programme using the cascading knowledge model, as discussed in Chapter 1, alongside a guide for training facilitators and workbook for participants. In addition to the training, a 'AfL package' for teachers was developed that included handbooks for each focal subject in Grades 1-4 that included items such as a question bank, visual learning ladders, and templates for feedback and student progress recording. Finally, a continuous assessment manual for teachers was developed as a reference tool for the techniques included in AfL. The 'AfL package' materials described above were contextualised and translated into appropriate languages. In this section we discuss the findings on these inputs.

We present the findings relating to the cascading knowledge model in Sections 4.1.1 and 4.1.2. Starting from the highest level, we first discuss the ToT training and move down the cascading

knowledge model onto ToT training (for selected teachers, school directors and other non-school-based staff) and then finally onto the lowest level of the model, the cluster and school training.

4.1.1 TOT TRAINING

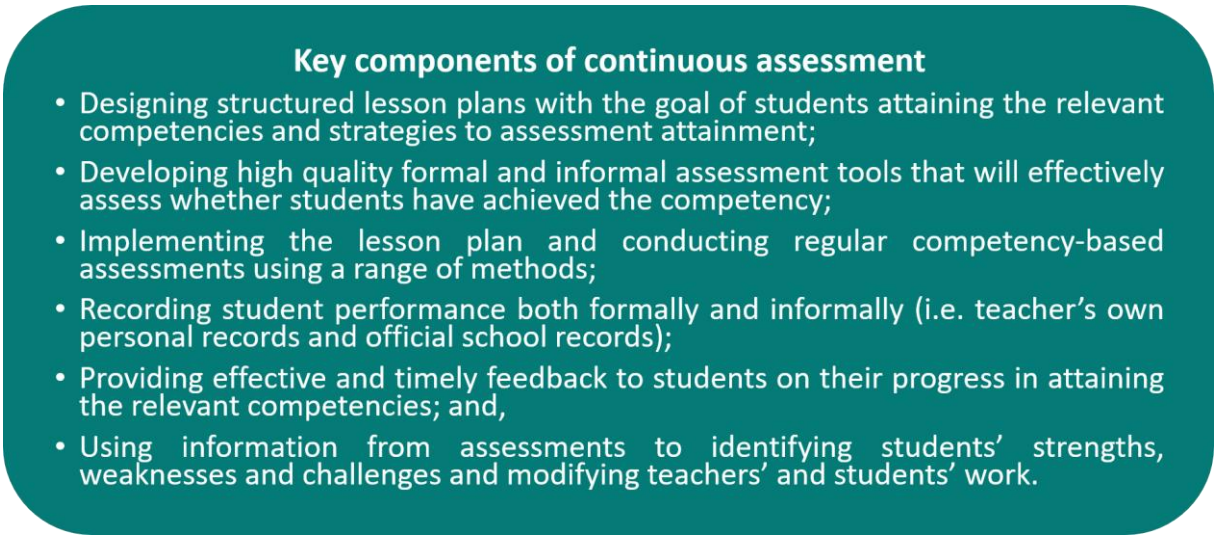
For every cluster school that participated in the AfL programme, one teacher per focal subject (up to four per school) was invited to participate in a ToT training that was facilitated by trained staff from local CTEs. According to the programme documents we reviewed, the ToT training comprised of two five-day trainings and a three-day refresher training and was conducted at the woreda level. In addition to the teachers selected to attend the ToT training, school directors of participating schools, WEO curriculum/assessment experts and cluster supervisors overseeing the schools were all invited to attend. According to the AfL dashboard developed by the implementing partner ELIXIR, between September 2018 and June 2019, 330 teachers and 110 school directors across 110 schools in Oromia (averaging three teachers per school) were supported in AfL ToT training sessions. In addition, 22 experts and supervisors from WEOs and REBs participated, though this was lower than expected due to the focus in Oromia being shifted onto the integration of the AfL programme into CTE pre-service courses.

In the quantitative survey in Oromia, all teachers and school directors (including teachers from satellite and control schools in case of spillovers due to teacher turnover) were asked whether they ever attended the ToT training. If they reported that they had attended, a set of questions relating to their experiences of the training were asked. In addition to asking teachers themselves, school directors were asked to indicate if individual teachers in their school had attended the ToT training. In total, 53 teachers and 24 school directors (all from cluster schools) reported that they had attended the training.

TRAINING CONTENT QUALITY

Using the definition of continuous assessment as a set of various components, as discussed in Chapter 1 and outlined in Figure 11, we were interested in understanding how effective the integration of these components into the AfL training programme was.

Figure 11. Key components of continuous assessment



To measure how effective the integration of these components into the AfL training programme was, we first asked participants (FTT and school directors) for their perceptions toward the quality of training they received by asking them whether they felt they had received enough training in each of the components, with the student performance recording combined with the use of information component. The respondents were read a positive statement that they had received enough training in each of the components and then asked whether they agreed with the statement and answer on a five-point Likert scale ranging from 5 (strongly agree) to 1 (strongly disagree). The results are shown in Table 9.

Table 9. Perceptions of ToT training quality on key components

Training Area	Strongly Agree (5) (column %)	Somewhat Agree (4) (%)	Neutral (3) (%)	Somewhat Disagree (2) (%)	Strongly Disagree (1) (%)	Average agreement (1-5)
Developing structured lesson plans	37.7	39	14.3	9.1	0	**4.1**
Developing high quality questions of my own for students	31.2	45.5	14.3	7.8	1.3	**4.0**
Using a wide range of assessment methods	29.9	35.1	22.1	10.4	2.6	**3.8**
Providing students with effective feedback	21.2	35.1	15.3	18.2	1.3	**3.8**
Using information from assessment to alter my teaching practices	24.7	45.5	18.2	10.4	1.3	**3.8**

N=77. Due to rounding, numbers and percentages in this report may not add up precisely to the totals indicated.

Overall, AfL FTTs and participant school directors perceived that the training covered content on each of the key components to a high level and were satisfied with the quality, with over half expressing agreement (somewhat or strongly) for each of the components. The components that respondents had the highest satisfaction levels in the training were developing structured lesson plans and developing high quality questions for their students as assessment tools.

In the qualitative interviews, we did not ask respondents across regions about their perceptions on specific training components. Nevertheless, in general many respondents' overall impressions of the ToT training were positive, with many interviewees across regions noting that the training content was relevant for their work:

> *It is very helpful to my day-to-day activities. In fact we took as a course while we were in college but we didn't use it for the teaching-learning process but after we took the training we understand how to prepare questions.* (Tigray15)

> *The content of the AfL training materials was delightful. It will change the results and behaviour of students if it is given for all teachers.* (BeGu13)

TOT TRAINER AND FACILITATION QUALITY

In addition to the content of the training, we are also interested in understanding how the training was facilitated and the quality of the ToT trainers providing the training. To measure this, we asked AfL FTTs and participant school directors to provide their perceptions of the facilitation of the training, firstly whether the training was interactive. Training that is interactive is more effective for participants to learn through inducing active learning (Kinzie, 1990) as well as to encourage participation and a provide wide range of views and opinions. The respondents were asked whether they agreed with the statement "The Training was interactive (i.e. The trainer encouraged participant involvement and included questions to participants)" and answer on a five-point Likert scale ranging from 5 (strongly agree) to 1 (strongly disagree).

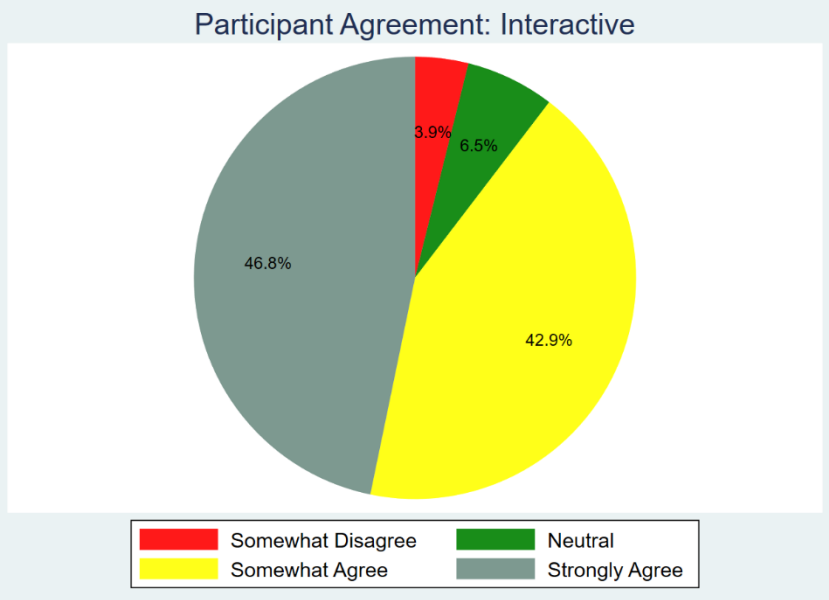
The results are shown in Figure 12. Almost 90% of participants agreed to some extent that the trainer provided a training programme that was interactive and fostered participation and questions from those attending.

Indeed, this finding was supported by the qualitative analysis, where some teachers and school directors who had attended ToT training noted that the trainers had used interactive and participatory methods:

> *[The trainers'] approach was also interesting, they explored the previous experience we have, identified the areas of the gap we have to fill and the like. Overall, it was interesting.* (Oromia25)

> *The trainer was very good in the process of the training; he had a good approach to the trainees [...] and he was participating us.* (Tigray9)

Figure 12. Participant perception on training interactivity (ToT)



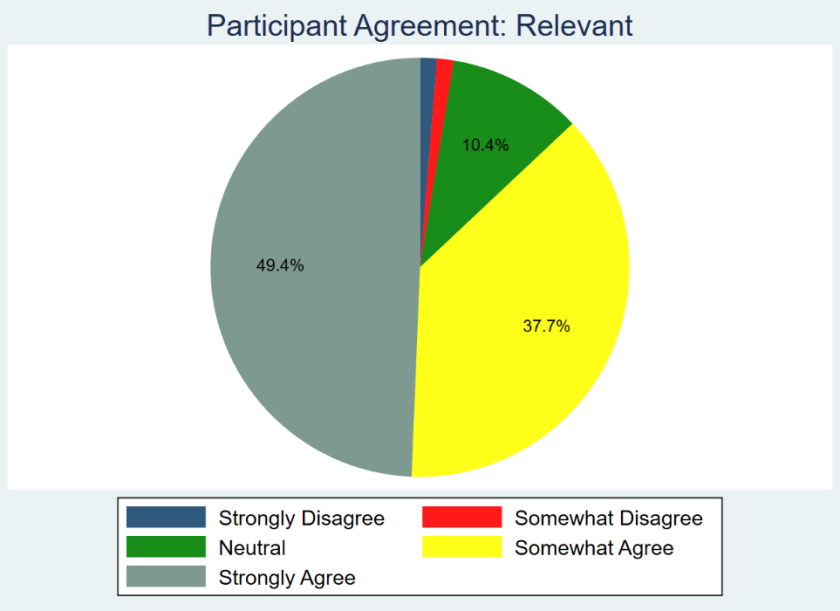
The second way in which the quality of the facilitation was considered through participant perceptions was whether they believe the content was taught in a way that was relevant to participants. It can often be the case that in teacher training, the focus on theoretical pedagogical concepts can mean that it is difficult to translate training content into practical and actionable changes that a teacher can implement in the classroom. Respondents were asked whether they

agreed with the statement "The Trainer used clear and relevant examples when discussing AfL techniques". Again, as shown in Figure 13, the response was clearly positive with 87% of respondents agreeing to some extent and less than 3% disagreeing. This suggests that the training included context-relevant information that resonated with teachers in Ethiopia.

However, there was emerging evidence from the qualitative analysis that training materials could still benefit from further tailoring to the national context. For instance, some school staff noted that the issue of practically implementing AfL techniques in large classes, with poor attendance, was not resolved during training:

> *When trainees ask questions and raise some issue about the anticipated challenges [such as large class sizes and absenteeism] that may hinder AfL implementations in the future, trainer's response was, "You have to do it". […] The questions were not raised due to lack of understanding, but it was from the fear and concern we have on the practicality of the programme. The AfL idea was imported from other countries, but the level and setup of our country is not similar with that of other.* (BeGu10)

Figure 13. Participant perception on training relevance (ToT)



Note: 1.3% of participants strongly disagreed and 1.3% of somewhat agreed

GENDER EQUALITY

We can also evaluate the programme through the lens of whether it was implemented fairly and to do so, we examine potential issues in gender equity of the programme. The first possible barrier for female teachers in attending the training was that education administrators and/or school directors may consciously or subconsciously select male teachers ahead of their female counterparts. From the quantitative sample, there appeared to be no gap between the proportion of male teachers that attended (51%) and the proportion of female teachers that attended (48%). An additional way to examine whether females were less likely to attend the ToT training is to look at the relationship between school size and the rate of females attending the training. Larger schools that have more teachers will have greater competition for places in attending the ToT training and more open to potential gender discrimination. However, the data shows no evidence

that female attendance at the ToT training decreased at larger schools, while gender composition of teachers remained constant.

The second source of potential issues in gender equity were that barriers to attendance, such as childcare, disproportionately affected female teachers. With regards to the barriers to attendance, no teachers indicated childcare being an issue to attending. Indeed, in the qualitative interviews, there was anecdotal evidence that concessions had been made in order to remove the barrier of childcare. For instance, a ToT trainer in Tigray described providing additional sessions for participants who needed to leave the training to breastfeed their child. A teacher in Benishangul-Gumuz who brought her child to the ToT training reported that a babysitter was provided so that she could "attend the training freely" (BeGu3).

It is possible that some gendered barriers were nested within the survey response option of teachers not being able to attend training due to private commitments, and it should be noted that the qualitative interviews did not try to capture gendered barriers in a generalisable manner. However, based on the data available, there did not appear to be any evidence of greater barriers to either being selected to attend the training or actually attending the training for female teachers.

4.1.2 CLUSTER AND SCHOOL TRAINING

The final level of the cascading knowledge model is the cluster and school training on AfL. The plan was for those who attended the ToT training (FTT, school directors of cluster schools, cluster supervisors, WEO officials) to organise and support the delivery of the AfL training programme to teachers both within cluster schools who did not attend the ToT training and to staff in nearby satellite schools within the same cluster. Similarly to the ToT training, in the quantitative survey for teachers and school directors (including teachers in control schools in case of spillovers due to teacher turnover) were asked whether they had ever attended a cluster or school training relating to the AfL programme. If they reported that they had attended, a set of questions relating to their experiences of the training was asked. Again, teachers who were interviewed self-reported attendance and school directors were asked to indicate whether individual teachers in their school had attended the cluster or school training. In total, 25 teachers and three school directors who were surveyed reported that they had attended the training.

In an ideal implementation of the cascading knowledge model, the knowledge should seamlessly cascade from the top to the bottom without any loss in quality and be as effective as the ToT training in contributing to teacher, and ultimately, student outcomes. However, in practice, if not properly supported and monitored, the model has a significant risk of dilution or distortion at each level that leads to a quality gap with the ToT training programme, as has been observed in other in-service education training programmes using a cascading knowledge model (Solomon & Tresman, 1999; Kennedy, 2005).

TRAINING CONTENT QUALITY AND QUALITY GAP WITH TOT

To assess the effectiveness of the cluster and school training alongside understanding the magnitude of any quality gap with the ToT and the magnitude of any quality gap with the ToT training programme, AfL TTSs and school directors participating in cluster and school training were asked the same questions as AfL FTT and school directors who attended the ToT training. In terms of a gap in the quality of content, the questions again asked respondents on their

perceptions on the quality of training in relation to the key components of continuous assessment as outlined in the previous section (see Figure 11 on page 35). The results presented in Appendix I include the average scores from the ToT training participants for comparison.

Across the key continuous assessment components, the responses indicate that on average, respondents felt that they received sufficient training at the school and cluster training sessions, with over half of respondents agreeing (either strongly or somewhat) with the positive statement that they had received enough training in the component. However, in terms of a quality gap, the rate of agreements (either strongly or somewhat) is lower in each component when compared with the ToT training participants' perceptions. The average Likert scale score is also lower for each area with a 'quality gap' of between 0.2-0.6 in the Likert scale.

TRAINER AND FACILITATION QUALITY AND QUALITY GAP WITH TOT

One other potential source for a gap in quality between the ToT training and the cluster and school training is the facilitation of the training. For the ToT training, the facilitator was either from the programme implementing partner, ELIXIR, or a member of staff from a CTE who had received training directly from ELIXIR. For the cluster and school training, although it was expected for there to remain a link with the ToT trainer (discussed later in the section), it was expected that the training would be facilitated by those who participated in the ToT training, that is the AfL FTTs, cluster school directors and cluster supervisors. To understand any gap in quality in relation to the facilitation of the cluster and school training, AfL TTSs and participant school directors, similarly to those who attended the ToT training, were also asked to provide their perceptions of the facilitation of the training on both whether it was interactive and relevant.

The results are shown in Appendix I and, again, include the average scores from the ToT training participants for comparison. The cluster and school training participants indicated that they agree (4 on the Likert scale, on average) that the training was both interactive and relevant. However, for both elements, the agreement level was lower than for ToT participants (0.3 on the Likert scale in both cases).

While the findings suggest a gap in quality between the ToT and the cluster and school training in terms of content and the facilitation of the training, the size of the gap is relatively small given the potential for quality dilution in a cascading knowledge model. Possible reasons for this include trainer quality and length of the training which are examined below. One limitation of using comparisons between perceptions of the cluster and school training participants and the ToT training participants is that the expectations of quality and the participants themselves may not be similar. That is, participants of a formal ToT training at the woreda/CTE level may have higher expectations of a training than someone attending short, ad-hoc internal training at their own school and may adjust their expectations accordingly. Given the findings relating to the implementation challenges discussed below, we expect the results outlined above to understate the gap in quality between the cluster and school training and the ToT training. Nevertheless, it appeared that participants were satisfied with the cluster and school training on the whole, despite any quality gap.

4.1.3 AFL PACKAGE MATERIALS SUITABILITY

MATERIAL ACCESSIBILITY

In terms of accessibility, several ToT-trained informants noted that the training material had been provided in a language appropriate for their context (for instance, Afan Oromo in Oromia, Amharic and English in Benishangul-Gumuz). A small number of informants also positively noted that the language of the material was changed following their feedback. However, some interviewees in Oromia also noted that some of the materials were dense and difficult to understand. A ToT trainer from Benishangul-Gumuz reflected that perceptions on the accessibility of the AfL handbook and the workbooks differ:

> *The training material that incorporates the facilitator's note [the AfL handbook] seems difficult to understand and is very bulky. The workbook is smaller and easy to understand. However, we have shown them on how to use it while they cascade it in their school. (BeGu23)*

Language aside, some interviewees noted that the continuous assessment formats were complicated and difficult to use in the classroom. Several of these informants reported that they or their school had adapted the formats to make them more accessible and user-friendly. For instance, an FTT in Benishangul-Gumuz noted that:

> *We have developed our own formats by minimising the content. Some of the schools around are using the formats we developed here. We are discussing with teachers to decide which part has to be reduced and which one not. We do not have the mandates to do so, but since the content on the format are many, and redundant, we are contextualising it as per the nature of education given in our area.* (BeGu13)

Similar, a cluster-trained teacher in Oromia reported being provided with and using a modified format:

> *We took [the format] from the school. It is also found in the module given during training. But it is somewhat modified. The school gave us the form, I prepare daily lesson plan using this.* (Oromia1)

LANGUAGE BARRIERS

Despite efforts by ELIXIR to provide the training materials in 15 different instructional languages – among them English, Amharic, Afan Oromo, Somali and Tigrinya – some teachers reported not having materials in the correct language. This issue was raised particularly in interviews with informants in Benishangul-Gumuz (rather than in Tigray or Oromia). FTTs from Benishangul-Gumuz for instance reported:

> *The training was given in English, but there were translators with our mother tongue languages and it was not difficult to understand. What make the issue difficult was that our mother tongue was left behind and formats were not developed with it.* (BeGu14)

> *The training […] was given in English, but we were asking them to give us the training in Amharic especially for teachers teaching from Grade 1-4. As per our request, the material was changed into Amharic in the second round training, but they didn't give it to us. Anyway, I have started to apply the lesson plan in my teaching.* (BeGu17)

When asked whether he would continue using AfL tools and materials in the long run, an FTT stressed the importance of materials in the appropriate language:

> *If I got the module translated to Amharic and other materials fulfilled, I am really happy to continue implementing it.* (BeGu17)

Moreover, a ToT-school director from Benishangul-Gumuz reported that overall, the implementation of AfL was going well in his school. However, he noted that that they only had material in Amharic and that they had to translate it themselves into one of the mother tongue languages used in the school. Additionally, he remarked that the school was not currently implementing AfL in the second mother tongue (Maonga) due to the low ability of teachers. According to ELIXIR there are 42 different instructional languages in Ethiopia, presenting a challenge when scaling up the programme, as materials must be translated.

However, irrespective of AfL, teachers in Benishangul-Gumuz reported facing language barriers in general. Some teachers highlighted that they did not speak the mother tongue of their students and thus could not give feedback to students directly. More generally, they faced challenges while teaching and often had to rely on students who spoke the language of the teacher and the local language to translate during class. This also hampered not just the implementation of AfL but teaching in general.

4.1.4 RELEVANCE OF THE AFL PROGRAMME TO THE EDUCATION CONTEXT

CLASS SIZE

A challenge in the implementation of AfL techniques in the classroom is the student teacher ratio. The schools included in the quantitative survey in Oromia had a median class size of 64 students, and 55% of teachers reported that class size was a 'big challenge' for them in being able to implement continuous assessment. In addition, school directors were asked what the biggest single challenge in implementing the AfL programme into their school was and 60% responded with class size.

In the qualitative interviews, some teachers in Benishangul-Gumuz and Oromia reported student teacher ratios exceeding 100. Given these circumstances and the limited time of one period many informants argued that it was challenging to implement AfL techniques such as continuous assessment or recording of student progress record:

> *Number of students in class is not manageable. When we have to do recording, for example in Grade 3, we have 89 students, to assess and record all this much student is very difficult. Maybe I can try to record for some 30-40 students if didn't reach their level of competence. This means tomorrow before I proceed to the next lesson, I have to go back and teach those 30-40 students again the previous lesson. If I did so, how can I manage my time to complete the course within this academic calendar? This is our challenge.* (BeGu6)

> *It becomes difficult to apply with a large number of students in the class. For instance there are 105 students in Grade 3. Imagine how much it is difficult to address all these students within one period, 40 minutes.* (Oromia16)

Given that the AfL manual states that teachers should ensure that each student understands the taught subject before moving on, several teachers, especially in Oromia, raised the issue of not

being able to individually follow up on students in large classes. For example, a ToT-trained teacher from Oromia reported that with a class of 70 students, it was very difficult to follow and monitor each student (Oromia30).

It was also asserted that AfL training does not adequately address techniques on how to address large class sizes. Like other informants in Oromia and Benishangul-Gumuz, a ToT-trained school director reflected that:

> *The trainings were provided taking class size of 50 students as a standard. Yet, this class size standard diverges from concrete realty of many schools. For example, many of our classes are a 100 and something students which is much larger than the standard class. We raised this challenge in the trainings. This is still unresolved challenge. Even if we still have the challenge, we have not abandoned the programme at all.* (Oromia7)

In addition, this school director argued that the assessment methods for large class size presented in the training were not suitable for young pupils in Grades 1-4:

> *In the training we learned about six assessment techniques of large class size. Yet, I personally think that the techniques are not effective to enable students because students, who are in Grades 1-4 in most cases, are relatively younger.* (Oromia7)

When training attendees in Benishangul-Gumuz approached their ToT trainer with the issue of large class sizes, he suggested that they divide their class into different groups and to concentrate on one group at time. In fact, an FTT from Benishangul-Gumuz reported she is using such an approach – however, with 80 students in one class, it takes her between two weeks and a month to assess the entire class. A cluster-trained teacher from Oromia reported a further adaptation of this method, by grouping students into different levels and focussing primarily on low-performing students:

> *Mainly I spent my time on group C and D students, the reason I do this is, for example I have 64 student in one section though it's difficult to manage and giving support for all student within 45 minutes. (*Oromia28)

Notably, only respondents in Benishangul-Gumuz and in Oromia reported class sizes to be an issue. In the quantitative survey in Oromia, teachers who attended the ToT training reported slightly higher rates of indicating class size being a problem for implementing continuous assessment. Qualitative respondents in Tigray did not spontaneously raise class size as a challenge in implementation.

TIME CONSTRAINTS

Closely related to the issue of class size is time constraints. Several school-level and higher-level stakeholders argued that teachers lacked time to implement the AfL properly. In the quantitative survey, 41% of teachers in Oromia indicated that time constraints were a 'big challenge' in implementing continuous assessment in their classes.

Some teachers argued that the length of lessons per se is not long enough, as the AfL requires teachers conduct additional tasks such as identifying the level of students and recording their level. They conveyed that this is not possible within a 40-minute period, especially if class sizes are large (see 'Class size' above). Furthermore, a single period is often not enough to implement

more time intensive teaching methods, as teachers from Oromia and Benishangul-Gumuz highlighted:

> *Using true/false and cards to assess them may not take up several minutes but checking their exercise books and making them practice in class cannot be conducted within 40 minutes.* (Oromia6)

> *I did not apply jigsaw and lecturing. It is group work. The time given for one session is 40 minutes, which is not enough to apply these methods. Both methods are applied as a group teaching method, which takes time.* (Oromia25)

> *If the education sector actors agree on time increment from federal to woreda level, it will be easy for every teacher to assess every individual student. […] teachers will get time to touch on or address all important packages. When we teach in 40 minutes time, we are reducing many things due to time constraint.* (BeGu18)

To be able to implement AfL methods properly, teachers suggested that class time be increased, for instance to 50 to 90 minutes .

Another issue raised was that the general workload of teachers is increased when applying AfL methods. In particular, teachers in Oromia and Benishangul-Gumuz noted that preparing daily instead of weekly lesson plans was very time intensive. One FTT from Benishangul-Gumuz complained:

> *Since it needs to prepare plan daily it is going to be 24hour work for teacher. I shout there the whole day and will start to plan again for tomorrow. There will not be rest at all.* (BeGu3)

In that regard she wished to get further training on time management. Other teachers argued that due to teacher shortages they were already expected to teach around 30 hours per week, making it difficult to implement daily lesson planning. Moreover, a few teachers from Benishangul-Gumuz also pointed out that besides teaching, they have other duties at school such as Continuous Professional Development or taking care of clubs.

STUDENT ABSENTEEISM AND DISCIPLINE

Teachers, especially in Benishangul-Gumuz, complained about problems with student attendance. Absenteeism was one of the common reasons that teachers might initiate meetings with parents. Absenteeism made it increasingly difficult for teachers to bring students to the same level of understanding before moving on to the next topic. According to these teachers, students attending classes irregularly impeded the teaching-learning process and made it challenging for teachers in general to make sure that students were learning.

A few respondents in the Benishangul-Gumuz also reported that they faced problems with the discipline of students, making it hard for the teachers to implement AfL methods in their daily teaching. For instance, one school director from Benishangul-Gumuz reported that students removed teaching aids from the walls when teachers were not around.

RESOURCES AT SCHOOL'S DISPOSAL

A further constraint that hindered proper use of AfL in schools was the lack of resources in schools. In the quantitative survey in Oromia, 46% of teachers indicated that resources at school were a 'big challenge' in implementing continuous assessment in their teaching.

When asked the most important type of resources are necessary to implement continuous assessment in their classrooms, teachers identified basic writing equipment such as pens and paper as the most important resource required. Similarly, in the qualitative data collection, teachers and school directors in Benishangul-Gumuz and Oromia indicated that they lacked paper and other stationery to be able to prepare and use daily, rather than weekly, lesson plans. One ToT-trained school director from Benishangul-Gumuz argued that the programme required a lot of resources, making it challenging for poorer schools to implement:

> *It requires and consumes paper. It needs at least 10 pieces of papers for one period and one subject. Because, it contains lesson plan, recording, and teaching aid and references. Therefore, the programme is not applicable for poor school like us. It is better if implemented in rich schools that have budget and resource. Otherwise, they have to provide all necessary inputs if implementation is needed.* (BeGu5)

Another ToT-trained school director from Benishangul-Gumuz argued that financial resources at the school were very limited due to its small size and that there was neither a computer nor enough paper to print materials. Given their small budget of 10,000 Birr, the school director reported that they could only implement AfL tools because they were "actively working and requesting assistance from partners" (BeGu13). The school director further added that this was only possible because they were convinced of the programme and thus willing to actively support it.

Despite being convinced that daily preparation of lesson plans was beneficial to students, a ToT-trained school director from Oromia reported that at his school, teachers had gone back to preparing weekly lesson plans due to shortage of resources.

Furthermore, different teachers in both Oromia and Benishangul-Gumuz reported that the implementation of AfL was hampered by the lack of adequate classrooms and too little classroom equipment such as tables and chairs. For example, a ToT-trained school director from Benishangul-Gumuz argued:

> *The students' chairs are not suitable to use for group work. Chairs are made of bamboo sticks and fixed on the floor from wall to wall. Chairs are immobile and uncomfortable to use for group work. The setup looks like a meeting hall, not a classroom.* (BeGu5)

4.2 EFFICIENCY AND EFFECTIVENESS OF DELIVERY OF THE PROGRAMME

To understand the efficiency and effectiveness of the delivery of the AfL programme, in this section, we initially focus the delivery of the programme in Oromia and the results of the quantitative results in terms of the attendance, and length of the ToT and the cluster and school trainings alongside the distribution of package materials. We then examine the regional difference in how the programme was delivered and the successes and challenge of the various modes of implementation.

4.2.1 DELIVERY OF THE TOT TRAINING

TOT TRAINING ATTENDANCE

To understand the efficiency of implementation of the ToT training alongside understanding levels of intensity of receiving the treatment AfL programme, we are interested in measuring how many ToT sessions that AfL FTTs and school directors attended. As discussed previously, the ToT training programme comprised of two-five days sessions and a three-day refresher session. Respondents who reported that they had attended a ToT training session were asked to report how many AfL ToT training sessions they had attended. Just over half (52%) of respondents indicated that they only attended one session and only 13% indicated that they had attended all three sessions of the programme.

Respondents who indicated that they did not attend three training sessions were asked to report the reasons why they did not attend all three sessions. The majority of respondents (55%) indicated that the reason for not attending all sessions was that they were unaware that there were further sessions. There are some possible explanations for this finding, firstly that teacher turnover in Ethiopia, and indeed in this sample, is high and many respondents may have only arrived recently and secondly, school policy of rotation between teachers for who attends the training. The issue with FTTs not regularly attending all three ToT training session is discussed in more detail in Chapter 4.2.8 Other reasons why FTTs reported as the reason that they did not attend included private or school commitments (18%) or that they were not compensated for attending the training (6%).

LENGTH OF TOT TRAINING

A second way, aside from number of sessions attended, to understand how efficient the ToT training programme was is to examine whether the total length of training was sufficient to cover all the materials. The total length of the ToT training programme was 13 days (two five-day sessions and one three-day session). To measure whether the training programme was long enough, we asked survey participants for their perception on whether the training they received was long enough to cover all of the contents and materials. Only one-third of participants felt that the length of training was sufficient.

While this is useful in providing information on participants' perceptions on training they *actually* received, as we identified previously the majority had not received the full 13 days. It is thus more challenging to extrapolate the finding to the ToT training programme as a whole. Although the qualitative analysis found that in Oromia in particular, many informants felt that the training length was insufficient, it may have been that they had only attended one session. In addition, while a small number of FTTs complained about their training only lasting three days, it is plausible that these participants attended the refresher instead of a basic training session.

Further quantitative analysis was conducted to check if satisfaction with the length of the training was related to the number of training sessions attended. However, the rate of respondents reporting that the length was sufficient did not change as they attended more sessions. This finding indicates that the reason for the dissatisfaction with the training length cannot be solely attributed to not attending all the sessions, and supports the assertion that the ToT training programme was perceived to be insufficient in length to cover all the materials.

4.2.2 DELIVERY OF THE CASCADING KNOWLEDGE MODEL

CLUSTER/SCHOOL TRAINING ATTENDANCE

As a measure of efficiency of the cascading knowledge model, we examine the attendance and implementation of the cluster and school training to identify bottlenecks and also to get an understanding of the intensity of the AfL training participants received. Table 10 and Table 11 show the level of attendance at both the ToT and the cluster and school training by the teachers and school directors surveyed in Oromia. Table 11 outlines attendance at training through teacher self-report and teacher attendance as reported by the school director[6]. It is clear from the findings that the satellite schools in the sample, almost universally, did not report attending any cluster training, with only three (6%) teachers at satellite schools indicating that they themselves did attend or were even invited to one. In addition, no school directors at satellite schools reported being invited or attending a training on AfL within their cluster.

Table 10. Total sample of participating teachers for ToT and cluster/school training

Treatment group	Source	Attended ToT training	Attended cluster/school training	Attended no training
Cluster school	Teachers Interviewed	53	25	16
	School director Reported	70	34	38
Satellite school	Teachers Interviewed	0	3	42
	School director Reported	0	1	71
Control school	Teachers Interviewed	0	0	42
	School director Reported	0	0	70

Table 11. Total sample of participating school directors for ToT and cluster/school training

Treatment group	Attended ToT Training	Attended cluster/school training	Attended no training
Cluster school	24	0	6
Satellite school	0	0	15
Control school	0	0	15

When school directors were asked if their school had ever provided cluster training on AfL, 22 school directors (71%) reported that they did not. This, along with the findings above, indicate that the lowest level of the cascading knowledge model, specifically the cluster training, was not successful in the dissemination of the AfL programme to satellite schools. That this assumption, a key part of the ToC, appears not to hold in our sample, also has consequences for our quantitative evaluation design and the analysis will need to be interpreted as an intention-to-treat analysis. This is likely to underestimate results of the semi-treatment in comparison to the control schools. Despite reported lack of cluster training, satellite schools may still have benefitted from more informal dissemination of the AfL programme.

[6] Note that school directors reported on five randomly selected teachers at the school who may not have been the same as teachers participated in the survey.

When examining the school training, that is the training held internally in cluster schools for staff who did not attend the ToT training, 56% of school directors reported that this training had taken place. When interpreting this result, it should be noted that in many cases where no school training was held the school had five or fewer teachers in Grades 1-4 and most, if not all, had attended the ToT training programme and thus had no need for a school training. The evidence indicates that while there may still be bottlenecks in the cascading knowledge model even within treated schools themselves, this aspect was more successful than disseminating to other schools in the cluster.

Issues with the cascading model also emerged in Tigray and Benishangul-Gumuz during the process of sampling schools for qualitative data collection. In Tigray, we targeted satellite schools that had supposedly received cluster training, and therefore did not expect reports from informants that such training had not taken place. Nevertheless, upon our visit to some satellite schools, it emerged that none of the teachers had obtained cluster training. Anecdotally, a school director in Tigray reported that the supervisor and the FTTs had forgotten about AfL and only started planning a training after we called the school to inform them about our data collection visit – though it should be noted that this emerged in only one of the 14 schools visited for the qualitative data collection. In Benishangul-Gumuz, no cluster school directors we contacted reported that cluster or school training had been given – although this information was contradicted by the qualitative data, highlighting that closer monitoring could have been needed in the region.

TRAINING SUPPORT

Potential challenges in an efficient cascading knowledge model at the cluster and school training level are the resources and support required to organise and provide the training. According to the programme documents, it was the responsibility of the WEO, cluster supervisors and the cluster schools themselves to organise and provide the cluster or school training. To understand the support, or lack thereof, provided to cluster schools in implementing this training, we asked school directors at cluster schools that had that provided either cluster or school training about the sources of support they had received in setting it up. Around three-quarters (74%) reported that they had received support from the cluster supervisor in setting up these sessions. Just under two-thirds (63%) of the school directors received support from education officials at either the WEO or REB level in providing the school or cluster training. These findings imply that whilst support was provided, it was not universal and may have led to challenges in implementing the cascading model. The link between the ToT trainers and cluster schools appears to have been effectively sustained outside of the ToT training, with 84% of school directors at cluster schools that provided either cluster or school training receiving direct support from the ToT trainer in implementing further training.

Table 12. Support in Implementation of cluster/school training by stakeholder

Stakeholder	Supported the cluster school in implementing cluster or school training		
	Yes	No	Total
Cluster Supervisor (row %)	14 (73.7)	5 (26.3)	19
REB / WEO	12 (63.2)	7 (36.8)	19
ToT Trainer	16 (84.2)	3 (15.8)	19
Total	72	70	10

LENGTH OF CLUSTER/SCHOOL TRAINING

In the above discussion we examined whether there was a gap in quality between the ToT training and the cluster and school training through perception. Though there appears to have been a gap, due to expectation differences in the training it may be a biased measure of difference in quality. A more objective measure of quality is to compare the intensity of the cluster and school training programme compared to the ToT training programme. The cluster and school training should have been modelled on the ToT structure (i.e. two five-day training sessions and an additional three-day refresher training).

To provide a measure of length, school directors who reported providing cluster or school training were asked what the typical length, in days, of a full training session was. In practice, there appeared to be no consistent implementation of the cluster and school training across schools, with training sessions lasting between one and eight days with over two-thirds (68%) lasting two days or fewer.

To calculate the total average intensity of training AfL TTSs received from attending cluster and school training, all AfL TTSs were asked how many training sessions they had attended. Teachers attended up to three cycles, with the overwhelming majority attending one or two. With this information we can calculate a rough expectation of the average total days training an average AfL TTS underwent. AfL TTSs completed an average on 1.5 training sessions that lasted, on average, 2.8 days, therefore meaning an average training intensity of 4.2 days compared to 7.4 days for AfL FTT. This is clear evidence of a quality gap between the ToT and the cluster and school training as TTSs receive far less training. Several respondents in the qualitative data collection also criticised the cascaded school and cluster training for being too short; for instance, one school-trained teacher said:

> *Due to shortage of training time it's difficult to cover all the content easily and deeply. To cover the whole content of training manual needs much time.* (Oromia29)

4.2.3 PROGRAMME MATERIAL DELIVERY

As part of the ToT, cluster and school training, participants should have received a hard copy and if possible, a soft copy of the AfL handbook, as well as subject-specific AfL workbooks for Mother Tongue, English, Mathematics and Environmental Science. In this section we present findings

from the qualitative and quantitative analysis on the relevance, availability and accessibility of the AfL specific training material.

Table 13 presents descriptive statistics on the availability of materials at ToT training in Oromia from our quantitative survey. Out of the 53 teachers who attended ToT training in Oromia, 83% received the hard copy of the AfL handbook, while only 25% of participants obtained the soft copy. Note, however, that teachers might not have had the necessary technical equipment such as flash drives or computers to receive the soft copy, as an FTT from Oromia pointed out in one of the qualitative interviews. The data further suggests that there was a shortage of workbooks at ToT training in Oromia, as only every other teacher obtained a workbook for Mathematics and Environmental Science. The availability of Afan Oromo workbooks appeared to be considerably higher.

Table 13. Materials received at ToT trainings in Oromia

AfL package material received	Rate received by participant (%)	Std. Deviations	Observations
AfL Handbook (Hard Copy)	83	0.38	53
AfL Handbook (Soft Copy)	25	0.43	53
Workbook (Afan Oromo)	79	0.41	53
Workbook (English)	6	0.49	53
Workbook (Mathematics)	49	0.5	53
Workbook (Environmental Science)	47	0.5	53

Broadly speaking, many ToT trainees felt that the material provided at the training was relevant for their teaching practice. However, it seemed that views on and experiences with the availability of the materials were more mixed. Some interviewees reported that there had been a shortage of material during the training they attended. In some cases, no hard copy material was provided during the training and participants had to rely on what the trainers were projecting to follow along with the training. One FTT from Benishangul-Gumuz was particularly critical, asserting that the lack of material compromised the quality of the training:

> *I don't think [the participants] understood [the AfL concepts] very well. If we were given hard copy manual at least we could have gained better concept. They were showing us what they have understood by projector. When any training is organised, training manuals have to be given. We can use that material as reference when we came back home.* (BeGu19)

Although many interviewees did report that material was provided after the ToT training, so that attendees had reference material they could take back to their schools, the way in which material was provided was variable. Many informants across regions reported receiving at least one soft or hard copy per school, so that the material could be shared and further disseminated during the school and/or cluster training. There was, however, some evidence that not making materials available to all ToT participants created a risk for the later availability and dissemination of the material. For instance, one FTT from Tigray reported that her school only had one copy of the manual, which could no longer be found:

> *They give us the training manual and it was very helpful […] but in the recent time we miss it. It was with the principal […] I had a plan to give a training to the new teachers this year but I couldn't find the manual that we get from the TOT training. Due to this reason I was discouraged to deliver the training.* (Tigray15)

Moreover, some ToT-trained informants, particularly in Benishangul-Gumuz, reported having requested but not received training materials and continuous assessment formats following the training. According to one FTT, for instance, despite seven months having passed since the ToT training, "no one sent it to us as per their promise" (BeGu19).

According to school- and cluster-trained interviewees, the availability of material was also variable. While some informants reported receiving (or providing) copies to each training participant, sometimes even in both hard and soft copy format, others noted that a shortage of material either meant that not all attendees received a copy, or that attendees had had to share the material during training and even return it afterwards.

4.2.4 REGIONAL DIFFERENCES IN IMPLEMENTATION

In this section, we highlight different approaches used to implement the AfL programme, as informed by our qualitative data collection.

The AfL programme took a very flexible approach to implementation. Interviewees at ELIXIR pointed out that regions with a more established background in human resource management (such as Amhara, Oromia and Tigray) as well as emerging regions (including Afar, Benishangul-Gumuz and Gambella) participated in the programme. Thus, different needs across regions had to be addressed through flexible and differentiated implementation:

> *When we offer training when we go in some of the regions to perform the AfL programme if there is like peculiar needs in that region or if there are things that we felt didn't work so we have to negotiate and modify so that it will be suitable.* (KII1)

While all the programme materials followed the same structure, the content and level of detail differed by region and local demand. For instance, the programme was recently rolled out to schools in pastoralist communities and refugee camps, where a considerable share of teachers has not obtained any formal teacher training. As informants at ELIXIR highlighted, some teachers in refugee schools are themselves Grade 8 or 10 students. Especially for these schools, ELIXIR designed a visualisation of the core AfL ideas on posters to make the material more accessible.

Given the different regional approaches to the implementation of the programme, experience sharing between the regions played a central role in improving day-to-day implementation, as an informant at UNICEF described:

> *One region will get an idea and run with it another region will look at it and say 'but it is too difficult to do' and they drag until they hear the other region has done something like this and it's 'oh we need to learn from them' so there is a lot of metamorphosis […] across the regions.* (KII2)

Experience sharing has also played a central role in the integration of the AfL programme into CTE pre-service curriculum. The Oromia REB pioneered this approach and inspired other regions such as Afar, Tigray and Amhara to also integrate AfL into their regional teacher pre-service training.

In Oromia and Tigray, the REB is playing a major role in the integration of AfL into pre-service training, while the process has not yet been initiated in Benishangul-Gumuz (see section 4.4.2).

In Oromia, the REB informant highlighted that the AfL programme has been implemented with increasing effectiveness and efficiency over the past few years. The key informant at the REB exemplified this with the switch from centralised training provision to more cost-effective decentralised training provision:

> *In the years 2016 and 2017 we were conducting AfL trainings at regional level. The trainings were costly or expensive as teachers are supposed to come to the training place (i.e. Adama and/or Sebata); covering all their accommodation was up to us. And in this we have invested a tremendous amount of money. It is after we discussed budget related issues at a workshop held at the end of 2017. Based on this we have changed the training areas to the colleges and at woreda level. With this, we increase the coverage of the training and came to train more teachers with less cost.* (Oromia23)

4.2.5 DIFFERENCES AT THE DECENTRALISED LEVEL

Within the AfL programme, the WEOs were generally in charge of coordinating school and cluster training and with following up on the use of AfL methods and tools in schools. However, the coordination of school and cluster training from the woreda differed substantially across regions, woredas and even across schools within the same woreda.

For instance, in Bambasi woreda in Benishangul-Gumuz, we visited two cluster schools. In one school, the WEO and cluster supervisor organised jointly, with the school director, a cluster training with 56 participants including AfL modules and other "important points on education" (BeGu21). In contrast, at the other school, teachers complained that neither the WEO nor the school director supported them in providing school or cluster training. Moreover, an FTT teacher at the latter school reported feeling not sufficiently prepared to conduct the training by himself as he had only attended one five-day training. Consequently, he only gave a short briefing to the other untrained teachers in the school on how to prepare and use lesson plans. In Homosha woreda, a different school followed yet another approach to conducting the school training. Instead of conducting the training in one session, the school director organised different training sessions, spread out over two days, during the free periods of teachers. A disadvantage of this structure was that the training was interrupted several times. The school director also acknowledged that there had not been sufficient time to discuss all the relevant content and preferred to call the session an "orientation" rather than a "formal training" (BeGu13). Again, the school director did not mention any support from the WEO.

In Tole woreda in Oromia, the WEO took a central role in organising and conducting the cluster training. As part of the qualitative data collection, we visited three schools within a cluster where teachers from all schools had obtained ToT training. Instead of facilitating a school training at each school, the WEO opted to host a centralised training session for all teachers within the cluster to better use the resources allocated for school and cluster training. A ToT-trained school director from the cluster explained:

> *We have trained about 50 teachers at woreda level using PowerPoint. Woreda Education Office had allocated a budget of 5000 Birr for school training. As the budget was meagre, we opted to conduct cluster/woreda training.* (Oromia8)

The WEO also reported provided per diems to the attendees. However, several participants complained that the training was too short (only two hours), the group size too big and that prior knowledge on AfL methods differed among the participants. An FTT even lamented that paying per diems to attendees reduced the resources available for the training and thus negatively influenced the quality of training, particularly its length.

In Ambo woreda, the WEO was also fully engaged in organising school training at all three cluster schools that we visited. For instance, at one school the WEO reportedly organised the entire school training; at another school, the WEO was said to have played a crucial role in helping the FTT to motivate the other teachers to attend school training without a per diem.

To avoid school closure due to AfL training at the school level, the school director and teachers at one school took advantage of harvest week. During harvest week students are mostly busy helping their parents to collect crops and thus not attending school.

In contrast, based on the qualitative data collected in Tigray, the WEO did not appear to take an active role in cascading knowledge to untrained teachers at cluster and satellite schools. Teachers from two cluster schools that we visited in Gulomekeda and Kilete Awelallo woredas reported that their WEO was not involved in providing school or cluster training. FTTs at one school in Gulomekeda woreda noted that they had provided the cluster training for 44 teachers together with their cluster supervisor, who was also based at their school. Given the lack of follow-up from the WEO, FTTs at one school in Kilete Awelallo woreda had not provided a formal school training. Instead, similar to a school in Benishangul-Gumuz (described above), the FTTs only conducted an informal briefing on AfL methods with their untrained colleagues within their school.

Overall, there are stark differences in how WEOs support schools in providing school and cluster training. The extent to which the WEO supports schools appeared to influence whether cluster schools cascade the training or not. Regardless of the WEO support, teachers in all three regions and woredas underscored that school and cluster trainings are by far too short to cover the AfL concepts taught in the ToT training (see Section 4.4.1 for a more detailed discussion).

4.2.6 PROGRAMME MONITORING

To monitor the overall rollout of the AfL programme, the implementing partner, ELIXIR, tracked performance indicators of delivery of the programme through a dashboard. The dashboard outlined the situation across all regions where the programme was implemented, showing the number of UNICEF supported Woredas, schools, teachers, directors, supervisors and CTE instructors, including the pastoralists and refugee camps, per each region. Training participants are also disaggregated by sex. In addition to the monitoring statistics, feedback from ToT trainers and participants in the training were also used to inform implementation reports developed by ELIXIR and submitted to UNICEF that outlined achievements, challenges and recommendations to improve the AfL programme in real time.

REGIONAL MONITORING AND SUPERVISION

In order to monitor the in-class use of AfL tools and materials, UNICEF created a supportive supervision tool to coach teachers and mentor them in the classroom while they are using AfL methods. Via this channel, UNICEF stakeholders argued, it was possible to ensure that teachers are implementing the intervention. The responsibility to conduct classroom supervision differs

from region to region. While in Benishangul-Gumuz and Tigray the REBs reported that they sometimes supervised teachers themselves, the Oromia REB did not engage in direct classroom supervisions. Instead, they provided CTEs with a mandate to supervise the implementation of AfL. Reportedly, CTEs supervised AfL implementation two to three times per semester, including the overall implementation supervision at the woreda level. Additionally, in all three regions, the WEO was commissioned to supervise the in-class use of AfL methods and tools. Moreover, the REB in Oromia encouraged peer-observations at the school level and prepared a spreadsheet to compare the implementation of AfL across different woredas. In Oromia, in 2018 and 2019, based on the supervisions conducted by the CTE and WEOs twice a year, a supervision report was produced to assess the overall project rollout in the region. Based on the results of the supervision and standardized checklist scores, schools were categorized into 5 categories with 'model school' given to those schools with highest quality implementation and schools that needed further follow-ups or support. However, our key informant at the REB in Oromia stressed that they lacked the budget to organise supervisions to the extent that they wished (see also Section 4.2.8 on 'Cost'). Also in Oromia, a study, commissioned by the REB, was conducted shortly after the rollout of the programme to guide the implementation and understand the current situation of continuous assessment within the region and provide recommendations for the programme going forward.

Along with supervising the implementation of AfL, REBs in all regions were also in charge of validating AfL materials, selecting ToT-trainers jointly with CTEs and selecting intervention schools. In addition, the REBs in Oromia and Tigray also supervised the implementation of pre-service training at CTEs. Overall, the implementation of AfL appeared to be more institutionalised in Oromia than in the other regions. The Oromia REB had even created an AfL steering committee, in charge of budget allocation and supervision of in-service as well as pre-service training.

WOREDA MONITORING AND SUPERVISION

Besides supporting schools in providing school and cluster trainings, the WEO was also commissioned to supervise the implementation of AfL methods at schools. As with facilitating cluster and school training, there was a lot of heterogeneity across woredas as to how they conducted supervisions.

In Benishangul-Gumuz, it seemed that the WEO regularly supervised the implementation of AfL in schools. Particularly in Bambasi woreda, teachers reported having been supervised several times by WEO staff on their implementation of AfL. Teachers and school directors in Bambasi reported that WEO staff came about two times per semester to evaluate the in-class use of AfL methods and tools along with general teaching practices. Feedback was mostly given orally to teachers immediately after the classroom observation. A few teachers in Bambasi also reported receiving written feedback, which was stored in the school director's office. Though to a lesser extent, the WEO in Homosha also reportedly conducted regular AfL-related classroom supervisions. Some teachers from both woredas even stated that they were occasionally supervised by the REB, which was also supported by the Benishangul-Gumuz REB interviewee.

In Oromia, the evidence on WEO supervisions on AfL implementation was more mixed. In Ambo woreda for instance, even though the WEO official interviewed reported that there was a dedicated person at the WEO to conduct AfL related supervisions and follow-up, and a school director stated that the WEO incorporated the AfL supervision into the general school supervision checklist at woredas, only a few teachers reported having received WEO supervision. In contrast

in Tole woreda, the WEO regularly supervised the implementation of AfL in schools. According to a teacher:

> *They come to school and follow teaching learning process, visit the class, observe the way we were teaching and manage our students, gave us a copy of lesson plan and evaluated how we are following up our students.* (Oromia1)

In Tigray, on the other hand, none of the interviewed teachers in either woreda visited reported having received AfL-related supervision from the WEO. Several teachers expressed their desire for more follow-up from higher-level institutions such as the woreda, ToT trainers or the REB . The only observation that appeared to be common at the school-level was peer supervision and supervision by the cluster supervisor. However, as a cluster-trained teacher pointed out, higher-level follow-up is a key part of ensuring quality of implementation and teaching in schools:

> *In my opinion the woreda education has to come at least once a year but the past semester no one came here but they have to know if the teacher is working well or not and they have the responsibility to assure quality of education. But a report cannot show the actual condition. The higher officials have to control the implementation.* (Tigray2)

PEER SUPERVISION AND SCHOOL ACTION TEAMS

In all three regions, classroom observations and supervision for teachers appeared to be an integral part of the teaching process at schools. However, the extent of classroom observations differed starkly within regions, woredas and schools. Moreover, classroom observations did not always include an evaluation of teachers' use of AfL methods and tools.

In Benishangul-Gumuz teachers and school directors in all schools reported that AfL-related classroom observations were common practice in their schools:

> *They observe our teaching style, how we participate students in class activity and group work, and also they observe our classroom management.* (BeGu15)

> *What we observe is everything based on the prepared checklist, that includes use of blackboard, dressing code, class hygiene, student control, group discussion, class work, checking students exercise book, use of lesson plan, is that teacher moving faster or behind annual plan, or he is on the right track, his presentation skill and attitude he has towards student.* (BeGu5)

In some schools, teachers and school directors formed an AfL action learning team, which supervised the implementation of AfL at schools. Teachers and school directors conducted classroom observations and shared mostly oral feedback on a monthly basis. However, this structure did not appear to be formalised in all schools we visited. Additionally, cluster supervisors and WEO staff also supervised the implementation of AfL in some schools in Benishangul-Gumuz.

One cluster supervisor in Ambo woreda, Oromia, reported organising supervision teams in each school to oversee teaching. However, because the cluster supervisor had not received any AfL training and therefore could not prepare appropriate items accordingly, it was emphasised that the checklist provided to teachers to conduct classroom observations was not related to AfL. Indeed, a teacher from the cluster reported that:

> *We have a regular observations and feedback system in our school. I have been having observations and receiving feedback multiple times but not as part of the AfL.* (Oromia19)

Yet in another school in Ambo woreda in Oromia, teachers reported a structured, cascading approach to AfL supervision in their school. ToT-trained school directors supervised and observed FTTs, who in turn supervised school-trained teachers following a common AfL-related observation checklist.

In Tigray, teachers reported that peer-supervision is common at schools, however, often unrelated to AfL. Only teachers at one school reported that the AfL action learning team at their school supervised their day-to-day use of continuous assessment techniques in teaching. Moreover, according to the school director at that school, the teacher supervision system also included informal and random student assessments:

> *We give feedback to teachers after we made assessment when he/she give class we some time use some students to assess the teachers performance we ask them some question to the students randomly then based on the answer we receive from the students we give feed back to the teacher.* (Tigray21)

Few schools across all three regions reported that they had set up a question bank at the school level. A question bank would allow teachers to submit their questions to be evaluated by selected teachers, or the AfL action learning team, on the appropriateness of the question, before approving it for assessing students and storing it thereafter. However, this practice was not widespread in any of the woredas and clusters that we visited.

4.2.8 PROGRAMME DELIVERY CHALLENGES

In this section, we describe the challenges in delivery that have been faced by the AfL programme and potential solutions, as informed by our qualitative data collection, in addition to findings from the quantitative survey with teachers and school directors in Oromia with regards to the contextual barriers to implementing effective continuous assessment.

LACK OF ADHERENCE TO THE CASCADING KNOWLEDGE MODEL

A major challenge in the implementation of the AfL programme has been the deviation from the cascading knowledge model.

First, as already described in Chapter 4.2.1, the majority of teachers and other participants responding to the quantitative survey did not attend all three sessions of the ToT training, resulting in lack of training and hampering the continuity of the implementation. Based on the qualitative data, this pattern also appeared to be common in Benishangul-Gumuz, and we did not see strong evidence for it in Tigray. Not attending all sessions has a clear impact on the ability of FTTs to implement and cascade the programme. For instance, according to one FTT from Benishangul-Gumuz:

> *I didn't understand [the AfL concepts] very well. It was a little bit better for me since it is my second round, otherwise, it will be difficult for those trainees who joined the training for the first time. The new trainees seemed confused when exercises were given in the class. It will be difficult for the first-time attendants to gain better concepts.* (BeGu19)

Key informants from ELIXIR postulated that on the one hand, some schools might have sent different teachers to the ToT training in order that more than four teachers would have the opportunity to attend and benefit. On the other hand, staff turnover (discussed further below) may have meant that it was not possible to send the same teacher to all three sessions. According to a teacher in Oromia, insufficient capacity at ToT training sessions may also have excluded some who had previously attended from completing more sessions.

Potentially in response to some of the issues outlined above, or potentially simply due to administrative negligence or error, ToT training sessions did not appear to be differentiated by content or attendees' previous experience, as asserted by an attendee from Benishangul-Gumuz:

> *When region invite teachers for training, they didn't distinguish between the previously training and untrained teachers. They simply call and conduct training both previously trained and untrained teacher all together. This type of training methodology is not fair since all teachers have no similar awareness on AfL. Those who attended the training previously is going to take it for the second time, whereas the new teachers are there for the first time, in this case it will confuse the new teachers. Therefore, training should be called separately and conducted separately […] Otherwise it will create knowledge gap and affect programme implementation on the ground and also affect the quality of the training.* (BeGu16)

As already described in Section 4.2.2, participants of the ToT training did not always pass on their knowledge through school and cluster training. Some teachers in Oromia reported not organising a school or a cluster training because their colleagues were not motivated to attend the training (see section 4.2.8). For instance, according to one ToT-trained school director, teachers would not attend the training without being offered a per diem:

> *We didn't try [to organise the school/cluster training] because if it's called training there is the expectation of payment and we could not do that.* (Oromia33)

Other teachers, also from Oromia, emphasised that they lacked the budget to provide the school and cluster training.

Even if the training was passed on, the length of the training often was not appropriate for similar reasons that training was not conducted – teacher motivation and budget constraints (see section 4.2.8). Teachers in all three regions argued that school staff would be more motivated to attend a direct training than a school or cluster training. Following pressures by teachers to receive per diems for their attendance to the cluster training in Oromia (see section 4.2.8), a woreda official in Oromia reallocated the assigned budget to provide per diems to teachers. A ToT-trained school director who participated in providing cluster training said that the programme's focal person at the woreda education office handed 90 Birr to every attendee (Oromia8). However, the training only lasted 2 hours and according to a ToT-trained teacher, who was also invited to attend the cluster training, this sparked anger among the attendees. According to him the attendants were upset because they felt that the budget was wasted (Oromia18).

More broadly, informants criticised the integrity of the cascading model, arguing that 'downstream' training does not suffice to implement AfL into daily teaching. A common demand among teachers in all three regions was to give direct training to all teachers and not to rely on the cascading of knowledge. An untrained teacher from Tigray summed this concern:

> *As to me, if the region education bureau trains to the woreda education bureau then to the schools, change will not be seen though this chain rather, the training should delivering directly to the ground at a school level or else if the CTE trains well and should deliver the training to the lower level as it is. Otherwise if the region to the woreda then the woreda education bureau also training to the supervisor then the supervisor passes an order though this chain I don't think we can change.* (Tigray 21)

As a result of the problems related to the cascading knowledge model, many teachers report a lack of training being a major challenge in implementing AfL methods in all three regions. Moreover, even if the school or the cluster conducted an AfL training, the length and quality of the training is often not as intended (see section 4.2.2). In addition, even ToT-attendees often wish to obtain more and longer training to increase their knowledge on AfL (see section 4.2.1). Finally, teachers do not always attend the full ToT training cycle as they are supposed to, resulting insecurity about AfL concepts.

We explore the reasons – staff turnover, motivation and cost – underlying the lack of adherence to the cascading knowledge model below.

STAFF TURNOVER

As mentioned above, staff turnover at schools was cited as an obstacle to the sustained implementation of AfL in schools. One of the interviewees at ELIXIR noted that "we expect them to give training, but the next time we go they are not there" (KII1).

Indeed, informants in all three regions similarly saw teacher turnover as a challenge in the implementation of AfL. Both teachers and school directors conveyed that teachers might leave after having attended the ToT training, making it impossible to cascade the knowledge to other teachers at the school. For instance, an FTT from Benishangul-Gumuz expressed these concerns:

> *The other challenge is there are teachers attended the AfL training for two rounds but transferred to other school without starting any implementation activity. Some teachers have also resigned after receiving training.* (BeGu6)

Indeed, the quantitative survey with school directors in Oromia indicated that there was a teacher turnover rate of 18% per year in their schools and the average tenure of a teacher at their current school was 3.4 years, meaning that the average teacher would have joined after the initial implementation of AfL in Oromia.

According to a school director from a remote school in Tigray, teacher and school director turnover is particularly problematic in remote and rural areas. A ToT-trained teacher from Oromia expressed particular concern about school director turnover, underlining the importance of school leadership buy-in for the sustainability of the programme:

> *There are some factors which may bar us from conducting school training in the future: removal of school leadership which perhaps engender new leadership with insufficient awareness about the programme and new teachers with reluctance to take the training. These and other factors may affect provision of school training in the future.* (Oromia6)

To mitigate the issue of school staff turnover, one school director proposed to train village officials as well, who could give training to newly arriving teachers in remote areas. One school in Oromia

has developed an effective approach by ensuring that all new teachers first received school training. As the school director explained:

> *Whenever a new teacher is employed by our school, he or she should receive the training before starting teaching. This is a mandatory prerequisite in our school. Otherwise, he or she cannot enable his or her students to perform well. This trend for sure will persistently continue in the future too.* (Oromia7)

TEACHER MOTIVATION AND PER DIEM

The motivation of teachers to attend school and cluster training without a monetary payment was reported to be poor. As several teachers, school directors as well as higher-level officials pointed out, teachers expected per diems for attending the cluster and school training, mainly because their colleague FTTs received daily payments while attending the ToT training. The CTE instructor in Oromia, where numerous teachers reported this issue, described the general problem:

> *The motivation of the teachers to practice AfL is scanty. It is affected by the benefits they think they should get when they take the training. Currently, school/cluster trainings are provided without per diem payment against the expectation of the teachers. Only some refreshment is provided for the teachers when they sit for the training. This has huge damage on the implementation of the programme for it kills their interest at the beginning stage.* (Oromia20)

Some teachers in Oromia complained that this circumstance deteriorated the quality of the school and cluster training. This issue was also prevalent in Benishangul-Gumuz and Tigray albeit to a lesser extent:

> *Most teachers do not want to get the training by their colleagues at their school or working station. Instead, they want to go out of their site and receive training and get paid for that. Otherwise, they don't want to hear from others.* (BeGu23)

Indeed, ELIXIR was aware of the issue of per diems affecting teacher participation and motivation. In response to this concern, ELIXIR recently shifted their focus from the cascading knowledge model to pre-service training in CTEs:

> *In some [regions] they require for per diem for example at school level. This is the reflection of teachers, the key teachers, when they organise a workshop the other teachers [say], "oh you go to the CTE you will have per diem" […] so this is one challenge really openly that's why now we give to CTEs.* (KII1)

COST

Budget constraints on the school and on the woreda level also limited the implementation of AfL. Woreda education officials in all three regions remarked that woredas and schools lack budget to properly implement cluster and school training as well as school supervision. As already described above, teachers demanded per diems to attend cluster and school training, which were not reflected in the budget. A ToT-trained teacher from Oromia described that she was only able to provide a one-day training due teachers' requests for per diems and the limited budget. In addition, the WEO staff member from Oromia who was interviewed reported that the budget allocated for training only considered refreshments, and not materials and stationery. Moreover,

the WEO interviewee argued that the allocated budget for school training should also consider the costs that the woreda incurred to supervise the training.

LACK OF SPECIFIC SUPERVISION FOR AFL

The highest level of supervision of the AfL programme are joint supervision visits undertaken by CTE deans and supervisors, REB officials, UNICEF and ELIXIR. The aim was to visit the school, interview the school director, observe lessons and provide feedback to the staff. According to the Oromia REB Supervision report, the majority (91%) of AfL schools were subjected to a supervision by the CTE (it should be noted that Oromia REB delegated the task of supervision to CTEs). The joint supervisions had clear structures for the reporting of the findings at a high level and produced yearly reports.

Despite the joint supervisors described above, based on the qualitative interviews, another major bottleneck in the implementation of AfL has been the lack of follow-up and supervision from the WEO and the cluster supervisor. Teachers and school directors in all three regions complained that after having attended the training, they were often left alone with the practical implementation of AfL. They stressed that the lack of engagement by higher level officials hampered the continuity of the programme's implementation. As some FTTs argued, supervision was needed in order for the programme to be sustainable:

> *To implement this programme effectively, we need a responsible person who cares for the implementation of programme; especially in our woreda there is a little bit carelessness on monitoring the progress by visiting the school where the programme is applied.* (Oromia32)

> *There were people from woreda who get trained with us from administration and supervisors but they are not doing their role still they didn't even visit us, but they have to know what is going on the ground.* (Tigray10)

This argument was backed up by other interviewees. For instance, another FTT from Oromia felt that teachers needed to be supervised to ensure that they were implementing AfL, as "some teachers are not responsible enough" (Oromia4). A ToT-trained school director from Benishangul-Gumuz even admitted that he does not follow up on the teachers in his school because nobody supervises him – it was also found in the quantitative survey that 23% of cluster school directors did not conduct observations themselves to assess the implementation of AfL. According to this school director, there will be no sustained change in teaching practices unless teachers and school directors are not properly supervised. The low rate of AfL specific supervision is also found in the quantitative survey in Oromia, with 44% of teachers in cluster schools reporting that they had been observed by a cluster supervisor in the past 12 months, however only 15% reported that the observation and feedback was related to the AfL programme. Again, cost may have been an underlying reason for the perceived or actual lack of supervision. In this regard, the Oromia REB informant asserted that:

> *We have conducted supervision up to a level of our capacity with what resources we have so far. But I feel that we need more oversight to increase its effectiveness and efficiency. However, the budget remains a big bottleneck; had more budget been allocated for the supervision activities, we would have up lifted its implementation and contribute more for its successful implementation.* (Oromia23)

However, stakeholders also pointed out that it was the quality and focus of existing general school supervision that need to be adapted. According to informants at ELIXIR, some regions have already included AfL adherence as part of the general performance appraisal of teachers. Yet, as a CTE instructor in Oromia noted, standard school supervision:

> *focusses on coverage of the books rather than the learning level of students. The supervision concentrates on how much the students learn, not how much they understand. The question of the supervisors to the teachers is how much portion they have covered, not how much have their students understood.* (Oromia20)

A ToT-trained teacher from Benishangul-Gumuz supported this claim and asserted that the goals of the standard school supervision undermined the continued implementation of AfL:

> *Most of the time, there is supervision done on teachers to evaluate them whether they are teaching their course based on the academic calendar. Do the annual lesson plan and the book they teaching now match or not are what they evaluate. So, if you struggle to capacitate similar student, it will pull you back and prevent you from finishing your subject on the planned schedule.* (BeGu6)

Moreover, some teachers in Benishangul-Gumuz and Oromia criticised the general school supervisions. Some teachers argued that the feedback from supervisors was mostly negative and not encouraging or constructive, while other teachers complained of not receiving any feedback after supervision. More generally, a ToT-trained teacher from Benishangul-Gumuz criticised the woreda for not taking action upon identifying gaps during supervision:

> *They have to come for supervision and identify existing gaps and give appropriate response for that gap. They are coming sometimes for supervision, but I didn't see them taking rectifying measures for the gap identified.* (BeGu15)

LACK OF TEACHERS' MOTIVATION

Beyond lacking motivation to attend the school or cluster training, school directors from Oromia and Benishangul-Gumuz and a ToT trainer from Benishangul-Gumuz also reported that some teachers lacked the motivation to implement AfL in their teaching. In the quantitative survey with teachers, only 17% reported that teacher motivation was a 'big challenge' to implementing continuous assessment, although 36% suggested that it was somewhat of a problem, indicating that motivation may still be a barrier to the success of AfL.

The ToT trainer contended that even though they followed up on teachers, teachers were not implementing the tools as they are expected to. Student progress recording and lesson plan preparation were raised as particular issues. However, while he personally linked the problem to teacher motivation, the underlying reason for not using the forms as expected might be that teachers face large class sizes and time constraints (see time constraints and large class sizes, below).

An FTT from Oromia and a ToT-trained school director from Benishangul-Gumuz also postulated that teachers might not be motivated to implement AfL methods in their teaching if they had not participated in the ToT training:

> *Most teachers hesitate from implementing the programme for the only reason that they haven't been invited to ToT and paid per diem. They think that the teachers who are paid per*

diem are responsible to implement it. They perceive that the programme belongs to those teachers. Thus, it is better if the training is given for all the teachers uniformly. The trainers better come to this area and provide the trainings. All the teachers will feel responsible. (Oromia18)

4.3 EFFECTIVENESS OF THE PROGRAMME

In the following section, we examine the effectiveness and impact criteria in terms of the AfL programme. We begin by assessing the ability of teachers to implement continuous assessment techniques in their daily teaching practice and then attempt to understand the AfL programmes role in any changes. To investigate teachers' ability to use continuous assessment, as per the ToC assumptions, we firstly assessed whether teachers possess the knowledge to implement continuous assessment in their classrooms and secondly to what level they actually did implement it and the associated techniques. In addition, we also investigated the level of parental engagement in education such as the channels and challenges faced by both schools and parents and whether AfL was effective in engendering parental engagement in their child's education. To do this, we conducted both descriptive analysis through using data from both the quantitative teacher survey, classroom observations and qualitative interviews. and parental engagement in education. Finally, we looked at student performance as an ultimate goal of the AfL programme in improving education and learning outcomes in Ethiopia. For each area, we initially conduct descriptive analyses and then attempt to capture programme effects through multivariate regression analysis.

4.3.1 TEACHERS' KNOWLEDGE OF CONTINUOUS ASSESSMENT

One of the underlying assumptions for teachers to be able to use continuous assessment techniques is that they have the capacity to do so and have a solid understanding of the key components described in Chapter 1 that underpin continuous assessment and techniques included in implementing them in the classroom.

While findings from the qualitative data cannot be considered comprehensive or generalisable, overall, AfL FTTs expressed that they had gained knowledge about techniques from the training and understood that they should implement continuous assessment in their teaching. According to school directors, the ToT training had helped teachers embrace and implement continuous assessment techniques:

> *The older continuous assessment method was focussing on student behaviour, attendance and personal hygiene was the criteria of the student to get mark to be promoted to the next grade. But now AfL training has helped teachers to get out of such traditional assessment technique and shift to the modern way of assessing student competencies.* (BeGu13)

> *When we see the [AfL FTT] there is a change, we can see this from the questions she prepared on weekly based and provide to the school which is found in the office. I can notice the change from her answers she gave me when I ask her.* (Tigray21)

However, a small number of FTTs interviewed expressed reservations about their ability to implement continuous assessment in the classroom. For instance, one teacher felt that the training lacked practical exercises while another asserted that the training was not adequate and that

"there are many things that I want to know about AfL" (BeGu15). A school director in Oromia expressed the following concerns:

> *Because of the shortage of the training time most of us didn't fully capture some of the AfL concepts; for instance, about MLC and how to fully handle a continuous assessment.* (Oromia33)

Another FTT felt that further follow-up training and support was necessary, especially as the time from when the ToT training was conducted widened:

> *The training was three years ago. Some form of follow-up was done during the first year of implementation. But after a while, no one came and saw us. When the follow-up and support is disconnected, we gradually started to go back to our previous and accustomed teaching methodology.* (BeGu3)

The general preparedness expressed by FTTs to implement AfL was in contrast to the views of school- or cluster-trained teachers who were interviewed. While some reported that they had gained knowledge on and felt prepared to use continuous assessment techniques, other interviewees asserted that the training they received was either too short or insufficient to enable them to understand and implement AfL:

> *I don't feel prepared as such, I have to learn more […] I was even complaining on to my school director that I need more training. It is still difficult for me to apply recording that categorise student as per their competence with level 'A' or 'B' and so on. I have started only preparing the lesson plan and it is almost one week since I started AfL implementation. The school director advised me to apply at least the lesson plan to be uniform with other teachers. He has evaluated me and gave me feedback. He has promised me to train me next on recording and the remaining packages*. (BeGu12)

> *No, I didn't feel that the training fully prepared me adequately for implementing AfL in my teaching. For instance, we are overwhelmed by large class size; and they didn't explicitly tell us how we have to manage the continuous assessment concept for large class size like ours […] I will continue using the AfL tools in my teaching but in a fragmented way as I am not well equipped with the usage of these tools.* (Oromia19)

The findings from the quantitative survey in Oromia also generally suggest that teachers in treatment schools are more likely to be prepared to implement continuous assessment than other teachers. In the survey, teachers were tested on their understanding in the following areas of knowledge. In all areas, teachers in treatment schools (who may have received ToT or school training) performed better than teachers in satellite schools (who should have received cluster training).

To attempt to quantify teacher knowledge on continuous assessment we used a set of test questions in the teacher surveys. Again, to examine this we refer back to the key components of continuous assessment described previously in Figure 11 on page 35. We assessed teachers ability to explain the various key components of continuous assessment and then we examined more deeply their understanding in two of the components that had more nuanced theoretical ideas behind them, that is the development of high quality questions for assessment tools and the provision of effective feedback. The areas we assessed are described in Table 14. To source

appropriate questions, we used Ethiopia's National Educational Assessment and Examinations Agency (NEAEA) continuous assessment manual.

Table 14. Knowledge areas of continuous assessment

Knowledge Area	Description
Continuous Assessment Components	• Ability to describe various components of continuous assessment.
Question Development	• Demonstrable understanding of the different elements (validity, reliability, objectivity and fairness) that make a question a useful assessment tool.
	• Ability to identify good practice for conducting assessments and setting questions within a classroom of students.
Feedback Provision	• Demonstrable understanding of the difference between constructive and unconstructive feedback.
	• Ability to identify examples of constructive and unconstructive feedback.
	• Ability to identify good practice for providing effective feedback to students based on assessments.

CONTINUOUS ASSESSMENT COMPONENTS

Of all the teachers interviewed, 99.5% indicated that they understood what was meant by the phrase 'continuous assessment' indicating that it was a concept known almost universally by teachers. However, to assess the knowledge deeply, as a follow-up, teachers were asked to describe fully what their understanding of the concept of continuous assessment by describing the various components involved, without prompting. Aligned with the previous discussion on the key components of continuous assessment, the following checklist was developed for this question:

1. Regular assessment of students' learning progress within the classroom with assessment tools that are aligned with the curriculum;
2. Using a mix of formative and summative assessment methods;
3. Structuring lessons with clear learning aims to meet competencies;
4. Utilising a range of assessment techniques in the classroom;
5. Providing timely, and often immediate, feedback to students on their progress;
6. Recording student progress; and,
7. Evaluation of student performance to make teaching and learning more effective.

Each of the items in the checklist was given a score of one if raised and described by the respondent, meaning a total score of between 0-7 for the teacher. After each description, the enumerator probed and asked if there were any further components of continuous assessment they felt they had not already described before completing the question. On average, teachers were able to include three of the above items into their response. The majority of teachers were able to explain that continuous assessment as being regular assessments of all students' ability and using a wide range of assessment methods. However, less than one-quarter of teachers discussed the use the information gained from regular assessments to alter their teaching for classes and individual students as part of continuous assessment. As shown in Table 15, teachers from cluster schools scored the highest on this question (3.4 out of 7) and satellite school teachers performed the weakest (2.9)

Table 15. Continuous assessment component score by treatment

Treatment group	Continuous assessment components score (0-7)	Std. Deviations	Observations
Cluster school	3.4	2.1	94
Satellite School	2.9	2.2	45
Control school	3.2	2.0	42
Pooled	3.2	2.1	181

QUESTION DEVELOPMENT

To assess a deeper level of knowledge of continuous assessment, we selected the component relating to using high quality questions as assessment tools of students learning to test teachers' knowledge on further. The development of high-quality questions was also a key technique focussed on in the AfL programme. If a teacher possessed the ability to develop their own high-quality questions, linked to specific MLCs, this would not only help the programme be effective but also sustainable in the longer term. The development of teachers' own questions would provide greater flexibility to classrooms, particularly in employing a range of assessment methods, as opposed to relying on textbooks or reference books for their assessment tools. The questions developed, if recorded and stored properly, would also provide some institutional knowledge within the school with an inventory of questions developed by teachers that would remain even after the teacher left the school. Whilst teachers developing their own questions is seen an effective tool in improving teaching practice (Ball & Feiman-Nemser, 1988) and a more sustainable method for teachers throughout their careers, having the necessary ability to develop high quality questions is often a challenge and was identified in the context as a challenge in the midline evaluation (CfBT, 2015).

For a question to be considered an effective assessment tool, it should meet the criteria of being valid, reliable, objective and fair (Brown et al., 1996). To measure teachers' understanding of these criteria, they were asked to explain what each of the criteria meant in reference to developing questions to use in class with students. Enumerators, who received extensive training on the criteria and provided with a model answer, used a rating scheme to mark the teacher's response on a scale of 0-2 for each criterion. A score of zero meaning that the respondent was not able to respond at all, 1 meaning the respondent showed partial knowledge and 2 meaning the respondent was able to explain fully what each of the criteria meant in terms of question development. Teachers displayed the strongest knowledge on what makes a question fair and the weakest on being able to describe what makes a question objective. Table 16 shows, again, that teachers in cluster schools performed the best on demonstrating understanding of high-quality questions.

Table 16. Question development score by treatment

Treatment group	Question development score (0-8)	Std. Deviations	Observations
Cluster school	5.8	1.7	94
Satellite School	4.8	1.8	45
Control school	4.5	1.9	42
Pooled	5.2	1.8	181

In addition to understanding the criteria of effective questions, we wanted to examine teacher's knowledge on wider good practice of developing and setting questions in class. To do so, teachers were asked a set three of true/false questions based on whether they believed a set of statements represented good practice for setting questions in class or not. Teacher were typically able to identify correctly that questions should vary in difficulty to allow for an item to discriminate between students on their ability (89% correct) and that questions should always directly link to the curriculum (83% correct). However, many teachers (43%) incorrectly responded in agreement that questions should always be answered individually by students, as opposed to allowing students to work in groups or pairs to assess knowledge. As shown in Table 17, control teachers performed the highest and slightly outperformed cluster school teachers, with satellite school teachers performing the poorest.

Table 17. Question practice score by treatment

Treatment group	Question practice score (0-3)	Std. Deviations	Observations
Cluster school	2.3	0.7	94
Satellite School	2.0	0.9	45
Control school	2.5	0.6	42
Pooled	2.3	1.8	181

FEEDBACK PROVISION

The second component selected for a deeper dive into teachers' knowledge related to feedback provision. Feedback plays a decisive role in learning and development of students in an educational setting and students perform faster and more effectively when they have a clear understanding of where they can improve (Hounsell, 2003) and studies have shown timely and effective feedback has been shown to have large effect size (Black and William, 1998a). Feedback can be categorised into either constructive feedback or unconstructive feedback (London, 1995). Feedback is considered to be constructive if its content is able to identify and inform the user on where they performed strongly and where they performed poorly, with clear instruction on how they can improve (Ovando, 1994). Similarly, as with the criteria of a high-quality question, teachers were asked to explain their understanding of the difference between constructive and unconstructive feedback. Enumerators used a rating scheme to mark the teacher's response on a

scale of 0-2 for each criteria with 0 meaning that the respondent was not able to respond at all, 1 meaning the respondent showed partial knowledge and 2 meaning the respondent was able to explain fully what the difference between constructive and unconstructive feedback. Just over half (51%) of teachers displayed some understanding of difference while 43% were able to explain fully, only 6% were not able to explain the difference at all. Table 18 shows that cluster school teachers slightly outperformed their counterparts in satellite and control schools in demonstrating knowledge on the criteria of constructive feedback.

Table 18. Feedback criteria score by treatment

Treatment group	Feedback criteria score (0-2)	Std. Deviations	Observations
Cluster school	1.4	0.6	94
Satellite School	1.3	0.7	45
Control school	1.3	0.5	42
Pooled	1.4	0.6	181

As an addition to knowledge on constructive and unconstructive feedback, to demonstrate that knowledge, teachers were provided four different examples of feedback and were asked whether it constituted constructive or unconstructive feedback. Overall, the teachers were able to demonstrate some ability to discriminate between constructive and unconstructive feedback with teachers correctly identifying two out of four examples on average. The area teachers struggled with, with only 27% answering correctly, was believing a positive statement alone was sufficient to be considered constructive feedback, even if it lacked information on what the student did well or suggestions for improvement. In addition, almost half (49%) believed that a vague instruction for the student to try harder was constructive even though the statement did not give the student sufficient direction on improving their work. Table 19 shows that teachers in cluster schools were better able to differentiate examples of constructive and unconstructive feedback (2.4 out of 4) compared to teachers in satellite and control schools (both 2 out of 4).

Table 19. Feedback identification score by treatment

Treatment group	Feedback identification score (0-4)	Std. Deviations	Observations
Cluster school	2.4	1.0	94
Satellite School	2.0	1.0	45
Control school	2.0	1.0	42
Pooled	2.2	1.0	181

Finally, alongside being able to explain the difference between constructive and unconstructive feedback and demonstrating an ability to differentiate examples of each, teachers were asked to identify wider good practice in the classroom for feedback provision. Teachers were read three statements and asked to indicate whether they agreed or disagreed that the statement

represented best practice. Teachers on average were able to correctly identify the best practice in around half of the cases with an average score of 1.5 out 3. One area that teachers performed particularly poorly in was believing that feedback should include direct comparisons to other students within the class when providing a student with feedback and 72% believed that was best correct. Table 20 reports the average scores by treatment group and shows that in relation to feedback best practices, cluster school teachers scored slightly lower than satellite teachers and control teachers.

Table 20. Feedback practice score by treatment

Treatment group	Feedback Practice Score (0-3)	Std. Deviations	Observations
Cluster school	1.4	0.9	94
Satellite School	1.5	1.0	45
Control school	1.6	1.0	42
Pooled	1.5	1.0	181

PROGRAMME EFFECTS ON TEACHER KNOWLEDGE

The findings above discuss the various components of continuous assessment that teachers were tested on and attempts to quantify their knowledge on continuous assessment as a whole. However, simply comparing the numbers between schools may be misleading due to differences between teachers at the schools. The reason is that it might be that teachers in cluster schools are more experienced, higher educated or more motivated than those from satellite or control schools (e.g. because of different funding or other institutional effects). Hence, to capture the effect of the AfL programme on improving teachers' knowledge on continuous assessment, we can compare outcomes for teachers from cluster and satellite schools against teachers from schools from the matched control schools and control for observable differences such as teacher characteristics and school characteristics and see if any differences between the treatment groups remain. The indicator used for teacher knowledge on continuous assessment practices is a composite measure of the teacher responses in the areas described above. Teacher scores on the domains described above were standardised and aggregated to give a single, composite outcome measure of teacher knowledge on continuous assessment.

Table 21 shows the estimated coefficients for the cluster and satellite schools against the control schools from the OLS regression.

Table 21. AfL treatment effects on teacher knowledge composite score

Treatment group	Estimated treatment effect
Cluster school	0.362** (0.168)
Satellite school	-0.347 (0.238)
Total Observations	176
Adjusted R-squared	0.11

Notes: Standard errors in parentheses; Regressions clustered at the school level; * p<0.1, ** p<0.05, *** p<0.01

Controls includes a set of teacher characteristics including age, gender, motivation score, self-efficacy score, digit span score, education and experience.

The results show that the AfL programme at cluster schools has a positive ($p<0.05$) relationship with teacher knowledge scores on continuous assessment. The magnitude of the effect is of 0.36 Standard Deviations (SD) of the composite score summarising teacher knowledge in continuous assessment compared with teachers in control schools. On the other hand, there is no evidence that teachers in satellite schools saw an increase in their knowledge on continuous assessment.

When interpreting the results above, it is important to note the limitations of the study design, that is while we can control for observed school and teacher characterises, there may still be unobservable differences between teachers and schools that could bias results and under or overestimate the effect of the AfL programme. It should be noted that as the unit of analysis is at school level, the coefficients represent the average treatment effect across the treatment groups and it's not possible to disentangle the effects of ToT training and cluster or school training. The findings above provide some promising evidence that AfL at cluster schools gives teachers better preparation and foundation of knowledge to implement continuous assessment in their classrooms.

4.3.2 USE OF CONTINUOUS ASSESSMENT TECHNIQUES

For continuous assessment to be effective in improving educational outcomes for students, teachers must not only understand but implement continuous assessment techniques in their classroom regularly and effectively. Hence, this section focusses on the application of the AfL knowledge and whether it transferred into actual changes in practice. Again, we use the key components of continuous assessment to guide the identification of various dimensions of teacher use of continuous assessment to investigate through quantitative and qualitative data. To quantify teachers' use of continuous assessment, we use self-reported and observed use of continuous assessment in classrooms.

ASSESSMENT RANGE AND RATE

Assessments can take many forms in both delivery by the teacher and response by the student and broadly falling into either oral, written or practical assessments (Le Grange & Reddy, 1998). Different types of assessments each have theoretical and practical benefits and limitations (N'Namdi, 2005; Guilbert, 1998) as well as varying in appropriateness depending on the MLC they are looking to assess. Teachers were asked to describe (without prompting) the different types of assessment methods they typically used in their classroom when assessing student learning. The most common methods of assessing students were homework and using oral questions in class in which there was little difference between the treatment groups, though homework was less common in satellite schools. Teachers from cluster schools were the most common in reporting that they used individual written assignments, oral presentations, debate and role plays and group tasks.

Table 22 outlines the count of the assessment types reported to be used by teachers, broken by treatment group, and shows that teachers at cluster schools, on average, use a wider range of assessments than their counterparts at satellite and control schools.

Table 22. Range of assessments by treatment

Treatment group	Average number of assessments used	Std. Deviations	Observations
Cluster school	4.0	1.4	94
Satellite School	3.4	1.5	45
Control school	3.5	1.1	42
Pooled	3.7	1.4	181

For each type of assessment that teachers reported that they typically used, they were asked to report how often they used the assessment method for an individual class to understand the regularity of assessment. Teachers responded on a six-point scale, with six indicating it was used every lesson and one indicating it was only used once per school term. Oral questions were the most regular method of assessment on average followed by homework and the least regular assessments types are debates and role-plays. The regularity of assessments was similar across treatment groups for most assessment types, although oral presentations, debates and role-plays were more common in cluster schools compared with satellite and control schools.

QUESTION DEVELOPMENT

Almost all teachers (98%) reported that they develop their own questions for use of assessing students in their classes and said that they used textbooks to find questions to use in class. During classroom observations, whilst reviewing the lesson plans, teachers were asked to indicate the various sources of questions for the assessments in that specific lesson. Textbooks were overwhelmingly the most common source 91% compared to 58% of questions the teacher had developed themselves. This suggests that whilst most teachers do develop their own questions, teachers still rely heavily on textbooks for assessment tools. This sentiment was expressed by an interviewed FTT in Benishangul-Gumuz, who noted: “I don’t have to develop questions, since the textbook already contains continuous assessment question in it” (BeGu7).

Apart from developing their own questions and using textbooks, the most common source of questions was from reference book (67%). Just under half of teachers (49%) reported using questions that colleagues developed and only 23% said that they took questions from a school question bank, indicating that resource-sharing amongst teachers in schools for assessment tools is still not widespread even within schools. Cluster schools had far greater dissemination of assessment tools between teachers from different schools, with 34% of teachers in cluster schools reporting that they used questions developed by colleagues at other schools compared to just 6% of teachers in satellite schools and 19% of teachers in control schools. This could potentially be a result of teachers meeting with colleagues from other schools as part of either ToT or cluster training sessions.

One technique promoted through the AfL is using supplementary materials in teaching, such as learning aids that are linked to MLCs, to increase students' engagement. The vast majority of teachers (88%) reported that they utilise supplementary materials in their class. However, there was no evidence that this practice is regularly employed within classrooms, with only 6% of classes observed using supplementary materials at some point during the lesson.

STUDENT PROGRESS RECORDS

For teachers to use the information provided by regular formative and summative assessments, they must have a structured way of collecting and recording the data on learning progress. An effective student progress record allows teachers to analyse and adapt their teaching at both a classroom and individual student level (Black & William, 1998a). As shown in Table 23, when asked if they keep either a written or electronic record of individual student achievement and progress throughout the year, 62% of teachers indicated they did with control teachers having the highest rate (74%) of keeping student records compared to teachers in cluster schools (66%) and satellite schools (42%).

Table 23. Student progress record keeping by treatment

Treatment group	Keeps written student progress record (%)	Observations
Cluster school	66.0	94
Satellite School	42.2	45
Control school	73.8	42
Pooled	61.9	181

This finding indicates that while the majority of teachers do keep progress records of their students, it does not seem to be a uniform policy across schools. This suggests a potential bottleneck in the use of continuous assessment within schools as a lack of student progress records will hinder attempts by teachers to make changes to their teaching based on the results of the assessments in classes. In addition, half of teachers who reported keeping student records only updated them once a term, indicating that when they are used, they are typically used more in a summative style that records a student's progress at the end of a term rather than for continuous assessment and analysing student learning throughout topics. This pattern was similar across each of the treatment groups.

LESSON STRUCTURE

Continuous assessment includes both spontaneous, flexible assessments alongside structured assignments that require preparation. Regardless of which style an assessment takes, they should always be clearly linked to the aims of the lesson and ultimately the competencies being assessed. The only way that a teacher can ensure that this is the case is to prepare clear and structured lesson plans prior to the class taking place that outlines the aims and activities of the lesson. Almost all (91%) teachers surveyed indicated that they developed a written lesson plan for each

lesson. For those who did not develop a lesson plan for each lesson, the most common reason was that it took too much time to do so.

Time constraints appear to be a clear challenge for teachers in developed written lesson plans, with 44% of surveyed teachers reporting that they regularly find that they do not feel they have enough time to make a satisfactory lesson plan and on average each lesson plan took 20 minutes and required oversight from senior colleagues. For instance, a FTT in Oromia reported in their interview that:

> *One thing which makes preparing lesson plan daunting is the fact that it should be prepared daily. On top of that, many detailed information should be included in the plan. This takes up much of our time. I feel the form should be designed in simpler and clearer way.* (Oromia9)

Though, as described above, almost all teachers surveyed reported that they developed written lesson plans for each lesson, during classroom observations only around one-third (32%) of teachers could produce a written lesson plan to show the observer. Teachers in cluster schools were most frequently able to produce lesson plans for the observed class

Table 24. Observed lessons with lesson plan by treatment group

Treatment group	Observed lesson had lesson plan (%)	Observations
Cluster school	44.4	94
Satellite School	13.0	45
Control school	27.3	42
Pooled	32.2	181

Lesson plans may take many forms, but effective lesson plans, using a behaviourist model, will typically cover the 3 A's – activating learning, acquiring new skills and applying or assessing knowledge. To measure the quality of a lesson plan, the following criteria were used as a checklist for lessons plans reviewed during classroom observations:

Table 25. Lesson plan checklist

Planning Areas	Checklist Item
Activate	Clear lesson topic
	Rationale behind the topic
	Prior knowledge of students
Acquire	Teacher activities
	Student activities
Assess	Competency Indicators
	Assessment tools
	Supplementary materials included

When investigating what was typically included in a lesson plan, topic of lesson and teacher activities were reported by almost all surveyed teachers (91% and 95%, respectively). Unsurprisingly, given the relatively low use of supplementary materials in the classroom, the least

common section was a list of supplementary materials. In addition, around three-quarter (78%) of lesson plans included clear indicators of competencies and 71% had a list of assessment tools to use. The qualitative study found evidence that the ToT training had led teachers to carefully consider the activities of themselves and their students. According to a FTT from Oromia, the content of the lesson plan has changed since the AfL training to reflect this:

> *We have been preparing both daily and weekly lesson plan even before the training. The difference is the contents of the lesson plan. Student activities and teachers' activities are separately put in the lesson plan after the training.* (Oromia16)

Through aggregating the checklist items outlined above (scoring no lesson plan as zero), we can have a rough measure of lesson plan quality, with a higher score indicating a higher quality lesson plan. On average, lesson plans were of a much higher quality (2.8 out of 8) in cluster schools when compared to satellite (0.8) and control (1.6) schools.

Table 26. Lesson plan quality score

Treatment group	Lesson plan quality score	Std. Deviations	Observations
Cluster school	2.8	3.3	90
Satellite School	0.8	2.2	46
Control school	1.6	2.8	44
Pooled	2.0	3.0	180

LEARNING INTRODUCTIONS

In addition to lesson planning, it is important for teachers to outline the aims of the lesson by introducing them to students at the beginning of a class (Levine et al., 2008). Students that know what they are expected to learn can focus attention on those areas, increase student engagement (Armbruster et al., 2009) and improve the holistic education experience by giving learners a sense of purpose of education (Reed, 2012). The use of learning introductions was measured using classroom observations and recording whether the teacher took time to formally introduce the lesson to students, make the learning objectives of the lesson clear and whether they explicitly linked to the MLCs identified in the lesson plan. In the large majority of classes (88%), the teacher took time to formally introduce the lesson to students and in 80% of classes the teacher informed the students about the learning aims of the lesson. Despite this, only 65% of classes had introductions that included explicit links to MLCs. Classes observed in cluster schools had a slightly higher rate of explicitly linking to MLCs (67%) compared to 61% in control schools and satellite schools.

The qualitative data analysis revealed that the focus on MLCs was perceived to be a key contribution of the AfL training; some teachers may have only learned about MLCs during the ToT or school/cluster training:

> *In fact, continuous assessment is not new thing I learned from AfL. I have been using it for many years now. Rather, a new concept I learned from the AfL is the MLC. Previously, we would use objectives instead of MLC. Now, MLC is used to rank students based on the level of their competency.* (Oromia9)

I even don't know the term minimum learning competencies before I took the school training. It is only after the school training that I have started to practice the minimum learning competencies. Generally, teaching gradually became a student centred one. (Oromia19)

TIME ON TASK: ASSESSMENT ACTIVITIES, TEACHER ENGAGEMENT AND STUDENT ENGAGEMENT

A teacher that is implementing continuous assessment effectively should have a classroom environment and lessons that are different to those who do not. Whilst it is a significant challenge to measure continuous assessment use quantifiably, we identified common themes of classrooms that used continuous assessment. Firstly, greater emphasis should be placed on spending time assessing students' knowledge throughout the class as opposed to 'chalk and talk' traditional approach of teachers simply dictating to students and/or students simply listening or copying.

For teachers that implemented continuous assessments effectively in their classroom we would expect greater levels of active assessments taking place than in a classroom without continuous assessment. Active assessments are considered to be teaching through oral questioning, in-class discussions, setting class assignments or any other tasks such as role-play or group tasks rather than practice, drilling through repetition or the teacher simply reading out loud from a textbook.

To capture these elements, enumerators used an adapted version of the Stallings classroom observation tool (Stallings, 1977) when conducting classroom observations. This provided 'snapshots' of the classrooms observed at regular intervals (typically every four minutes). The tool captures student and teacher activities and a distinction was made between active assessment activities (oral questions, discussions, group work and class assignments) and other on-task learning activities (e.g. teacher reading out loud, students copying from the blackboard or teachers using repetition drills). If the teacher was engaging in non-academic activities, this was split into either classroom management (e.g. discipline or handing out textbooks) or off-task if the teacher was not present, or partaking in activities that were neither academic or related to classroom management. The data showed that cluster school classrooms typically had a larger proportion of their time in active assessments (40% of the total class time) compared with just 34% at satellite schools and 36% in control schools.

Student engagement was also used as an indicator of continuous assessment. The teacher was considered to have student engagement if at least a large group or all of the students in a class were engaged with the activity the teacher was involved in. In addition, teacher interaction was measured by considering the proportion of time that teachers were interacting with students in their teaching by directly asking them questions or being involved with academic discussions. Table 27 reports the proportion of the proportion of class time where students and teachers were engaged with each other.

Table 27. Classroom snapshots and student engagement and teacher interaction

Treatment group	Proportion of Class Time (%)	
	Student engagement	Teacher interaction
Cluster school	77.7	29.7
Satellite School	72.8	24.3
Control school	82.0	24.3
Pooled	77.5	26.7

PROGRAMME EFFECTS ON TEACHER PRACTICE

Whilst the above analysis has outlined descriptive statistics of teaching practices, any observed difference between different school types should not been interpreted as causal. Hence, to capture the effect of the AfL programme on improving participating teachers' knowledge, we have to compare teaching practice outcomes outlined below for teachers from cluster and satellite schools against teachers from schools from the matched control schools. In addition, we will control for observable characteristics that may influence teaching practice as well.

For in-lesson teaching practices indicators of continuous assessment, we use the proportion of time observed in classrooms to fall into the categories of (1) Active Assessment, (2) Student Engagement, and (3) Teacher Interaction. Table 28 shows the estimated coefficients for the cluster and satellite schools against the control schools from the OLS regression with clustered standard errors at school level.

Table 28. AfL treatment effects on teacher practice indicators

Treatment Group	Estimated Treatment Effect		
	(1) Active assessment	(2) Student engagement	(3) Teacher interaction
Cluster school	0.066* (0.034)	-0.030 (0.040)	0.072* (0.037)
Satellite school	-0.008 (0.042)	-0.077* (0.045)	0.030 (0.044)
Control mean	35.5	82.0	24.3
Total Observations	177	177	177
Adjusted R-squared	0.12	0.11	0.12

Notes: Standard errors in parentheses; Regressions clustered at the school level; * p<0.1, ** p<0.05, *** p<0.01

Controls include the number of students in a classroom and a set of school characteristics.

The results indicate that the AfL programme at cluster schools had a positive effect on teachers using active assessments in their classrooms with a magnitude of 6.6 percentage points (p<0.1), a relatively large effect given it represents an 19% increase on the average time spent in the classroom on assessment in control schools. A similar effect is seen on teacher interaction with an estimated treatment effect of 7.2 percentage points (p<0.1), a 30% increase on the average time the teacher spent interacting with students compared with control schools. These findings indicate, though that the AfL programme was successful in providing teachers with the required tools to help teachers move away from the traditional practice of teachers dictating knowledge to students and replace it with regular interactions and continuous assessment of students learning throughout a lesson. There is no evidence of any positive treatment effect on satellite schools and indeed a negative relationship with student engagement by 7.7 percentage points, which translates as a decrease of 10% compared to the control schools.

These findings were also partially supported by the qualitative interviews, where teachers in several schools discussed moving towards a "student-centred" approach in their teaching practice:

> *First I was using teacher-centred teaching technique. I start it by myself and I finish it, but now I speak and the students also have the chance to speak […] especially for the kids they need more to practice by themselves.* (Tigray2)

Our former teaching way is one-directional, but now it become to student-centred teaching system […] in the previous mechanism everything is expected from the teacher but now it becomes a two-directional mechanism which gives chance of participation for student too. (Oromia27)

As it is clear that teacher use of continuous assessment can only be partly captured through in-class time-on-task measures, in addition, we also constructed outcome indicators for teachers practicing continuous assessment at their schools through the use of principal component analysis (PCA). As continuous assessment contains various dimensions, we can use PCA as a tool to overcome the problem of dimensionality and reduce a large set of variables that are correlated and represent an underlying trait to smaller set, whilst retaining most of the information from the large set of variables. Four components were identified for the teacher's use of continuous assessment: 1) Teacher use of assessment and feedback, 2) Reported teaching structure of continuous assessment, 3) Observed structure of continuous assessment in lessons and finally 4) Parental engagement with their child's learning progress. Table 29 outlines the dimensions of each component indicator and what is included the PCA.

Table 29. Component indicators of continuous assessment use

Component indicator	Component indicator dimensions
Reported Use of assessment and feedback	Teacher reported: • Range of assessments used • Rate of each assessments used • Range of feedback types used
Reported Continuous Assessment Structure	Teacher reported: • Use of written lesson plans • Use of written student progress records • Use of supplementary materials in class • Use of student progress records • Rate of update of student progress records • Sharing of student progress records between other teachers and parents.
Observed Continuous Assessment Structure	Teacher observed: • Use of written lesson plans • Quality of lesson plans • Use of supplementary materials in class • Introduction of class and learning aims to students • Explicit links to MLCs of assessments • Teacher coaching students at an individual level

Parental engagement with their child's learning progress	Teacher reported:
	• Rate of communication with parents:
	  ○ Low performing students
	  ○ Medium performing students
	  ○ High performing students
	• Rate of physical meetings with parents:
	  ○ Low performing students
	  ○ Medium performing students
	  ○ High performing students
	• Breadth of feedback provided to all parents (i.e. content of feedback includes student performance, areas for improvement, pathways for improvement).

The first component can be interpreted as the use of assessments and feedback in lessons and is based on the range of different assessment and feedback methods and regularity of assessment

that is self-reported by teachers. The second component and third components relate to the structure of continuous assessment in their teaching, that is how they use various tools from the AfL programme such as lesson plans, student progress records and inclusion of supplementary materials to enhance the provision of continuous assessment. One key difference is that the second component uses self-reported data whereas the third component relies on information from classroom observations.

The fourth and final component related to involvement of parents – though not always strictly typically considered part of continuous assessment, it was a key component that was identified as part of the AfL programme and it overlaps substantially with continuous assessment as it requires regular assessment of a student's learning and development of feedback to involve parents throughout a student's education. As we did not interview parent's, we will rely on teacher and school reported involvement with parents at an institutional level to measure parental engagement. The results for the analysis on this component are reported in Section 4.3.3 with a fuller discussion surrounding parental engagement. Table 30 shows the estimated coefficients for the cluster and satellite schools against the control schools from the OLS regression.

Table 30. AfL treatment effects on teacher practice components

Treatment Group	Estimated Treatment Effect		
	(1)[a] Reported use of assessment and feedback (PCA)	(2)[a] Reported continuous assessment structure (PCA)	(3)[b] Observed continuous assessment structure (PCA)
Cluster school	-0.223 (0.277)	-0.281 (0.391)	0.855*** (0.314)
Satellite school	-0.309 (0.354)	-1.119** (0.422)	0.126 (0.314)
Total Observations	169	169	177
Adjusted R-squared	0.08	0.13	0.25

Notes: Standard errors in parentheses; Regressions clustered at the school level; * p<0.1, ** p<0.05, *** p<0.01

[a]Controls include a set of individual teacher characteristics and a set of school characteristics.

[b]Controls include the number of students in a classroom and a set of school characteristics.

Interpreting the magnitude of coefficients of variables computed through PCA is less clear due the nature of dimension reduction, though the results from column 1 and column 2 suggest that there is no evidence that AfL has led to an increase in the use of assessment and feedback within a teacher's classroom nor on their practices on structuring their teaching based on self-reporting from teachers. However, when looking at a composite score for teaching structure based on classroom observations, we see a strong, positive effect of AfL for cluster schools (0.86 SD, p<0.01) and no effect on semi-treatment satellite schools. It is not so clear why cluster school teachers would conduct greater levels of structuring their teaching using continuous assessment techniques, but not report greater levels. One possible way to reconcile these findings is that as the measure of all the dimensions do not perfectly overlap (for example, observed continuous assessment structure did not include review of student learning progress records), this is what causes the difference in outcome. However, even in area where there was overlap – lesson planning for example – we saw from the descriptive analysis that there was a gap in the rate of those saying they always prepared a lesson plan and those being able to present them on request during a lesson observation. It is possible teachers may all overestimate the level and quality of

the continuous assessment techniques they implement in their schools on reporting but only cluster school teachers were found to demonstrate it in their teaching.

Analysis of the qualitative data supports the finding that the AfL programme has made a difference to teacher practice. While across the sampled regions, there were school and non-school informants who were sceptical about whether the AfL training had introduced anything new, overall, interviewees expressed that the programme had led them to change several of their practices, including: basing their teaching and assessment on MLCs; using daily, formative assessments and recording students' results; providing constructive feedback; and being more attentive to students' needs in their teaching. The quotes below from various interviewees provide a snapshot of these perceptions:

> *I started implementing assessment for learning after I took training. The difference is, for continuous assessment we ask any question that suddenly comes to our mind. But in assessment for learning questions should not be out of students learning competency. We have already prepared questions on the form. It is well organised.* (Oromia16)

> *In the previous time, we were following students but it was general. Additionally, there was no identification based on knowledge attitude, and skill. Student evaluations were not conducted based on the above-mentioned domains, we simply evaluate out of the given total points, if the test is given out of five points, we evaluate the maximum score not based on the domains. We started this after training. This is very important, compared to the previous approach.* (Oromia25)

> *I give feedback to the students. Before I had no tolerance but now I have good relationship with the students […] There is a progress in my teaching techniques relative to when I begin. At first I encourage the students who answer and insult those who do not but now I do not insult students because I know I have to treat the students equally.* (Tigray1)

> *Yes they change my system highly, for example I'm not giving attention for an individual student previously, I'm just giving lecture and out from class, but nowadays I become more familiar with my student to understand each of their potential and limitation, this helps me to use appropriate and feasible means of teaching that fit my student capacity.* (Oromia28)

Overall, the quantitative and qualitative results provide some evidence that the AfL programme improves teacher practice on continuous assessment. The improvement is both in terms of providing teachers with the required knowledge on the various areas of continuous assessment and in helping to actually change teacher practice within the classroom. Changes in teacher practice is a crucial step in the ToC, as any impact on student learning outcomes will come through this pathway.

4.3.3 PARENTAL ENGAGEMENT

School staff (teachers and school directors) reported trying to involve parents in various ways. For instance, many of the schools included in the qualitative sample held regular conferences or meetings with parents between one and four times per year (although some schools did not have regular meetings in place). This is supported by the quantitative that showed that teachers had face-to-face meeting with parents, on average between once a term and once a year. Exceptionally, the four schools visited in Tigray reported holding at least monthly meetings between parents and the student's homeroom teacher.

The purpose of these meetings was to increase the involvement of parents in their children's education. One school in Benishangul-Gumuz reported 'formalising' parents' involvement to a greater degree by including them in the development of a plan for their child every semester. According to one teacher:

> *Usually we involve parents when we develop plan every semester. We plan by involving parent and student […] There is evaluation of the plan with parents at the end of the semester. During planning, everyone will know its role and responsibility, what is expected from parent, teacher, student, and the school for the success of the child. It will be read for everyone and put their signature.* (BeGu6)

School staff also reported that parents could be invited to the school in case further discussions were deemed necessary. For the schools that did not hold regular meetings with parents, formal letters of invitation to parents from school directors were the main way to engage parents. Based on the qualitative interviews with school staff, across regions, common reasons for these interactions with parents included disciplinary issues, absenteeism, and poor performance of their child in school.

Several teachers and school directors noted that these meetings with parents could be used to try to understand and resolve underlying issues that might be responsible for the student's behaviour or performance. According to school staff and parents, teachers could also discuss how the parents could support their child:

> *Moreover, we will call parents and evaluate the score of their children sitting together. We will show parents their child result in all subject and discuss the reasons and on the solutions especially if their child scored lower grade. There are some teachers who discuss with parents on how parents can help the child at home.* (BeGu5)

However, in Oromia and Benishangul-Gumuz, parents' actual engagement in either the regular or ad-hoc, invitational meetings was variable, according to teacher and school director informants as well as parents who participated in focus groups. While some parents were more involved, other were less responsive. One teacher suggested that parents whose children were performing well in school were more likely to have higher involvement, compared to those whose children were performing less well, although the direction of effects is unclear:

> *[The annual meeting involving parents is] only in June, some of us [go] if we want* (Parent in Oromia26)

> *There is problem in parental engagement in the area. Let alone for the normal discussion, parents are not willing to come when asked to discuss about the serious discipline problem of their own child. They don't want to come to school whether you sent them letter or not.* (BeGu10)

> *But, sometimes parents whose children have fallen behind do not come to school even when they are summoned. On the other hand, parents whose children are high achievers come to school without being summoned.* (Oromia9)

In addition, parents reported that they did not generally tend to initiate contact with schools:

> *There is no such practice [of proactively visiting the school] indeed. Further, the teachers will not be welcoming if we try to do that. The behaviour hasn't been practiced. The teachers may*

wonder about my intention if I go and ask them how the teaching-learning process is going. (Parent in Oromia2)

We contact with teachers when our children are asked to come with parents […] Otherwise, there is no way that we contact with teachers. (Parent in BeGu26)

Although there was some evidence that parents were interested to be more involved in their children's learning at school, some parents and school staff also said that parents may not have time to do so. School staff, additionally, asserted that parents were not interested in engaging with them. Some felt that parental engagement was particularly low in rural areas, potentially because parents had to travel for work:

Students' parents' participation in many government schools is very low. Particularly, in rural areas like our school, it is unlikely to get parents participation in learning teaching process. (Oromia16)

The teacher wants to discuss with parents, but parents do not come. They will go far to mine gold. (BeGu13)

As with holding regular meetings with schools and parents, the exception in terms of level of parental engagement appeared to be the schools visited in Tigray. According to several parents and school-based informants, parents were reported to be engaged in their children's education, with the majority of parents attending the scheduled monthly meetings, and some parents proactively making additional visits to the school:

As you can see the village settlement is so far apart but the participation of the parents is good. More than half of the total number of parents attend parental meeting once a month based on the schedule. (Tigray5)

We get in touch with [the homeroom teacher] every month and sometimes we meet with them anytime when I have time to visit we discuss about my children performance and everything and I do according to the recommendation from the teacher. The same is true with the other respondents. (Parent in Tigray24)

Finally, based on interviews and focus groups, it also emerged that some schools had parent-teacher committees. Those informants who brought up these committees noted that their role included evaluating the teaching-learning process and addressing drop-out. It was not clear, however, how widespread or active such committees were.

VIEWS ON PARENTAL ENGAGEMENT

Regardless of the status quo of parental engagement, many respondents (including school staff and parents) across regions believed that parental engagement should increase. In this regard, there appeared to be few differences across regions as well as training type (i.e. whether the informants had received ToT training, school/cluster training or no training).

Several school-based respondents felt that increasing parental engagement was one important mechanism for improving children's academic performance and other school outcomes such as attendance or motivation. Because school contact time is short, some teachers felt that their influence was limited and that parents had to play a vital, complementary role in their child's education. For instance, a teacher in Benishangul-Gumuz reported that:

> *Children stay in the school only for four hours, so parental thoughtfulness is important and has to be improved more on how to nurture his/her children. Even if he/she is illiterate they can at least ask whether they have learned or attended class today or not.* (BeGu2)

Teachers also felt that parents could provide information about their child that might be helpful for the teacher, or to shed light on underlying obstacles to the child's school performance. A teacher in Oromia, for instance, stated that:

> *It is very important to know their child's performance, know child's problem and identify the gap whether the problem is with the teacher or the student regarding student result.* (Oromia16)

Raising awareness of the importance of parental engagement was cited by school-based informants as a way to increase parents' engagement.

> *I think parental engagement can be increased by heightening their awareness about importance of their regular involvement in helping their children improve their academic performance.* (Oromia6)

One way to raise awareness, as suggested by a teacher in Benishangul-Gumuz, was to do so during kebele meetings and other public gatherings. Indeed, the kebele office was seen to have greater influence on parents than the school: a few teachers in Oromia also suggested leveraging kebele officials to reach parents, or to arrange for meetings to take place at the kebele office:

> *The school has to raise the awareness of parents during kebele meeting or other public gathering to come to school when they are invited to come, since parental engagement is important.* (BeGu17)

> *School board and committee can be involved in heightening the awareness of parents to increase their involvement. Even kebele development team can involve in this activity. I hope parental engagement can be augmented in this fashion.* (Oromia6)

PROGRAMME EFFECTS ON PARENTAL ENGAGEMENT

As part of the quantitative analysis, and as discussed in the previous section, one component of teacher practice relating to continuous assessment is the involvement of parents with their students learning. The components of the PCA for the parental engagement indicator are shown in Table 29 and include the rate of communication and physical meeting with parents of students of all abilities and the level of feedback provided to parents. Table 31 shows the estimated coefficients for the cluster and satellite schools against the control schools from the OLS regression.

Table 31. AfL treatment effects on parental engagement components

Treatment Group	Estimated Treatment Effect
	Reported Parental Engagement (PCA)
Cluster school	1.383*** (0.370)
Satellite school	0.915** (0.435)
Total Observations	169
Adjusted R-squared	0.19

Notes: Standard errors in parentheses; Regressions clustered at the school level; * p<0.1, ** p<0.05, *** p<0.01

Controls include a set of individual teacher characteristics and a set of school characteristics.

There is a strong and positive effect on the parental engagement component on cluster schools of 1.38 SD (p<0.01) and satellite schools of 0.92 SD (p<0.05). This suggests that the AfL programme was successful in encouraging and promoting parental engagement in Oromia. Looking beyond Oromia, one finding from the qualitative data was that parental engagement appeared to be higher in Tigray than in the other two regions included in our evaluation. Anecdotally, two teachers interviewed in Benishangul-Gumuz reported that parental engagement had increased in the past year, though this was not raised by other informants in the school. Overall, there were no clear differences in level of parental engagement between school staff who had received ToT training and school/cluster training, and the qualitative sampling approach does not allow comparisons between schools (unlike in the quantitative approach).

Based on focus groups with parents across regions, it also emerged that parents had not been explicitly informed about AfL. A school director in Benishangul-Gumuz also reported that details of the AfL programme had not been shared with parents and the community. Parents did not mention changes in how schools engaged with them since the introduction of the AfL programme, and, as they were not aware of it, would not have been able to attribute any changes to the programme.

However, if contact with parents through meetings and invitation letters can be attributed to the AfL training, then it seems reasonable that the programme may have been responsible for at least some of the parental engagement reported by informants. Based on responses from some ToT-trained teachers, AfL training may have changed the frequency of school-initiated contact with parents, as well as the content of these interactions:

> *[After the ToT training] I was expected to […] involve parents to school.* (BeGu14)

> *The trainers have educated us about parental engagement. We used to think that parents are involved when their children misbehave in school. The trainers taught us that this is not only the time we need to involve the parents. They should be involved if their children perform wonderful or poor. They should be asked about the conditions of their children if they perform poorly. They need to improve the household conditions of their children if the conditions are hampering their education. They will be advised to support the children harder. The parents will share their experiences with the teachers if their children are performing outstanding. And the teachers will utilise the experiences to advise other parents.* (Oromia3)

4.3.4 STUDENT OUTCOMES

As per ToC, improvement of student learning outcomes is the final step in the causal chain that should be achieved through schools and teachers changing their teaching practices and hence providing a higher level of educational quality for students. In Chapters 4.3.1 and 4.3.2 we examined changes in teacher practice and found promising evidence in the schools that tool part of the AfL programme teacher knowledge and practice of continuous assessment has been improved. To measure the effect on student learning outcomes, we focussed on test scores in Afan Oromo, English, Mathematics and Environmental Science as our primary outcomes. These subjects were selected as they are the focal subjects for the AfL programme. The tools used to measure student learning were low-stakes tests – that is, students were informed that these tests would have no impact on their grades or educational attainment overall - that students were randomly selected to take part in.

Table 32 and

Table 33 reported the average raw scores on each of the subjects in Grades 3 and 4, respectively and broken down by treatment group.

Table 32. Raw student test scores by subject in Grade 3

Subject	Cluster school	Satellite school	Control	Observations
Afan Oromo (%)	39.1 (28.3)	30.3 (26.6)	34.1 (28.6)	584
English (%)	29.8 (22.9)	25.6 (18.8)	26.2 (24.9)	584
Mathematics (%)	50.0 (32.1)	36.8 (30.0)	40.3 (29.9)	584
Environmental science (%)	40.5 (31.8)	31.8 (28.9)	33.7 (27.4)	584

Table 33. Raw student test scores by subject in Grade 4

Subject	Cluster school	Satellite school	Control	Observations
Afan Oromo (%)	43.5 (27.1)	36.7 (23.7)	35.4 (22.7)	577
English (%)	43.3 (25.0)	36.7 (24.1)	35.5 (24.8)	577
Mathematics (%)	56.2 (29.8)	53.5 (29.6)	47.6 (29.3)	577
Environmental science (%)	43.9 (28.8)	38.3 (28.0)	37.6 (25.3)	577

Notes: Standard deviations in parentheses.

Though students in cluster schools outperform students in satellite and control schools in each subject and grade combination, this may be due to observable and unobservable differences between the schools. In the following section we attempt to account for any observable differences between the schools and capture the programme effect on student outcomes.

PROGRAMME EFFECTS ON STUDENT ACADEMIC OUTCOMES

To compare student outcomes and attempt to capture the effects of the AfL programme, firstly we need to ensure comparability between grade and subject scores on the tests. To do this, we standardised the total scores for each grade-subject combination and then aggregated scores across grades and re-standardised to yield the overall subject standardised test scores for a given

subject. For the analysis, we jointly estimated the effects of both cluster schools and satellite schools[7] in a pooled regression and present estimates for all interventions together in Table 34.

Table 34. AfL treatment effects on student learning outcomes

Subject	Cluster school	Satellite school	Observations
Afan Oromo	0.296* (0.168)	-0.059 (0.172)	1,161
English	0.257 (0.204)	0.029 (0.184)	1,161
Mathematics	0.300** (0.140)	0.008 (0.145)	1,161
Environmental science	0.280 (0.169)	0.069 (0.189)	1,161

Regressions clustered at the school level; * p<0.1, ** p<0.05, *** p<0.01

Controls include student characteristics, school characteristics such as school inspection score, urban dummy, average class size, teacher turnover rate and total enrolment in Grades 1-4, proportion of female teachers, average staff teaching experience, average teacher education and dummy variables for each zone.

The results indicate that the AfL programme at cluster schools had a positive effect on student learning outcomes for Mathematics with an increase of 0.3 SD ($p<0.05$) and Afan Oromo ($p<0.1$) with a magnitude of 0.3 SD as well. There was no observed semi-treatment effect on satellite schools in any of the subjects. This is, again, unsurprising as the prior findings showed that teachers in the sampled satellite schools in the sample rarely benefitted from cluster training in their areas.

These quantitative findings are further validated by the qualitative analysis. School teachers and school directors who participated in qualitative interviews were generally optimistic that AfL had contributed to improvements in students' academic outcomes. While some interviewees reported that improvements had not been universal among students and that improvements varied between schools, others felt that the changes introduced by AfL, such as focussing teaching on MLCs and to helping students achieve them, and facilitating greater student participation and engagement in lessons had generally helped improve students' academic performance, compared to the previous practice of less regular (weekly or monthly) assessments. Some interviewees believed that fuller implementation might lead to greater effects being seen.

Even though parents across regions did not explicitly know about the AfL programme, there was some indication of their awareness of a change in teaching practice. Particularly, one parent participating in a focus group in Benishangul-Gumuz noted that:

> *The change I see from teachers' teaching methods is that students are asked by teachers whether they have understood what they learned or not and even children are given time to ask and reflect. We parents can easily notice that if our child is being changed by what they are learned in the class. We can understand if our children's knowledge is growing from time to time. Previously, teachers were not explaining subjects being discussed very well for children. Children may have learned many points but has no idea about that course when*

[7] For all regression analyses in this report, we use control schools as the base treatment group so all estimates are comparisons with the control schools.

asked. The way teachers teach matters to help children understood things. So it is now much better than ever for children. (Parent in BeGu25)

School staff who felt that AfL could be credited for positive changes in students' outcomes pointed to differences between current and previous cohorts or to reported differences between the grades of students whose teachers were using AfL techniques compared to those who were not (though these comparisons cannot be considered a counterfactual, as was constructed in our qualitative analysis). Such examples had also been related to the Oromia CTE staff member interviewed:

> *Reportedly, [the teachers] say that the competency of their students has been improved since they began to use AfL approach. In previous times, even students reach 5th and 6th grades before learning how to read and write. This has been reversed and students become able to read and write earlier because of AfL […] They report there are such changes despite the challenges of big class size, low teachers' motivation and shortage of teaching-learning materials.* (Oromia20)

It should be noted that a small number of respondents were more critical, noting that it was not possible without further investigation to attribute the changes to AfL:

> *There are some effects seen, but not satisfactory, due to time limitation. If we really get time and implement it fully, it will bring better effect. There is also problem in telling this change or effect is AfL-specific.* (BeGu18)

It may be difficult to attribute changes to AfL because of its complexity. As the programme has several components and potential mechanisms for effecting change, several of which do not necessarily represent a drastic change from previous practice, any changes in school and teaching practices, as well as student outcomes, may not be obviously linked to AfL. For instance, the Oromia cluster supervisor interviewed was sceptical that any positive changes could be attributed to AfL:

> *There have been changes […] I and the school directors of the 5 schools come together and make evaluations every month. We may notice improvements in some aspects of the education but the school directors have never mentioned that AfL programme is contributing to the change. You can check it out in our minutes that have been recorded on our meetings. The name of your program has never been mentioned. We explain the improvements from the angles of teachers' hard work, strong supervision and parental engagement.* (Oromia21)

NON-ACADEMIC OUTCOMES

Qualitative informants also reported that they believed that AfL had contributed to a range of non-academic outcomes for students.

Across regions, one of the more frequent outcomes raised by interviewees was that students' participation in class and engagement in their own learning had increased. Teachers also reported that students were more motivated. Among interviewees' proposed explanations were that AfL had facilitated a shift towards a student-centred teaching approach, that the assessment methods in the AfL package encouraged teachers to target all students in the class, rather than a small subgroup (of, for instance, more active students), and that constructive feedback meant that students were more willing to speak up in class.

Debate and dialogue are implemented in the student-centred teaching-learning process. This has contributed for boosting the confidence of the students for their engagement is enhanced by the programme. We can say that AfL has contributed much in this regard. (Oromia24)

All the mentioned AfL concepts made them to prepare themselves or to read in advance as it [the continuous assessment] made them believe that each of them would be asked in a class. It also creates an initiation of learning on them. Besides majority of the students were not attentive prior to the implementation of AfL but then after their engagement has also been changed tremendously. (Oromia19)

But now even if the student doesn't answer the correct answer still encouragement there. As a result the effect that I notice from this was students are encouraged to participate in the class they don't afraid whither they give the correct answer or not. (Tigray15)

Teachers and school directors also interviewed also reported seeing improvements in students' self-confidence because of their increased participation in class:

Just to share you my personal observation, there is a student who was so timid and lacked assertiveness. Yet, now as a result of AfL, he has totally become assertive and self-confident. (Oromia7)

Some school staff in Oromia reported believing that AfL had increased attendance and reduced absenteeism of students. Although some interviewees felt that "there is no observable relationship between the AfL and attendance/absence of the students" (Oromia24), others believed that students' increased engagement in school meant that they were less likely to absent themselves or play truant. In addition, because teachers would note and follow-up on absent students, underlying reasons for absenteeism could also be addressed:

[AfL] contributes to [improving attendance] for the reason that teachers pay attention to the life conditions of the students and they have the opportunity to identify the conditions that compel the students to miss class. We work on the challenges and will rescue the students not to miss class. (Oromia4)

We have a daily assessment and follow up at the individual level. So, we can identify who is absent and who is attending class consistently at an early stage, students who are missing class for no reason will improve their attendance. Additionally, we can address parents to drop out. (Oromia25)

It should be noted that even though this finding was not reported by informants in the other regions, the research design does not allow us to conclude that absenteeism was not also reduced in Benishangul-Gumuz and Tigray.

School-based informants' perceptions on the effect of AfL on dropout rates were mixed, with some noting that AfL had reduced dropouts while others felt that AfL had had no impact on dropouts.

Very few interviewees reported that in their view, AfL had had detrimental effects on non-academic student outcomes. Only two teachers interviewed noted that some students may be overwhelmed by the daily continuous assessment.

GENDER DIFFERENCES

Generally, across regions, interviewed respondents reported that they did not see or expect gender differences in children's outcomes as a result of AfL. This view was expressed both by interviewees who worked in schools (teachers, school directors) as well as non-school interviewees such as WEO staff and ToT trainers. Some interviewees pointed out that as a programme, AfL does not treat girls and boys differently. For example, a ToT trainer in Tigray noted that "I can't have an accurate answer to who AFL benefits more because we never see it in that point of view" (Tigray12).

However, there was also anecdotal evidence that girls may be more affected by AfL. It was suggested by some teachers that the programme would positively impact students whose motivation had increased as a result of AfL. Because girls reportedly tend to be more passive than boys, some teachers felt that they now put more effort into engaging female students. Subsequently teachers' increased attention may have led to girls' increased engagement and participation in class, and may have also had some effect on other outcomes such as attendance and dropout:

> *It enhances all students to participate actively. And it affects girls more by increasing their participation. Especially a continuous assessment or the follow up helps girls more because girls are shy and do not participate well in the class if there is no follow up. The follow up in turn helps me to identify their performance and do follow up again.* (Oromia1)

> *Previously girls were dropping out from school, but now girls' dropout has reduced. They even do not absent from school without permission. This shows that they their interest in education has improved. If the child wants to be absent from school, parents will come and ask permission. I may give one day permission after I evaluate their parents' problem, I will not give more than one day because, if I give them three days, they will make it one week. Boys are doing fine from the beginning.* (BeGu20)

4.4 SUSTAINABILITY OF THE PROGRAMME

In the following section, we use examine potential avenues that would allow for the AfL programme to become more sustainable in the future. Coupled in with this, is the assessment of the level that AfL has become embedded within the education system and the perceptions and buy-in of policy makers as the programme moves forward.

4.4.1 PROMISING AVENUES OF INTERIM SCALE-UP

There have been various approaches in Oromia to scale up the programme, including the integration of AfL into pre-deployment training and summer in-service training. Furthermore, ELIXIR and UNICEF have been undertaking efforts to implement the programme into the World Bank-funded nationwide GEQIP-E programme as part of the Continuous Professional Development with the MoE.

PRE-DEPLOYMENT TRAINING, IN-SERVICE SUMMER TRAINING AND CONTINUOUS PROFESSIONAL DEVELOPMENT

According to informants at ELIXIR, relying on the cascading knowledge model only reaches about 500 teachers per year in Oromia. One avenue to scale up the programme that is already being practiced in Oromia is the integration of AfL into pre-deployment training and in-service summer

training. Between finishing their CTE coursework and final graduation, CTE would-be teachers are expected to take a three-day pre-deployment training on a specific theme (such as gender equity, teaching ethics, HIV/AIDs education or positive classroom discipline). As the REB expert in Oromia explained, via integrating AfL into pre-deployment training they managed to reach about 37,000 teachers within the past two academic years. It should be noted that the pre-deployment training does not replicate the intensity of ToT training; however, it is around the length of an average training cycle for school or cluster training as found in the quantitative survey.

The integration of AfL into the Continuous Professional Development programme (CPD) from the MoE presents a further opportunity for scale-up. Each teacher in Ethiopia is required to attend at least 60 hours of CPD-related activities each year. As ELIXIR points out, *"there are other projects as well like gender sensitive pedagogy like child discipline and the like so we are not telling them all the time to do this but at least to give training as much as possible [… on] AfL"* (KII1).

Moreover, in Oromia AfL is also being offered as part of the in-service summer upgrading, where in-service teachers take courses at CTEs during the summer every four to five years to upgrade their certification (KII1).

GENERAL EDUCATION QUALITY IMPROVEMENT PROGRAM FOR EQUITY

In cooperation with the Ministry of Education, the World Bank has initiated the General Education Quality Improvement Program for Equity (GEQIP-E), which aims to assist the government in improving general quality in education across Ethiopia. It targets the improvement of the quality of education in all grades including in 0-class (World Bank, 2020). As an effort to scale up the AfL programme, UNICEF and ELIXIR have been working closely together with the World Bank and the Ministry of Education to integrate a module on Continuous Classroom Assessment (CCA) in the GEQIP-E programme. As UNICEF stakeholders explained:

> *Continuous Classroom Assessment that is under the GEQIP programme has benefited a lot from our experience in AfL. The […] teacher training manuals for instance has adapted a lot from the manual that we developed; the tools themselves have also been used to inform the kinds of tools that have been included in the GEQIP-E programme so which goes nationally, nationwide. So we can say yes, AfL has impacted a nationwide kind of continuous assessment programme.* (KII2)

Similarly, ELIXIR sees the incorporation of AfL methods into the CCA module of GEQIP-E as a success in efforts to scale AfL up. Our key informants emphasised that the World Bank programme aims to reach about 50% of all schools across Ethiopia and is a major milestone to guarantee the sustainability of AfL.

EXTENSION TO OTHER GRADES AND SUBJECTS

Integrating higher grades (i.e. Grades 5 to 12) into the AfL programme presents another potential avenue for scale-up. For instance, while the REB expert in Benishangul-Gumuz felt that AfL was most important for the lower grades, he also saw the programme as necessary at all levels of education, including at university. Stakeholders from UNICEF also expressed that the programme could be expended to more grade levels:

> *I think this is a question that actually sits in my head what happens to the children beyond Grade 4? […] the children are getting used to a certain way of doing things and then when they go to grade 5, they look around and it's like nobody is asking us questions.* (KII2)

This idea was also supported by some school directors and teachers in Tigray and Benishangul-Gumuz, who also advocated implementing AfL techniques beyond Grades 1-4. For instance, one cluster-trained teacher argued that:

> *It must include to all from low to high level even so it is important to focus more on the low grade level but we don't have to forget to scale up to the higher grade level.* (Tigray1)

Furthermore, a CTE instructor reported further efforts in Oromia to integrate the subject Aesthetics into the AfL programme, so that AfL tools can be applied in all subjects at primary school level. At time of writing, the Oromia REB and CTEs were working on it and are planning to integrate it soon. According to the CTE instructor:

> *Our intention is to integrate AfL into all subjects of study in order to ensure its sustainability. At the end of the day, all graduates will teach in light of the programme without taking additional on-the-job training.* (Oromia20)

ONLINE DATABASE

Lastly, ELIXIR envisions an online platform for teachers to not only exchange AfL materials and tools but also to share their experience in applying AfL in practice across Ethiopia. This would constitute a major contribution to the ability of practitioners to share their experiences and learn from others. ELIXIR plan to set up this platform in the course of the next year and to make it available to all teachers across Ethiopia as a path to scale up the programme and granting its sustainability.

4.4.2 INTEGRATION OF AFL INTO PRE-SERVICE TRAINING

The core assumption of the cascading knowledge model is that following ToT training, knowledge and material are passed on by ToT-trained teachers and school directors within their schools and clusters. However, as already described earlier, there are several barriers to the cascading of knowledge at the cluster level. Moreover, as the REB focal person in Oromia highlighted, the cascading knowledge model is very resource-intensive and benefits relatively few teachers. Informants at ELIXIR also noted that the cascading knowledge model is currently limited to 22 intervention woredas across Oromia. In order to address these issues, beyond the interim scale-up options of integrating AfL into pre-deployment and in-service summer training, the REB also recently integrated AfL into pre-service teacher training at CTEs. In fact, according to a CTE instructor from Oromia, AfL has become a major pillar of teacher training in CTEs across Oromia:

> *It has become the element of the main college training instead of on-the-job training as a project. The intervention thus is no more a project for we have integrated it into the regular training of the teachers. It has developed into a programme for it has been integrated into the main/regular training of the teachers.* (Oromia20)

Oromia has already trained 37,000 CTE graduates using the integrated AfL module and the long run, ELIXIR informants calculate that 10,000 AfL-trained teachers will graduate from CTEs in Oromia each year, making it considerably more effective than the cascading knowledge model.

During key informant interviews, UNCIEF and ELIXIR informants repeatedly stressed that the Oromia REB had independently developed this avenue to scale up the AfL programme, marking a major achievement in the implementation of AfL.

This shift in policy has not only taken place in Oromia, but upon knowledge sharing between the Oromia and other regions, REBs in Tigray, Amhara and Afar are also working to integrate it into the pre-service training at CTEs at time of writing. The key informant at the Benishangul-Gumuz REB also reported aiming to implement AfL methods into the curriculum at CTEs. For him, integration of AfL into pre-service training was an essential step towards ensuring that all teachers (not just those teaching Grades 1-4) were prepared to implement AfL. Moreover, the current financial cost of the cascading knowledge model was too high to be able to provide all teachers with the necessary training. However, he also stressed that while there are enough human resources at the CTE with substantial knowledge on AfL, they still lacked the financial means to proceed with the incorporation to CTE curricula.

Teachers and school directors across all three regions were strongly in favour of integrating the AfL methods into pre-service teacher training at CTEs.

> *Also it's better if [teachers] get […] AfL as one of the [teacher training] courses to get deep awareness, rather than giving them after they are out from the college.* (Oromia29)

> *As to me, first and foremost the AfL in CTE the training should be in-depth when these things happen teachers become capable of addressing AfL.* (Tigray22)

Any integrated AfL-CTE training should be thorough and sufficiently contextualised. For instance, a ToT-trained teacher from Tigray reported having taken an AfL-related course at college, but not understanding its importance and thus not internalising its approaches. Another ToT-trained teacher from Oromia was also critical of the integration of AfL into pre-service training:

> *I think that the programme is being provided for graduates at CTE as integral part of the regular courses. Delivering the course at that level isn't adequate.* (Oromia3)

4.4.3 EXPERIENCE SHARING

To improve the implementation AfL teachers and school directors in Oromia and Benishangul-Gumuz advocated to increase experience sharing both within and across regions. For instance, a ToT-trained school director from Benishangul-Gumuz supported sharing experiences with counterparts in Oromia because the director felt that the implementation of AfL in Oromia was going better than in Benishangul-Gumuz. An Oromia ToT-trained school director suggested experience sharing at the woreda level not only between schools that implement the programme but also between treatment schools and non-treatment schools.

4.4.4 CHANGES ACHIEVED IN RELATION TO ATTITUDES OF PRACTITIONERS AND POLICY MAKERS

This section described changes in relation to policy and practice arising from the implementation of AfL, include changes in attitudes of stakeholders.

CHANGES AT THE POLICY LEVEL

The AfL programme induced considerable policy change over the past four years. As outlined in Section 4.4.1 it impacted several programmes and policies at the national and regional level, including GEQIP-E, CPD, pre-deployment training, summer in-service training and the integration of AfL to pre-service training.

As UNICEF interviewees pointed out, the project particularly influenced education policies in Oromia, where decision-makers highly valued AfL:

> *[In] Oromia, we had planned a few selected primary schools in the UNICEF supported Woredas but as soon as the buy in had in literal terms say caught fire the region was like wait a minute; this is not only for UNICEF supported Woredas. We want this to go beyond.* (KII2)

Indeed, it was the REB in Oromia that spearheaded the incorporation of AfL into pre-service, pre-deployment and summer-in-service training (see Section 4.4.1. for greater detail). At time of writing, Afar, Amhara and Tigray are now also following this path. ELIXIR informants greatly valued the attained changes in policy:

> *Success is like inclusion of the CTEs. The success is the taking up of assessment for learning as a classroom continuous assessment by the World Bank funding [GEQIP-E], the attempt of including it in the teacher education programme at the ministry level [CPD]. So the fact that it is becoming a policy issue now is a big success.* (KII1)

CHANGES IN PERCEPTIONS OF AFL

The key informants at REBs across all three regions were convinced about the programme, even though their history with AfL methods and tools differed substantially. While the interviewee at the Oromia REB was already convinced of the programme before its implementation, the REB experts interviewed in Benishangul-Gumuz and Tigray only first learned about AfL techniques during the ToT training session they attended. Despite this difference in prior knowledge, all three unanimously underlined the importance of AfL in their respective regions. Indeed, the AfL focal person at the REB in Benishangul-Gumuz, who started his career as a school director, told us:

> *By the way in my previous assignment [as a school director], there were some techniques that I forgot and also ignored to apply. After my introduction with AfL, I have realised that I could have brought a change if I were implementing it earlier.* (BeGu22)

The REB expert from Oromia further described the positive perception of AfL at the REB in Oromia and the desire to scale AfL up into different grade levels:

> *My colleagues at the regional education bureau were neutral prior to the very commencement of the programme, but after they started to see its results […] they start to praise the programme. When we informally discuss about the programme they are very positive about it and even wanted to extend the programme into 5 up to 8 grades.* (Oromia23)

AfL focal persons at WEOs and cluster supervisors in all three regions generally shared positive perceptions about the programme. For instance, the WEO expert from Oromia spontaneously reflected on how their perception on the programme had changed throughout its implementation:

> *I had not perceived AfL as so important when I first heard about it. But after I visited schools and observed the improvement it brings to the teaching-learning process, I really changed my mind and appreciated the project.* (Oromia36)

One assumption within the ToC is that for teachers to implement continuous assessment effectively in their classroom they have to believe it is worthwhile and have a positive view of it (both idealistically and practically). Like other stakeholders, teachers from the three regions reported that their perception on AfL improved throughout the implementation of the programme. Indeed, prior to attending the training, school staff's expectations and views on the programme differed: some teachers and school directors were worried about additional workload, some had no knowledge about it and others were already inclined towards the programme, as colleagues shared their positive experiences with them. Nevertheless, the training positively influenced their perceptions in most cases. For example, a cluster-trained teacher from Oromia reported initially being sceptical about the usefulness of the programme prior to the training but subsequently finding very useful for students. Another school-trained teacher from Tigray told us that while he did not know about AfL before attending training, he was soon convinced and aimed to "become a model to other teachers" (Tigray2). A ToT-trained teacher from Benishangul-Gumuz reflected on this positive change in perception:

> *When I first hear about AfL, I was not clear about what it is. After I started attending the training, especially after about two days later, it became clearer and clearer, then after that I liked it and attended it very well.* (BeGu16)

The quantitative survey in Oromia found very positive results on buy-in from school directors, with 90% of cluster school directors reporting that they felt the AfL programme was either very (37%) or extremely (54%) effective at increasing the quality of teaching at their school. In addition, 77% of cluster school directors reported that they had tried to persuade colleagues at other schools or superiors on the benefits of the AfL programme.

Moving onto continuous assessment on the whole, as part of the quantitative survey in Oromia, teachers were presented with a set of perception questions to gauge buy-in for continuous assessment as a concept and its accompanying techniques. The first set of questions related to the teacher's perception of continuous assessment as a concept and whether it was an ideal strategy for improving education. Teachers, on average, strongly agreed that continuous assessment was a positive thing and was good practice in teaching. There was very little difference between treatment groups which indicated that in general, teachers' buy-in for continuous assessment was not a problem, regardless of which training they had attended.

PERCEPTIONS ON USING AFL TECHNIQUES

Among ToT-trained teachers in Tigray, Benishangul-Gumuz and Oromia there was widespread belief that the AfL has positive effects on the teaching-learning process. Numerous teachers in all three regions highlighted that they found it particularly helpful to categorise students based on their level of achievement. They argued that this helped them to adapt their teaching to diverse knowledge levels in the class and to support low-performing students to catch up with their classmates. For instance, one FTT in Oromia stated that:

> *The AfL approach supports us to focus on students of different achievement statuses, to enable the clever students can support those who fall behind for example.* (Oromia4)

Moreover, some FTTs in Benishangul-Gumuz and Oromia emphasised the usefulness of the daily lesson plan. Although time intensive to prepare, they highlighted that the plans helped them to better structure their lessons and to guide them in evaluating students on a daily basis. However, different teachers in the two regions also point out that given time and material constraints (not enough copies of the lesson plan for example), it was not feasible for them to implement (see also time constraints and resources at school disposal). One particularly negative teacher from Benishangul-Gumuz worried that AfL would impact negatively on high-performing students:

> *By anticipating students will proceed altogether to the next level, some active and fast learning students are being held back by force to simply wait until the slow learners improve and cope up with them.* (BeGu10)

The majority of cluster- and school-trained teachers interviewed were based in Oromia, while others were from Tigray. These interviewees shared a positive perception on AfL techniques and tools. However, like their ToT-trained counterparts, several in Oromia also found it difficult to implement it in a real world setting with large class sizes and time constraints (see Section 4.1.4).

Similar findings emerged from the quantitative survey. Teachers surveyed in Oromia were also asked about their views on the practicality of implementing continuous assessment. In stark contrast with the perception of it as a concept (presented above), teachers were on average negative about its application and implied that it caused a burden on them. Teachers at AfL cluster schools were similar in their feelings towards the practical implementation of AfL. This gap between the perception on AfL as a best practice concept alongside the perception of it as a burden suggests that if more work can be done to help teachers overcome implementation challenges, continuous assessment can be a more sustainable strategy in the longer term. When asked whether they believed the benefit to learners of using continuous assessment outweighed the extra demands on them as teachers and the school in general, teachers on average agreed (an average rating of 3.9 on a 1-5 Likert scale) and was consistent across treatment groups. This suggests that although the demands remain the same, teachers in schools that are involved in AfL buy in to the benefits of continuous assessment slightly more.

As reported in Section 4.3.3, parents who participated in the focus group did not know about AfL and thus could not share specific views on the programme. Nonetheless, in one focus group discussion in Benishangul-Gumuz, parents indicated that they value if teachers apply teaching methods beyond lecturing, suggested that AfL techniques could be received positively by parents:

> *Children are given assignment to make alphabets or letters by cutting from carton. This is not easy to do for children but it is very important for them to learn to do things easily.* (Parent in BeGu26)

CHAPTER 5: CONCLUSION AND RECOMMENDATIONS

This report documents the endline evaluation findings on the AfL programme in Ethiopia. Through mixed-methods analysis, we examined data collected in February and March 2020 from 68 schools across three regions (Oromia, Tigray and Benishangul-Gumuz) and from various stakeholders involved in the AfL programme. This section summarises the key findings from the evaluation, outlines the main lessons learned, and finally provides recommendations for future programme implementation.

EQ1. WHAT IS THE QUALITY AND RELEVANCE (INCLUDING GENDER EQUITY) OF PROGRAMME INTERVENTIONS (MATERIALS, MODULES, TOOLS AND TRAINING)? HOW CAN THEY BE IMPROVED FOR FUTURE USE?

The AfL package was, on the whole, successful in developing materials for the ToT training programme and reference material for education practitioners. There was a strong focus on ensuring that the materials and languages were contextualised to the specific regions and that feedback was regularly incorporated into revisions. Participants in the ToT training reported high levels of satisfaction on both the content – across all key components of continuous assessment – and facilitation of the training, believing it to be interactive and relevant to them. However, most expressed the feeling that the length of training was too short to fully understand and be prepared to implement AfL.

In addition to the training, participants received AfL packages for each of the selected subjects (Mother tongue, English, mathematics and environmental science) that contained field-tested assessment tools, both formative and summative, that were linked to MLCs, and various templates for structured continuous assessment (e.g. for student progress records). Generally, teachers and school directors felt that the AfL package materials were helpful and relevant to them in implementing continuous assessment in their schools, though many expressed concerns about the availability of such materials.

There were identified issues where the AfL programme design interacts with contextual challenges in the Ethiopian education system. An example of such is large class size, with Oromia for example, having on average 52 students in a class (Rossiter et al, 2017) and in our study some schools reporting over 100 students per class. Some teachers questioned whether this is an appropriate context for continuous assessment. Implementing teaching that is required to be interactive and individual student-focussed in very large classes can lead to teachers being overburdened and was identified by many teachers as one of their biggest challenges in implementing AfL in the classroom. Another issue relating to the wider Ethiopian education context is a high level of teacher turnover at schools. It was often be the case that teachers benefitted from the training and materials, only to move to another school or leave the teaching profession shortly after. Whilst it can be argued that teachers transferring between schools may lead to wider spillover benefits of the programme, it does hamper the ability of schools to sustainably institutionalise the AfL programme as well as leading to less-than-optimally trained teachers (as teachers may leave before attending all training sessions). Finally, many schools reported that their ability to utilise continuous assessment was constrained through a simple lack of basic resources such as stationery and chairs in classrooms.

Although it is a significant challenge to address the contextual issues in the Ethiopian education system such as large class sizes, teacher turnover and school resources, and indeed outside the

scope of this evaluation, there are some promising avenues of practice to at least limit how they interact with the AfL programme and continuous assessment. The effect of high levels of teacher turnover can be largely addressed through the scaling up of the programme into CTEs pre-service training as all new teachers will receive the training on AfL prior to starting their teaching career. Continuing to offer AfL modules in in-service summer training and mandatory CPD will also help to narrow the gap between existing and new teachers. For class sizes, potential methods to minimise the effect on being able to implement AfL in class were discussed, such as dividing the classroom into different groups and focussing on one group at a time. However, the effectiveness of this, given the very high numbers of students, may be limited and may only compound the time burden of teachers of implementing AfL.

EQ2. WHAT CAN THE PROGRAMME DO AT BOTH POLICY LEVEL AND DECENTRALISED STRUCTURE LEVELS TO IMPROVE PROGRAMME INTERVENTIONS AND IMPACT AND PROMOTE SUSTAINABILITY AND THE SCALE-UP OF PROMISING PRACTICES?

There were various several key areas of promise identified to improve the programme interventions. They offer potential to overcome the challenges identified which can be broadly be categorised into issues specific to the implementation of the AfL programme and contextual issues in education in Ethiopia. To address the issues relating to the effectiveness of the cascading knowledge model and its challenges for any scale up, several key areas of promise were identified particularly with regards to pre-service training. As already implemented by Oromia REB, and soon to be followed by three other regions, the integration of AfL modules into CTE programmes represents a cost-effective approach when compared to the cascading knowledge model. In addition, with the Oromia REB taking a lead on successfully integrating the AfL modules into the CTE pre-service training, there is opportunity for knowledge sharing with other REBs that are interested in following the same approach. Other potential avenues of promise can be seen as more interim measures for scale up. These include the inclusion of AfL concepts within short pre-deployment training for teachers, the incorporation of continuous assessment modules in the GEQIP-E programme, extending the programme within schools to further grades in primary outside of first cycle and the use of an online database to hold AfL materials and tools for wide access throughout Ethiopia.

EQS 3 AND 4. TO WHAT EXTENT HAVE PROGRAMME INPUTS MADE A DIFFERENCE TO TEACHERS' ABILITY TO USE CONTINUOUS ASSESSMENT TECHNIQUES? TO WHAT EXTENT ARE CHANGES IN TEACHER PRACTICE ATTRIBUTABLE TO THE AFL PROJECT ACTIVITIES?

In terms of AfL improving teaching practice, there was some promising evidence that the programme has been effective in helping teachers to implement high quality continuous assessment within their classrooms. AfL training appeared to improve the level of knowledge of teachers on the key components of continuous assessments, particularly the ability to develop their own assessment tools and provide effective feedback to their students. In addition, AfL was observed to improve the rate of active assessment and teacher-student interaction in class. Teachers were also observed to structure their lessons to continuous assessment more effectively through the use of clear and comprehensive lesson plans and in-class introductions to learning objectives.

There was also promising evidence that the AfL programme not only changed teacher behaviour, but also that through those changes, students were able to benefit through improved learning

outcomes. A positive treatment effect was found for the AfL programme in Mathematics and Afan Oromo, though no effect was found for the other two core subjects of English and Environmental Science. Due to the limitations of the quantitative research design and the nature of large potential confounding effects on student educational outcomes, any direct attribution to the AfL programme for the estimated effects should be taken with caution. It is potentially the case that the treatment effect on student outcome is biased due to unobserved differences between the treatment groups or other confounding factors such as their home lives. However, this positive finding is also supported by the qualitative findings and anticipated via the underlying ToC logic and the identification of improvements in teaching practice by the AfL programme. As student improvement happens through the pathway of teachers changing their teaching practices, if there was no observed improvement in teaching practice, then any finding relating to student outcomes would likely be spurious.

Apart from academic outcomes for students, the AfL was reported to have beneficial effects on the students, particularly in terms of their school engagement and self-confidence. These changes are attributed to a move to a more student-focussed teaching style that creates a classroom environment where students are better able to engage with their learning and participate in their lessons. In addition, there was a positive treatment effect found in increased parental engagement in their education with teachers and schools communicating with all parents more often and providing a wider range of feedback to them on their child's performance.

EQ5. WHAT WERE THE MOST EFFICIENT AND EFFECTIVE APPROACHES USED BY REGIONS, WOREDAS, SCHOOLS OR TEACHERS TO BRING ABOUT CHANGE? WHAT WORKED, WHAT DID NOT WORK, AND WHY?

There were some identified challenges in the implementation of the AfL programme. In terms of the ToT training, whilst it was generally positively received by participants. It was also observed that in many cases, teachers sent by schools to the ToT training rarely attended the full three sessions, therefore missing the full benefit of attending the ToT training.

There also appeared to have been serious challenges in the implementation of the cascading knowledge model. Focusing firstly on the school training, that is, ToT participants from cluster schools providing AfL training to colleagues in their school, it seems to have regularly taken place and participants reported high levels of satisfaction. It should be noted, however, that the levels of satisfaction were lower than those of ToT participants. The school trainings were also reported to be much shorter and less structured than the ToT training, suggesting a substantial gap in quality between the school training and the ToT training.

At cluster training level, there appeared to be significant bottlenecks in the cascading knowledge model. For example, almost all satellite schools sampled for the quantitative survey reported that they had not received training on AfL. This suggests that the benefit of the programme to schools that are in treatment clusters but not invited to the ToT is limited, and also backed up by the findings of the qualitative analysis. The main causes of this bottleneck, in providing the cluster training, were identified as the lack of budget to cover all the necessary costs, potentially partly because teachers requested per diems to attend the training.

In relation to the delivery of programme materials such as manuals and workbooks, the respondents of both the quantitative and qualitative research indicated that teachers regularly did not receive the full AfL package and believed that this limited the effectiveness with which

they could implement what they had learned. There were also clear bottlenecks in the supervision and monitoring of the AfL programme, particularly at the cluster and school level with school directors reporting a lack of follow-ups after the training. This meant that schools did not have regular supervision on their progress in implementing the AfL programme into their schools nor feedback on how they could improve.

The programme implementers, ELIXIR, took a very flexible approach when rolling out the programme, and adapted the AfL packages and training to local contexts as required. Therefore, while there was at least a loose common structure of the programme to all regions, the detail of the content of the training and materials differed depending on the area. Generally, this approach proved to be very successful, as it enabled the implementors to try out different approaches in the rollout and identify successful aspects of the local implementation to share with the other regions. This is especially true for the integration of the AfL programme into pre-service training at CTEs. This approach was pioneered in Oromia and opens the possibility of sharing knowledge and lessons learned with other regions wishing to take a similar route. While experience sharing is common practice at the higher level, it appears to be less utilised at more decentralised levels. This is one potential avenue for improvement, to incentive local practitioners to meet and share their experiences of implementing AfL on the ground amongst each other.

In our qualitative analysis we observed considerable differences relating to the follow-up and supervision by WEO, with their involvement varying greatly across woredas. As a result, the implementation success of the AfL programme changed depending on the area, with those that had lower levels of WEO involvement and follow-up struggling more to implement the programme. This becomes particularly evident when looking at the provision of school and cluster training. The qualitative data suggests that the extent to which the WEO supports schools greatly influences whether cluster schools cascade the training or not. Moreover, anecdotal evidence from teachers suggested that the follow-up and supervision of teachers' day-to-day use of AfL methods was an effective way of stimulating their use. Furthermore, strong involvement at all levels with all relevant stakeholders is key to the implementation of AfL. If WEOs and cluster supervisors are expected to supervise the implementation of AfL in schools, they need appropriate training. As experience from Oromia demonstrated, providing appropriate training in AfL to local administrators enables them to institutionalise the supervision of AfL via integrating the AfL supervision into their common supervision.

At the school level, some schools have introduced AfL action learning teams, embedding the supervision of AfL at the school level. This enables teachers at schools to exchange ideas with colleagues and, further, provides a first point of contact if teachers face issues in the in-class use of AfL materials and tools. Teachers reported mostly positively about AfL action learning teams when they were in place at their school.

EQ6. WHAT OVERALL LESSONS CAN BE LEARNED FROM THE DELIVERY OF THE AFL?

The key lessons to take away from the delivery of the AfL programme can be summarised as:

- There is some promising evidence that teachers in the Ethiopian education system are capable of applying concepts of continuous assessment within their classroom, both in their teaching practices but also within their structuring of lessons, after undergoing a training programme.

- There are, however, also substantial challenges to schools and teachers in implementing continuous assessment within their classrooms. These particularly relate to how continuous assessments interacts with longstanding issues in education in the Ethiopian context such as high teacher turnover, very large class sizes and poorly resourced schools. The issue of high teacher turnover at schools can be largely addressed through the scaling up of training programmes using pre-service training avenues and use of interim measures such as in-service summer training and CPD.
- There is also some evidence that sustained changes to teaching practices through continuous assessment can then have follow-on improvements in terms of student educational outcomes. Student's may benefit in non-academic areas such as such as increased self-confidence and school engagement.
- There is evidence that implementing continuous assessment and pursuing greater parental engagement can be effective in bringing parents' involvement in their child's education to a higher level. It is however, dependent on the motivation of parents to do so. There is also a need to further increase parents' awareness of the importance and benefits of their involvement in their child's education.
- The flexible nature of the implementation of the AfL project across the rollout led to various approaches being taken across the regions. There have been benefits to this strategy as it gave REBs and local education authorities to adapt materials to their context and needs. It also provided freedom to innovate and strengthen their practices based on the challenges they faced. The sharing of best practices and lessons learned should be a priority to help any sustainable scale up.
- In its current form, the cascading knowledge model may be not an effective way of attempting to scale up training programmes to a wider number of education practitioners such as school directors, teachers and local education officials. This is due to issues relating to financial constraints and issues with motivation for schools and teachers conducting and attending further trainings.
- There is, however, some emerging evidence that the extent to which the WEO and higher-level education officials follow up on the implementation can positively influence the cascading of knowledge for teacher trainings. There is also anecdotal evidence that more intensive follow-up from the WEO on the day-to-day teaching practices can positively influence their use.

EQ7. HAVE ANY CHANGES BEEN ACHIEVED IN RELATION TO POLICY, PRACTICE, ATTITUDES OF PRACTITIONERS AND POLICY MAKERS?

There has been clear progress in obtaining buy-in for AfL as a programme at a regional level, with substantial influence seen on the Oromia education system. During the implementation of AfL, the Oromia REB decided to extend the programme beyond the UNICEF-supported areas with their own funding. Looking towards the future, the Oromia REB has integrated and begun implementing the pre-service training in CTEs – which will be more sustainable than the cascading knowledge model for other REBs as well. As of 2020, three further REBs in Tigray, Afar and Amhara are in the process of integrating AfL into the curriculum at CTEs. Furthermore, the Benishangul-Gumuz REB also expressed interest and desire to integrate the programme into pre-service training. Additionally, the Oromia REB has integrated AfL into pre-deployment training and summer in-service training at CTEs.

Moreover, there have been considerable achievements in securing the sustainability of the programme beyond the funding from UNICEF. Firstly, the AfL programme strongly influenced the Continuous Classroom Assessment module under the World Bank-funded GEQIP-E programme. Covering 50% of schools in Ethiopia, it presents an important avenue for the scale-up of AfL. Secondly, the MoE has integrated AfL into the CPD programme for all teachers, giving teachers the opportunity to upskill themselves in AfL as part of their 60 hours of obligatory CPD per year.

The perception on AfL among high-level stakeholders such as officials at REBs was either already positive at the start of the implementation or improved after attending ToT training and beginning to see the initial outcomes of the implementation. Further down the education chain, teachers, school directors, cluster supervisors and WEO officials mostly shared this positive view of AfL. However, while teachers shared highly positive perceptions on the AfL and the concept of continuous assessment, they were vocal about the practical challenges they faced in implementing AfL techniques into their schools and classrooms. Teachers face numerous contextual challenges when trying to implement AfL, though on the whole, felt that the benefits of AfL outweighed the associated challenges.

UNEXPECTED EFFECTS

The main unexpected effects of the AfL programme relate to the challenges faced with regards to the cascading knowledge model and sharing knowledge with other schools in the cluster. Whilst the evaluation design, and the inclusion of the satellite schools, was developed to understand the wider benefits of the programme to schools that received cluster training. In the analysis, we found that there were no programme effects found on satellite schools. This can be understood as the quantitative and qualitative analyses both identified that the cluster trainings were not widely conducted.

IMPLICATIONS FROM THE STUDY

The implications from the study have been informed by the field evidences, stakeholders' suggestions and evaluators' own experiences. Each has been broken down into suggested actions to ease their implementation. Table 35 lists the relevant stakeholder to implement/support the suggested actions. To enable the implementers, these actions have been classified in terms of order of priority as immediate, short, medium and long-term:

Table 35. Implications from the study

Implications	Target	Priority
While the AfL programme in its current form has been shown to increase teacher knowledge in continuous assessment, the length of the training has been criticised as insufficient by many ToT participants. Hence, to ensure that benefits of the programme are maximised, implementers of the program should attempt to address the causes through potential incentives teachers or better monitoring of attendees. This finding also extends to potential scale-ups and integration into CTE courses to ensure that sufficient time is provided to fully learn the techniques and the components of continuous assessment included in AfL.	UNICEF, MoE, ELIXIR	Medium-term
If the cascading knowledge model is to be continued in certain regions, substantial improvements need to be made in the organisation and monitoring of the cluster and school training. Further budget to cover basic costs and incentives should be considered to ensure that the training is implemented as intended (especially with regards to training length and provision of materials) and that teachers are motivated to attend. Experience sharing workshops at regional or woreda level may be a cost-effective way of improving local implementation and maintaining or increasing the engagement of school staff. Other alternative avenues are, for example, funding small workshops in which at least one ToT trainer is present to ensure that more teachers from satellite schools are able to benefit from the cluster school's knowledge.	UNICEF, MoE, ELIXIR	Medium-term
Oromia has included the AfL programme in its CTE pre-service training. It is recommended that other regions consider doing the same. The pre-service training pathway to scale-up appears to be a cost-effective way of extending the reach of the programme to all newly trained teachers.	REBs	Medium-term
While the CTE pre-service scale-up option has clear benefits; it should be noted pre-service trainees are unlikely to have classroom experience. This is salient as training on the practical applications of continuous assessments seems to have been a key part of the AfL programme. This should not be lost in any integration into CTE modules in favour of theoretical pedagogical material.	REBs, CTEs	Short-term
To ensure long-term success and improvement AfL programme outcomes it is essential that the REBs take full ownership of the programme in their regions. Lessons learnt from Oromia and other regions can be provided to other REBs via experience sharing sessions.	UNICEF, MoE, REBs	Long-term
Efforts should be dedicated to strengthening the supervision and monitoring of the AfL programme at the cluster level. This includes	REBs	Short-term

conducting frequent supervisory support, and capacity strengthening for cluster supervisors and woreda education officers.		
With the outbreak of Covid-19 in Ethiopia and globally, the promotion and use of continuous assessment to overcome the challenges and uncertainty involved with high-stakes testing may become more salient. Specific consideration should be attached to strengthening the practice of recording of student progress for effective collection and use of data. With parents also potentially taking a greater role in their child's education due to school closures, increasing parental engagement and awareness may become even more important in the future.	UNICEF, MoE	Short-term

APPENDICES

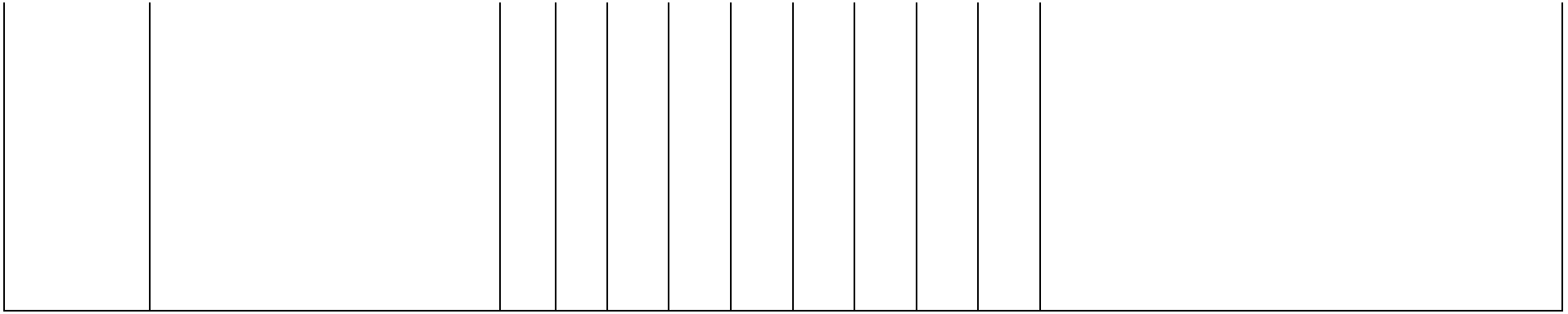
APPENDIX A. EVALUATION MATRIX

	Main Evaluation Questions and Sub-Questions	Data Collection Instrument									Indicators
		Quantitative				Qualitative					
		Student Learning Assessment	Teacher Survey	School Director Survey	Classroom Observation	Teacher SSI	School Director SSI	Trainer SSI	KIIs	Parent FGDs	
Relevance and Equity	**Main Evaluation Question 1: What is the quality and relevance (including gender equity) of programme interventions (materials, modules, tools and training)? How can they be improved for future use?**										
	SEQ 1.1: Was the design of the AfL training programme appropriate to providing knowledge on implementing continuous assessment within the classroom?		X	X		X	X	X			- Participant Satisfaction rate on Modules on key components of continuous assessment (ToT / Cluster) - Difference in satisfaction between ToT and Cluster trainings - Trainer understanding of components
	SEQ 1.2: Was the design of the AfL training programme appropriate to the capacity of participants?		X	X		X	X	X			- Participant satisfaction in trainer's ability to convey information - Participant's satisfaction in the accessibility of the content of the training - Participant's satisfaction with the pacing of the training
	SEQ 1.3: Were the AfL package materials suitable and accessible for schools and teachers?		X	X		X	X	X			- Satisfaction with the accessibility of the language in the materials - Satisfaction with the presentation of ideas in the materials - Number of languages materials translated into
	SEQ 1.4: How appropriate is the AfL programme and continuous assessment in the educational context?		X	X	X	X	X	X	X	X	- Regularity of use of concepts and techniques within classroom - Reported contextual challenges in the classroom to using continuous assessment

											- Interaction between AfL programme and contextual factors
	SEQ 1.5: How can the AfL package be improved for future use?		X	X	X	X	X	X	X	X	- Analysis of above
	SEQ 1.6: Was the AfL package implemented fairly that did not exclude marginalised groups		X	X		X	X	X			- Barriers to entry for participants and interaction with females and marginalised groups.
Sustainability	**Main Evaluation Question 2: What can the programme do at both policy level and decentralized structure levels to improve programme interventions and impact and promote sustainability and the scale-up of promising practices?**										
	SEQ 2.1: What are potential avenues for promoting sustainability of the AfL programme?					X	X		X		- Reported challenges of scale up: - Cost - Contextual factors - Buy-in - Dilution of programme - Infrastructure
	SEQ 2.2: How appropriate is the integration of AfL into pre-service training in CTEs as a potential scale-up option?					X			X		- Quality of integration of modules - Cost of integration - Buy-in of integration from stakeholders
Effectiveness	**Main Evaluation Question 3: To what extent have programme inputs made a difference to teachers' ability to use continuous assessment techniques?**										
	SEQ 3.1: Do teachers have the knowledge to implement continuous assessment in the classroom?		X	X		X	X				- Teacher testing scores on key components of continuous assessment: - Question Development - Feedback Provision - Knowledge of components
	SEQ 3.2: How do teachers use continuous assessment techniques in the classroom?		X	X	X	X	X				- Teacher reported use of range of assessment methods - Teacher reported regularity of assessments - Teacher reported development of their own questions - Teacher reported use of student progress records - Teacher reported use of structured lesson plan - Teacher reported rate of update of student progress records - Teacher reported rate of individual feedback for students - Teacher reported content of individual feedback for students - Observed use of structured lesson plans - Observed use of active assessment activities

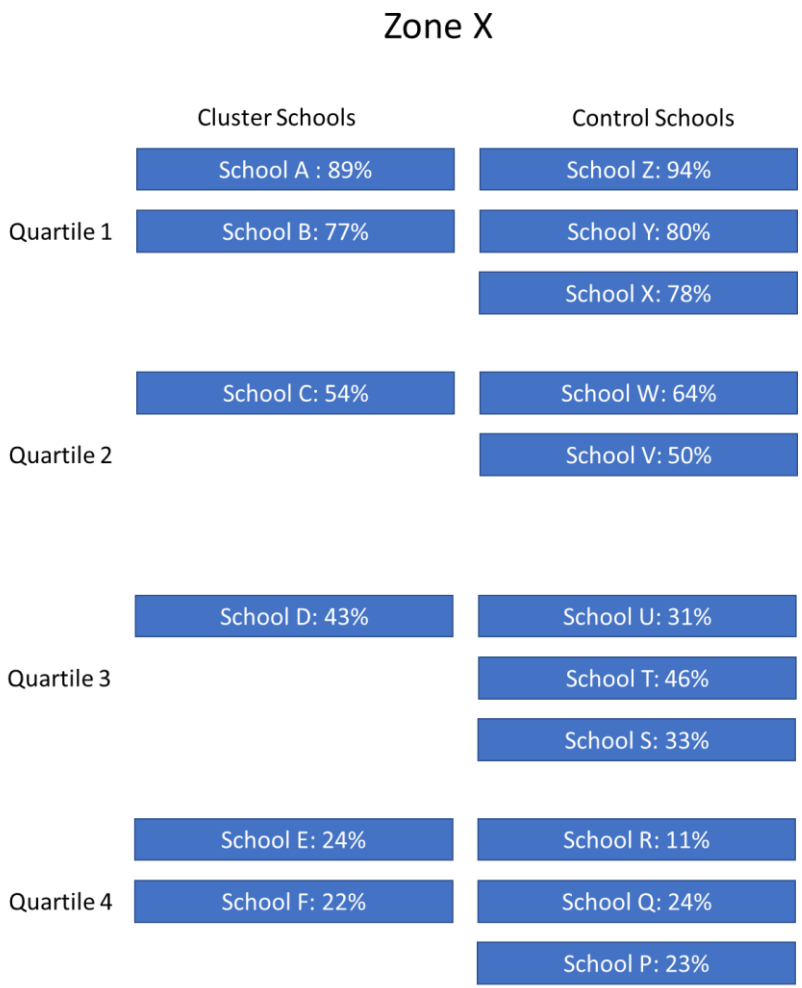
											- Observed teacher engagement with students
	SEQ 3.3: What is the quality of the continuous assessment techniques used in the classroom?		X	X	X	X	X				- Observed quality of structured lesson plans - Observed quality of learning introductions and links to MLC - Observed student engagement in classrooms
	SEQ 3.4: What is the level of parent-school involvement with their child's education?		X	X		X	X			X	- Reported frequency of school and parent meetings - Reported structure of school and parent meetings - Reported content of student feedback to parents - Use of student progress records in parent-school meetings - Differences in parental engagement depending on student attainment level - Parents views on their role in their child's education
Impact and Gender Equity	**Main Evaluation Question 4: To what extent are changes in teacher practice attributable to the AfL project activities?**										
	SEQ 4.1: To what extent are changes in teacher practice attributable to the AfL project activities?		X	X	X	X	X				See SEQs 3.1, 3.2 and 3.3
	SEQ 4.2: Have programme inputs had any effect on parent-school involvement with their child's education?		X	X	X	X	X				See SEQ 3.4
	SEQ 4.3: To what extent is the change in students' outcomes attributable to AfL?	X								X	- Student test scores: - Mother Tongue - English - Mathematics - Environmental Science - Dropout Rates - Repetition Rates - Attendance - Student perceptions on school and learning - Gender equity in changes in outcomes
Effectiveness / Efficiency	**Main Evaluation Question 5: What were the most efficient and effective approaches used by regions, woredas, schools or teachers to bring about change? What worked, what did not work, and why?**										
	SEQ 5.1: Was the ToT training implemented efficiently?		X	X		X	X	X			- Participant attendance rates - Perceptions on the suitability of the length of training

											- Amount of school time missed to attend - Graduation rates
	SEQ 5.2: Was the school and/or cluster training implemented efficiently?		X	X		X	X	X			- Rate of cluster trainings taking place - Rate of school trainings taking place - Participant attendance rates - Perceptions on the suitability of the length of training - Amount of school time missed to attend - Graduation rates
	SEQ 5.3: Was the monitoring of the programme done efficiently and effectively?		X	X		X	X	X	X		- Rate of supervisions from various levels - Rate of feedback provision - Satisfaction with feedback provision - Monitoring framework of the programme - Methods of monitoring of performance indicators
	SEQ 5.4: How did approaches to implementing AfL differ throughout areas?								X		- Reported differences in the implementation of AfL
All	**Main Evaluation Question 6: What overall lessons can be learned from the delivery of the AfL?**										
	Lessons Learned and Implications from the Study - Based on Findings:	X	X	X	X	X	X	X	X	X	- Analysis of all above
	- Appropriateness of AfL programme	X	X	X	X	X	X	X	X	X	
	- Acceptability of AfL programme	X	X	X	X	X	X	X	X	X	
	- Feasibility of AfL programme and scale-up	X	X	X	X	X	X	X	X	X	
	- Sustainability of AfL programme	X	X	X	X	X	X	X	X	X	
All	**Main Evaluation Question 7: Have any changes been achieved in relation to policy, practice, attitudes of practitioners and policy makers?**										
	What are the changes to policy, practice, attitudes of practitioners and policy makers with the current implementation of AfL: - MoE - UNICEF - REBs - Cluster Supervisors - ELIXIR - School directors - Teachers - Parents		X	X	X	X	X	X	X	X	- Buy-in of continuous assessment and wider AfL techniques - Level of social mobilisation and public awareness of programme - Influence on existing curriculum practices - Linkage improvement between assessment, examinations and curriculum - Partnership strengthening with other education institutions - Establishment of MoE AfL adhoc Committee - Establishment of REB AfL coordinating team - Establishment of School Level AfL Action Learning Team

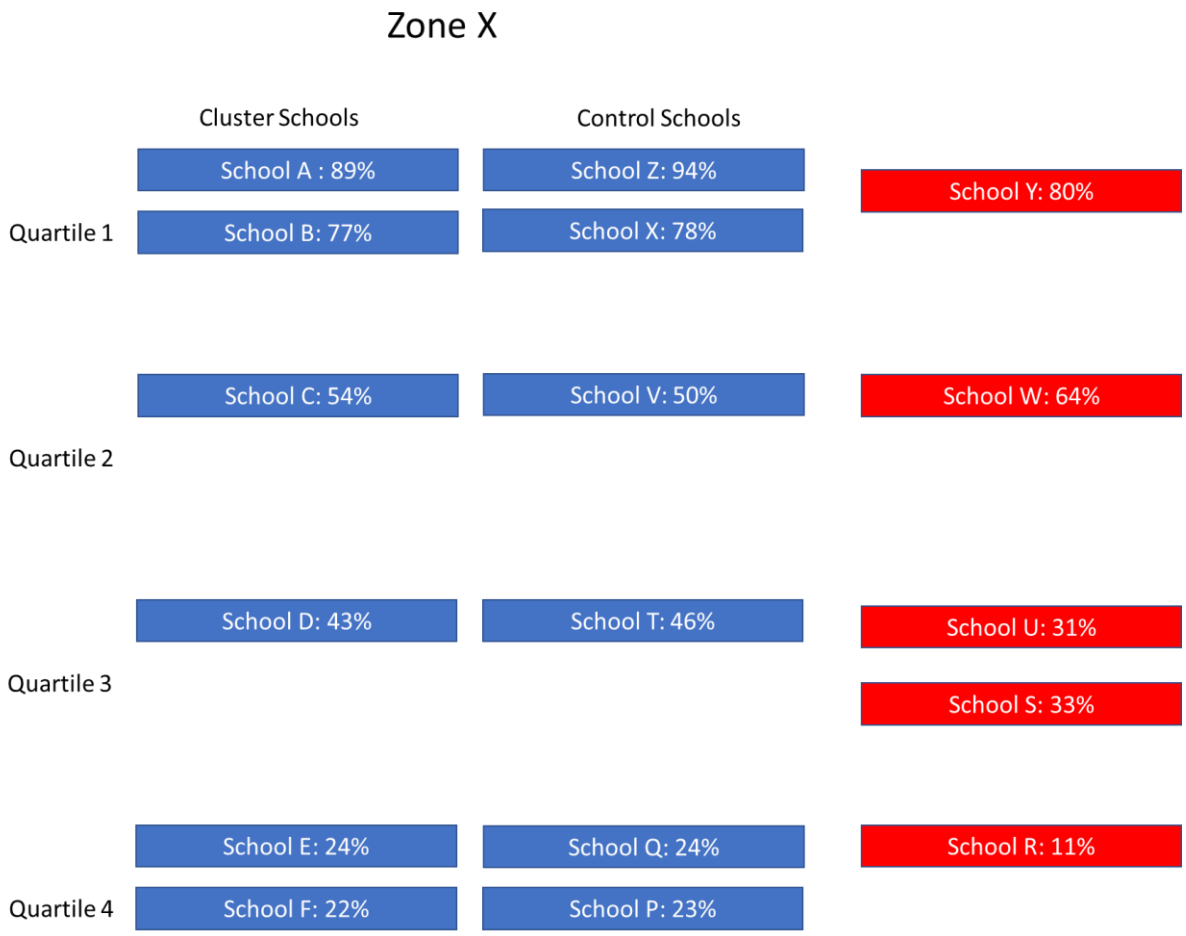


APPENDIX B. MATCHING PROTOCOL

Step 1: Outline all Cluster Schools and Control Schools in a given evaluation zone



Step 2: For each Cluster School, select it's nearest neighbour based on performance score



Step 3: For each performance quartile, select the best match (with the minimum difference between the matched Cluster and Control School)

APPENDIX C. QUANTITATIVE DATA COLLECTION TOOLS

The table below provides an overview of the tools used and the planned sample for quantitative data collection and all tools are provided in Annex.

Instrument	Description	Total per school	Total in sample	Sampling method
Student learning assessment	Per Grade 3 – 4: Afan Oromo, English, Mathematics, Environmental Science	20	1,200	Random selection of 10 Grade 3 and 10 Grade 4 students.
School director survey	School director questionnaire including the following sections: • School director characteristics • School characteristics and facilities • Teacher roster • Implementation of AfL • Continuous assessment practices	1	60	N/A
Teacher survey	Teacher questionnaire including the following sections: • Teacher characteristics • Grades and subjects' roster • Time use • Raven's and digit span test • Continuous assessment practices • Implementation of AfL • Self-efficacy • Motivation	3	180	Stratified Random Sampling of: 1 AfL FTT (Cluster) 1 AfL TTS (Satellite) Subsequent random sample of remaining teachers.
Classroom observations	In-class observation form with the following sections: • Class information • Time on task matrix • Classroom assets • Observations	3	180	Non-random but directed to observe at least three different subjects and teachers.

APPENDIX D. REPLACED SCHOOLS IN SAMPLE

Replaced			Replacement				
Woreda	**School name**	**t***	**Woreda**	**School name**	**t***	**Date**	**Comment**
Burqaa Dhimtuu	Baha Biftuu	1	Adaamii Tulluu	Battallee	1	02-Mar	Teacher Strike – no classes
Jaju	3d Intaye	3	Jaju	Shereyda	3	03-Mar	No school called 3D INTAYE in the woreda
Fantallee	Abadir	1	Siraaroo	Lokkee Shamballii	1	18-Feb	Inter-clan fighting caused school closure
Fantallee	Dire Seden	1	Burqaa Dhimtuu	Baha Biftuu	1	18-Feb	Inter-clan fighting caused school closure
Fantallee	Galcha	1	Tole	Tullu	1	18-Feb	Inter-clan fighting caused school closure
Fantallee	Gara Dima	2	Daro Labuu	Tamsaay Xuphoo	2	18-Feb	Inter-clan fighting caused school closure
Gimbichu	Tesfa Maekal	3	Gimbichu	Araddaa	3	24-Feb	Tesfa Maekal is a private school
Jibat	Gamo Yabalo	3	Jibat	Maru Leki	3	24-Feb	Gamo Yabalo no longer in sample woreda due to change in borders
Jibat	Maru Gombo	3	Jibat	Tutu Jibat	3	24-Feb	Maru Gombo no longer in sample woreda

* 1= Cluster 2=Satellite 3=Control

APPENDIX E. BALANCE OF MATCHING CHARACTERISTICS

	(1)	(2)	(3)	(4)	(5)	(6)
Variable	**Mean cluster**	**Mean satellite**	**Mean control**	**Cluster vs satellite**	**Cluster vs control**	**Satellite vs control**
School Inspection Score (0-1)	0.605	0.593	0.607	0.012	-0.002	-0.014
	(0.096)	(0.106)	(0.077)	(0.031)	(0.028)	(0.034)
Urban School	0.161	0.000	0.267	0.161	-0.105	-0.267**
	(0.374)	(0.000)	(0.458)	(0.097)	(0.127)	(0.118)
Observations	31	15	15	46	46	30

Note: * p<0.1, ** p<0.05, *** p<0.01

APPENDIX F. QUALITATIVE SURVEY TOOLS

The table below gives an overview of all instruments that the research team used for the qualitative data collection and all tools are provided in Annex.

Instrument	Instrument	Description
School based instruments	**Focus group discussion with parents**	The focus group topic guide was used with parents of Grade 1-4 school children whose school is implementing AfL.
	ToT-trained Teachers	This interview topic guide was used with teachers in cluster schools who had received the ToT training.
	School-trained or Cluster-trained Teachers	This interview topic guide was used with teachers in satellite schools who had received the cluster training, or with teachers in cluster schools who had received the school training.
	ToT-trained School directors	This interview topic guide was used with school directors in cluster schools who had received the ToT training.
	Cluster-Trained School directors at satellite Schools	This interview topic guide was used with school directors in satellite schools who had received the cluster training.
Non-school-based instruments	**ToT Trainers**	This interview topic guide was used with ToT trainers who conduct the ToT training.
	Cluster supervisors	This interview topic guide was used with cluster supervisors who had received the ToT training.
	Woreda Education Office Staff	This interview topic guide was used with WEO staff who had received the ToT training.
	CTE instructors	This interview topic guide was used with CTE instructors, ideally those who had been involved with integrating AfL into CTE pre-service teacher training.
	Regional Educational Bureau staff	This interview topic guide was used with Regional Educational Bureau staff who have been involved with AfL.
	UNICEF staff	This interview topic guide was used with key informants from UNICEF, who are in charge of the AfL programme.
	ELIXIR staff	This interview topic guide was used with key informants from the local implementing partner ELIXIR.

APPENDIX G. LIST OF RESPONDENTS IN THE QUALITATIVE ANALYSIS

No.	Name	Region	Woreda	Role	Type of training obtained
1	KII1	-	-	Key informants (ELIXIR)	-
2	KII2	-	-	Key informants (UNICEF)	-
3	BeGu1	Benishangul-Gumuz	Homosha	Cluster Supervisor	ToT training
4	BeGu2	Benishangul-Gumuz	Bambasi	Teacher	ToT training
5	BeGu3	Benishangul-Gumuz	Bambasi	Teacher	ToT training
6	BeGu4	Benishangul-Gumuz	Bambasi	Teacher	ToT training
7	BeGu5	Benishangul-Gumuz	Bambasi	School Director	ToT training
8	BeGu6	Benishangul-Gumuz	Bambasi	Teacher	ToT training
9	BeGu7	Benishangul-Gumuz	Homosha	Teacher	ToT training
10	BeGu8	Benishangul-Gumuz	Homosha	Teacher	ToT training
11	BeGu9	Benishangul-Gumuz	Homosha	School Director	ToT training
12	BeGu10	Benishangul-Gumuz	Homosha	Teacher	ToT training
13	BeGu11	Benishangul-Gumuz	Homosha	Teacher	ToT training
14	BeGu12	Benishangul-Gumuz	Homosha	Teacher	School training
15	BeGu13	Benishangul-Gumuz	Homosha	School Director	ToT training
16	BeGu14	Benishangul-Gumuz	Homosha	Teacher	ToT training
17	BeGu15	Benishangul-Gumuz	Homosha	Teacher	ToT training
18	BeGu16	Benishangul-Gumuz	Homosha	Teacher	ToT training
19	BeGu17	Benishangul-Gumuz	Bambasi	Teacher	ToT training
20	BeGu18	Benishangul-Gumuz	Bambasi	Teacher	ToT training
21	BeGu19	Benishangul-Gumuz	Bambasi	Teacher	ToT training
22	BeGu20	Benishangul-Gumuz	Bambasi	Teacher	ToT training
23	BeGu21	Benishangul-Gumuz	Bambasi	School Director	ToT training
24	BeGu22	Benishangul-Gumuz	-	REB	-
25	BeGu23	Benishangul-Gumuz	-	ToT trainer	ToT training
26	BeGu24	Benishangul-Gumuz	Bambasi	WEO	ToT training
27	BeGu25	Benishangul-Gumuz	Homosha	Parents	-
28	BeGu26	Benishangul-Gumuz	Bambasi	Parents	-
29	Oromia1	Oromia	Tole	Teacher	Cluster training
30	Oromia2	Oromia	Ambo	Parents	-
31	Oromia3	Oromia	Ambo	Teacher	ToT training
32	Oromia4	Oromia	Tole	Teacher	ToT training
33	Oromia5	Oromia	Ambo	Teacher	ToT training
34	Oromia6	Oromia	Ambo	Teacher	ToT training
35	Oromia7	Oromia	Ambo	School Director	ToT training
36	Oromia8	Oromia	Tole	School Director	ToT training

37	Oromia9	Oromia	Tole	Teacher	ToT training
38	Oromia10	Oromia	Ambo	School Director	ToT training
39	Oromia11	Oromia	Ambo	Teacher	ToT training
40	Oromia12	Oromia	Tole	Teacher	ToT training
41	Oromia13	Oromia	Ambo	Teacher	School training
42	Oromia14	Oromia	Tole	Teacher	ToT training
43	Oromia15	Oromia	Ambo	Teacher	School training
44	Oromia16	Oromia	Ambo	Teacher	ToT training
45	Oromia17	Oromia	Ambo	Teacher	School training
46	Oromia18	Oromia	Tole	Teacher	ToT training
47	Oromia19	Oromia	Tole	Teacher	Cluster training
48	Oromia20	Oromia	-	CTE Instructor	-
49	Oromia21	Oromia	Ambo	Cluster Supervisor	Untrained
50	Oromia22	Oromia	Ambo	School Director	ToT training
51	Oromia23	Oromia	-	REB	-
52	Oromia24	Oromia	Tole	School Director	ToT training
53	Oromia25	Oromia	Tole	Teacher	ToT training
54	Oromia26	Oromia	Tole	Parents	-
55	Oromia27	Oromia	Ambo	Teacher	ToT training
56	Oromia28	Oromia	Ambo	Teacher	Cluster training
57	Oromia29	Oromia	Ambo	Teacher	School training
58	Oromia30	Oromia	Ambo	Teacher	ToT training
59	Oromia31	Oromia	Tole	Teacher	Cluster training
60	Oromia32	Oromia	Tole	Teacher	ToT training
61	Oromia33	Oromia	Tole	School Director	ToT training
62	Oromia34	Oromia	-	ToT trainer	-
63	Oromia35	Oromia	Tole	Teacher	Cluster training
64	Oromia36	Oromia	Ambo	WEO	Untrained
65	Oromia37	Oromia	Tole	Teacher	School training
66	Tigray1	Tigray	Gulomekeda	Teacher	Cluster training
67	Tigray2	Tigray	Kilete Awelallo	Teacher	School training
68	Tigray3	Tigray	Kilete Awelallo	Teacher	School training
69	Tigray4	Tigray	Gulomekeda	Parents	-
70	Tigray5	Tigray	Kilete Awelallo	School Director	Untrained
71	Tigray6	Tigray	Kilete Awelallo	Teacher	Untrained
72	Tigray7	Tigray	Kilete Awelallo	Teacher	Untrained
73	Tigray8	Tigray	Gulomekeda	Teacher	School training
74	Tigray9	Tigray	Gulomekeda	Teacher	ToT training
75	Tigray10	Tigray	Gulomekeda	Teacher	ToT training
76	Tigray11	Tigray	Kilete Awelallo	WEO	-
77	Tigray12	Tigray	-	ToT trainer	-

78	Tigray13	Tigray	Gulomekeda	Teacher	ToT training
79	Tigray14	Tigray	-	CTE Instructor	-
80	Tigray15	Tigray	Kilete Awelallo	Teacher	ToT training
81	Tigray16	Tigray	-	REB	-
82	Tigray17	Tigray	Gulomekeda	Teacher	Cluster training
83	Tigray18	Tigray	Gulomekeda	Cluster Supervisor	ToT training
84	Tigray19	Tigray	Gulomekeda	School Director	ToT training
85	Tigray20	Tigray	Kilete Awelallo	Teacher	Untrained
86	Tigray21	Tigray	Kilete Awelallo	School Director	ToT training
87	Tigray22	Tigray	Kilete Awelallo	Teacher	Untrained
88	Tigray23	Tigray	Kilete Awelallo	Teacher	Untrained
89	Tigray24	Tigray	Kilete Awelallo	Parents	-

APPENDIX H. GENDER SPLIT OF TOT PARTICIPANTS

Gender	Attended ToT training (as reported by the school director)		
	Yes	No	Total
Male (row %)	36 (50.7)	35 (49.3)	71
Female	34 (47.8)	37 (52.2)	71
Total	70	72	142

APPENDIX I. PERCEPTIONS OF TRAINING QUALITY ON KEY COMPONENTS (CLUSTER/SCHOOL AND TOT TRAINING)

Perceptions of ToT training quality on key components (Cluster/ School and ToT)

Training area	Strongly agree (5) (%)	Somewhat agree (4) (%)	Neutral (3) (%)	Somewhat disagree (2) (%)	Strongly disagree (1) (%)	Average (1-5)	Average – ToT (1-5)	'Quality gap' (ToT – Cluster/School)
Developing Structured Lesson Plans	14.3	39.3	28.6	17.8	0	**3.5**	**4.1**	**0.6**
Developing high quality questions of my own for students	17.9	35.7	25	14.3	7.1	**3.4**	**4.0**	**0.6**
Using a wide range of assessment methods	10.7	50	21.4	17.9	0	**3.5**	**3.8**	**0.3**
Providing Students with effective feedback	17.9	39.3	25	17.9	0	**3.6**	**3.8**	**0.2**
Using information from assessment to alter my teaching practices	21.4	35.7	28.6	10.7	3.6	**3.6**	**3.8**	**0.2**

Participant perception on training interactivity and relevance (Cluster/School and ToT)

Facilitation criteria	Strongly agree (5) (%)	Somewhat agree (4) (%)	Neutral (3) (%)	Somewhat disagree (2) (%)	Strongly disagree (1) (%)	Average (1-5)	Average – ToT (1-5)	'Quality gap' (ToT – Cluster/School)
Interactive	32.1	39.3	25	3.6	0	**4.0**	**4.3**	**0.3**
Relevant	21.4	60.7	10.7	7.1	0	**4.0**	**4.3**	**0.3**

APPENDIX J. TEACHERS REPORTING ON CHALLENGES IN IMPLEMENTING CONTINUOUS ASSESSMENT

Teachers reporting on class size challenges to implementing continuous assessment

Treatment group	Not a challenge at all	Somewhat of a challenge	A big challenge
Cluster school (row %)	10.6	31.9	57.5
Satellite school	17.8	33.3	48.9
Control school	14.3	31.0	54.8
Pooled	13.3	32.0	54.7

Teachers reporting on time constraint challenges to implementing continuous assessment

Treatment group	Not a challenge at all	Somewhat of a challenge	A big challenge
Cluster school (row %)	13.8	44.7	41.5
Satellite school	11.1	51.1	37.8
Control school	4.8	52.4	42.9
Pooled	11.1	48.1	40.9

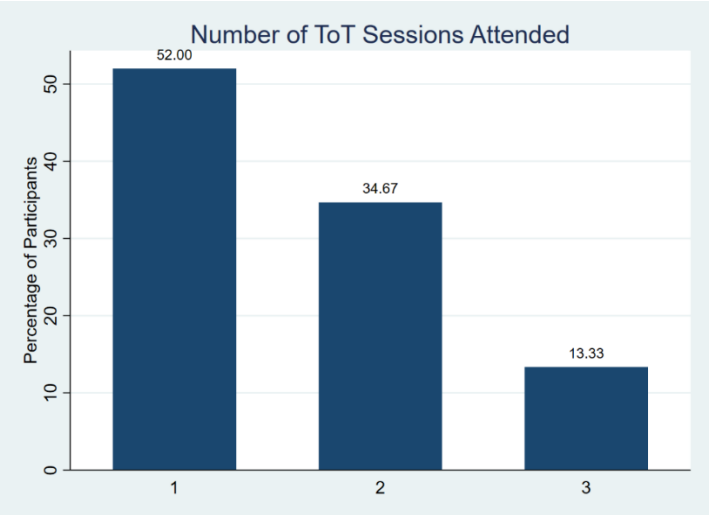
Teachers reporting on resource challenges to implementing continuous assessment

Treatment group	Not a challenge at all	Somewhat of a challenge	A big challenge
Cluster school (row %)	13.8	37.2	48.9
Satellite school	13.3	42.2	44.4
Control school	14.3	42.9	42.9
Pooled	13.8	39.8	46.4

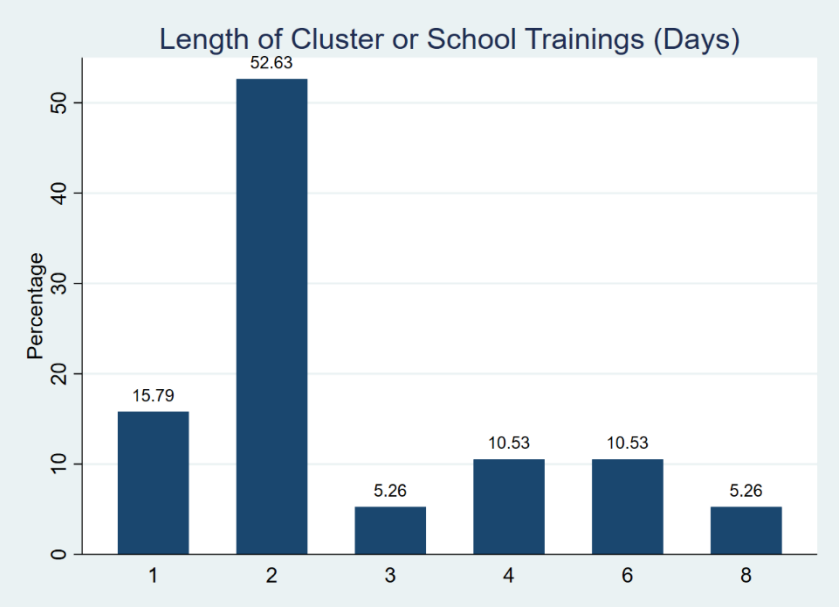
Teachers reporting on motivation to implement continuous assessment

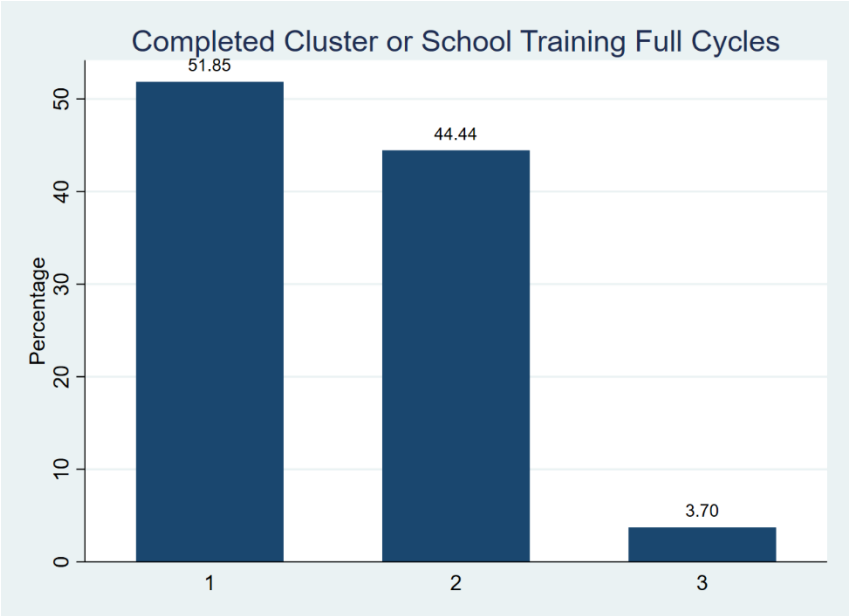
Treatment group	Not a challenge at all	Somewhat of a challenge	A big challenge
Cluster school (row %)	58.5	26.6	14.9
Satellite school	37.8	48.9	13.3
Control school	33.3	42.9	23.8
Pooled	47.5	35.9	16.6

APPENDIX K. NUMBER OF TOT SESSIONS ATTENDED

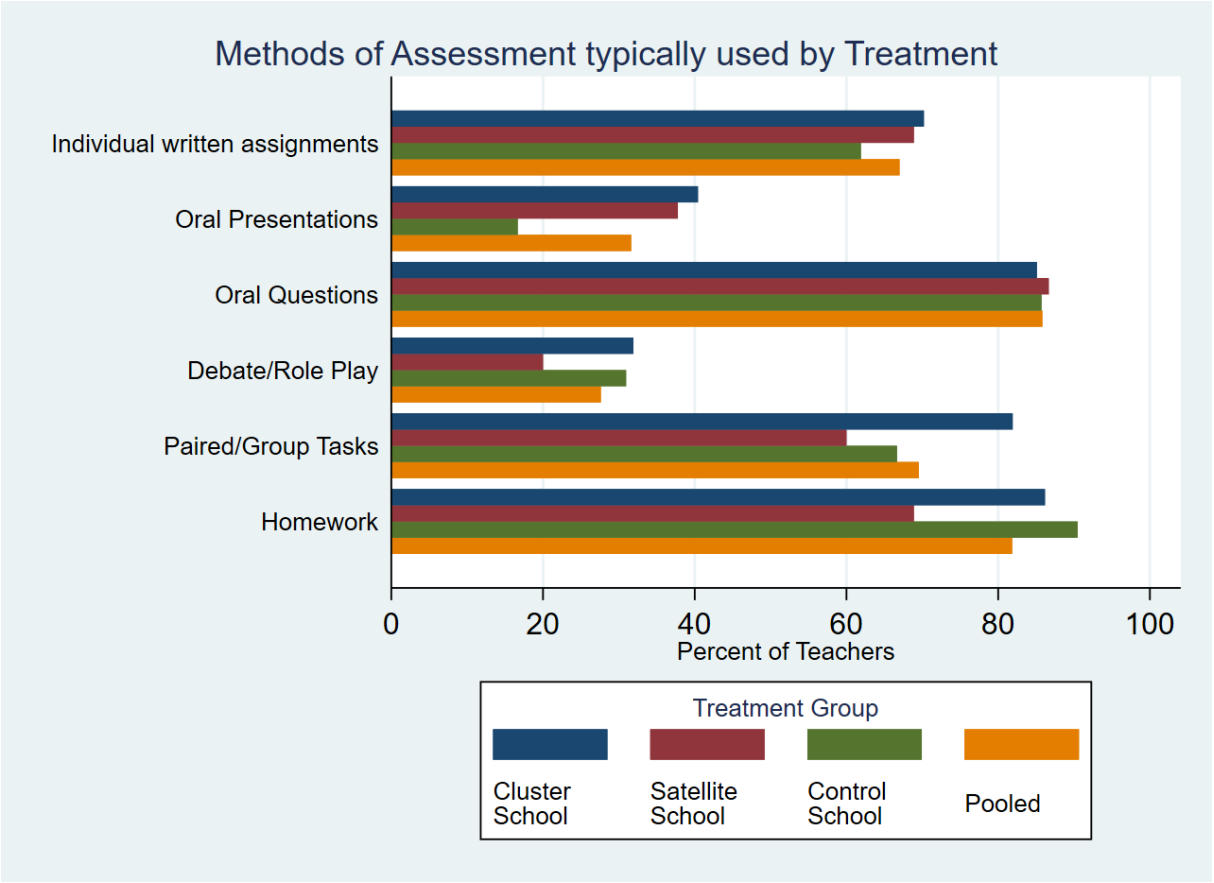


APPENDIX L. LENGTH OF CLUSTER OR SCHOOL TRAINING AND NUMBER OF CYCLES

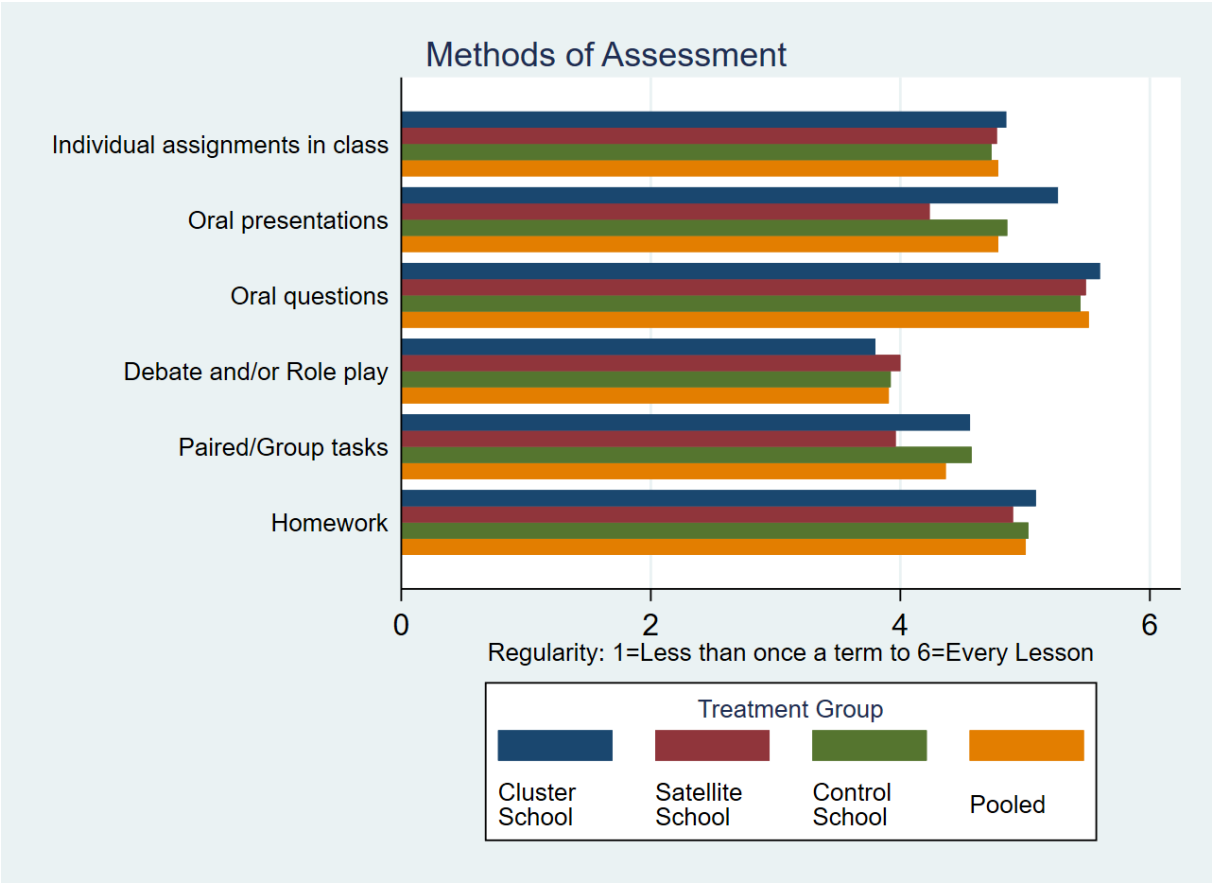




APPENDIX M. METHODS OF ASSESSMENT USED

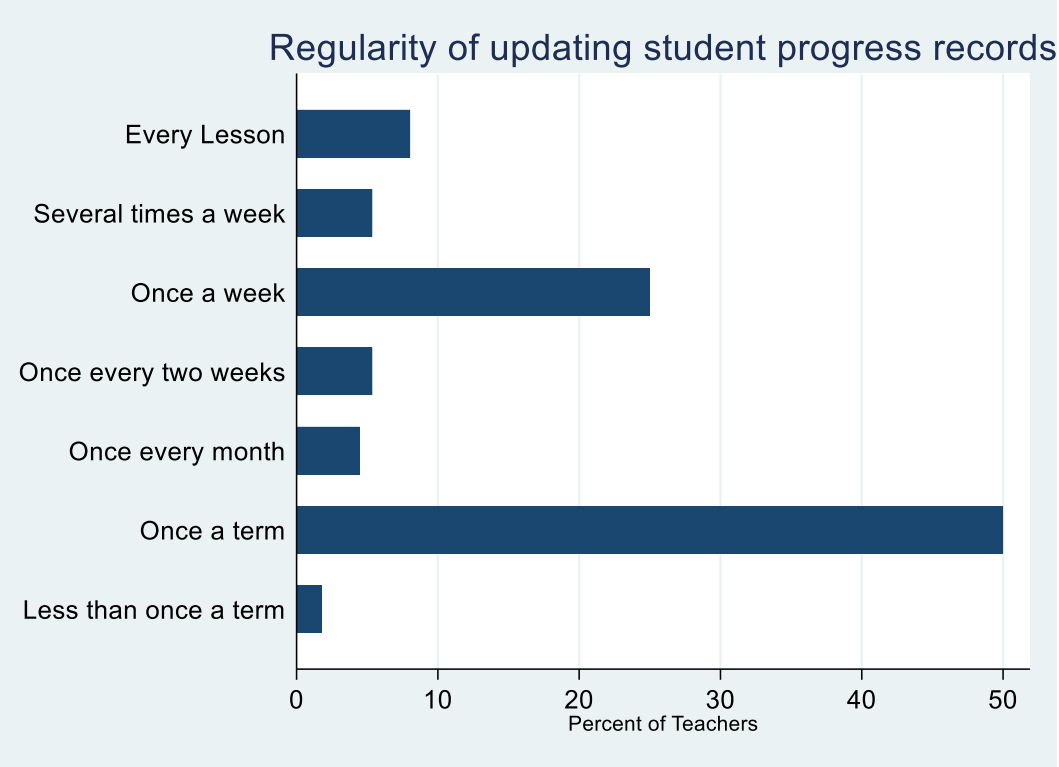


APPENDIX N. RATE OF ASSESSMENTS BY TYPE



APPENDIX O. REGULARITY OF UPDATE OF STUDENT PROGRESS RECORDS

Regularity of updating student progress records



REFERENCES

Armbruster, P., Patel, M., Johnson, E., & Weiss, M. (2009). Active learning and student-centered pedagogy improve student attitudes and performance in introductory biology. CBE Life Sciences Education, 8, 203-213.

Ball, D. L., & Feiman-Nemser, S. (1988). Using Textbooks and Teachers' Guides: A Dilemma for Beginning Teachers and Teacher Educators. Curriculum Inquiry, 18(4), 401

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2016), Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India, Working Paper, MIT

Banerjee, A., Duflo, E. (2011). Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty.

Benavot, A., and E. Tanner. (2007). The Growth of National Learning Assessments in the World, 1995–2006. Background paper prepared for the Education for All Global Monitoring Report, 2008, UNESCO, Paris.

Berry, J., Kannan, H., Mukherji, S., & Shotland, M. (2020). "Failure of frequent assessment: An evaluation of India's continuous and comprehensive evaluation program," Journal of Development Economics, Elsevier, vol. 143(C).

Black, P., & Wiliam, D. (1998a). Inside The Black Box: Raising Standards Through Classroom Assessment, Phi Delta Kappan, 80(2).: 139–144

Black, P., & Wiliam, D. (1998b). Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice, 5(1), 7–74.

Brown, S., Race, P. & Smith, B. (1996). 500 tips on assessment, London: Kogan Page.

Browne, E. (2016). Evidence on Formative Classroom Assessment for Learning. K4D Helpdesk Report. Brighton, UK: Institute of Development Studies

Burdett, N. (2017). "Review of High Stakes Examination Instruments in Primary and Secondary School in Developing Countries." RISE Working Paper 17/018.

Carlson, M.O., Humphrey, G.E., & Reinhardt, K.A. (2003). Weaving Science Inquiry and Continuous Assessment: Using Formative Assessment to Improve Learning.

Central Statistics Agency (CSA). (2016). "Agricultural Sample Surveys 2015/2016." In. Addis Ababa, Ethiopia: Federal Democratic Republic of Ethiopia, Central Statistics Agency (CSA).

Chisholm, L., & Leyendecker, R. (2008). Curriculum reform in post-1990s sub-Saharan Africa. International Journal of Educational Development, 28(2), 195-205.

De Lisle, J. (2016). Unravelling continuous assessment practice: Policy implications for teachers' work and professional learning. Studies in Educational Evaluation, 50, 33–45.

Dowrich, M. (2008). Teacher Perceptions of the Implementation of the National Continuous Assessment Programme in a Primary School in the St. George East Education District in Trinidad and Tobago, Submitted in Partial Fulfillment of the Requirement for the Degree of Master of Education (Concentration in Curriculum) of The University of the West Indies. https://www.researchgate.net/publication/279511071_Teacher_perceptions_on_the_implementation_of_the_national_Continuous_Assessment_Programme_in_a_primary_school_in_the_St_George_East_Education_District_in_Trinidad_and_Tobago (Accessed 7 June 2020.)

Dundar, H., Béteille, T., Riboud, M., & Deolalikar, A. (2014). Student Learning in South Asia: Challenges, Opportunities, and Policy Priorities. Washington, DC: World Bank.

Federal Democratic Republic of Ethiopia Ministry of Education (MoE). (2015). "Education Sector Development Programme V (ESDP V)." In. Addis Ababa, Ethiopia: The Federal Democratic Republic of Ethiopia, Federal Ministry of Education (MoE).

Federal Democratic Republic of Ethiopia Ministry of Education (MoE) (2019). Education Statistics Annual Abstract 2011 E.C. (2018/19).

Guilbert, J-J., & World Health Organization. (1998). Educational handbook for health personnel / J.-J. Guilbert, 6th ed. rev. and updated 1998. World Health Organization.

Harlen W., & Deakin, C. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning (EPPI-Centre Review, version 1.1*). In: Research Evidence in Education Library. Issue 1. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Hayford, S. K. (2007). Continuous assessment and lower attaining pupils in primary and junior secondary schools in Ghana (Doctoral dissertation). School of Education, University of Birmingham, Birmingham, UK.

Heyneman, S. (1987). Uses of Examinations in Developing Countries: Selection, Research, and Education Section Management. International Journal of Educational Development 7: 251–63

Hill, P 2013. Examination Systems. Paris: United Nations Educational, Scientific and Cultural Organization.

Hounsell, D. (2003). “Student feedback, learning and development”. In Higher education and the lifecourse, Edited by: Slowey, M. and Watson, D. 67–78. Buckingham: Society for Research into Higher Education & Open University Press.

Iipinge, S. M., & Kasanda, C. D. (2013). Challenges associated with curriculum alignment, change and assessment reforms in Namibia. Assessment in Education: Principles, Policy & Practice, 20(4), 424-441.

Kamangira, Y. T. (2003). Feasibility of a large-scale implementation of continuous assessment as a stimulus for teacher development in Malawi. American Institutes for Research: Improving Educational Quality (IEQ) Project.

Kapambwe, W. M. (2010). The implementation of school based continuous assessment (CA) in Zambia. Educational Research and Reviews, Vol. 5, No.3, pp. 99-107.

Kellaghan, T., & Greaney, V. (1992). Using Examinations to Improve Education: A Study in Fourteen African Countries. Technical Paper 165, Africa Technical Department Series, World Bank, Washington, DC

Kellaghan, T., and Greaney, V. (2003). Monitoring Performance: Assessment and Examinations in Africa. Background paper commissioned by the Association for the Development of Education in Africa (ADEA) in the framework of The Challenge of Learning Study, ADEA, Paris

Kellaghan, T., & Greaney, V. (2019). Public Examinations Examined (English). Washington, D.C. World Bank Group.

Kennedy, A. (2005). Models of Continuing Professional Development: A Framework for Analysis. Journal of In-Service Education 31 (2): 235–250.

Kingston, N., & Nash, B. (2011). Formative assessment: a meta-analysis and a call for research. Educ. Meas. 30(4), 28-37.

Kinzie, M.B. (1990). Requirements and benefits of effective interactive instruction: Learner control, self-regulation, and continuing motivation. Educational Technology Research and Development. 38(1).

Le Grange, L., & Reddy, C. (1998). Continuous Assessment: an introduction and guidelines to implementation. Cape Town. South Africa: Junta.

Levine, L. E., Fallahi, C. R., Nicoll-Senft, J. M., Tessier, J. T., Watson, C. L., & Wood, R. M. (2008). Creating significant learning experiences across disciplines. College Teaching, 56, 247-254

London, M. (1995). Giving feedback: Source-centered antecedents and consequences of constructive and destructive feedback. Human Resource Management Review, 5(3), 159–188.

Modupe, A. V., & Sunday, O. M. (2015). Teachers' perception and implementation of continuous assessment practices in secondary schools in Ekiti-State, Nigeria. Journal of Education and Practice, 6, 17-20.

Moller, Lars Christian. (2015). " Ethiopia's great run : the growth acceleration and how to pace it (English)." In. Washington, D.C., USA: World Bank Group.

Muskin, J. (2017), "Continuous Assessment for Improved Teaching and Learning: A Critical Review to Inform Policy and Practice", Current and Critical Issues in Curriculum, Learning and Assessment No. 13

N'Namdi, K. A. (2005). Guide to teaching reading at the primary school level, Paris: UNESCO.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1978). The Belmont report: Ethical principles and guidelines for the protection of human subjects of research. Bethesda, MD.

National General Education Inspection Framework. MoE. Available at http://www.moe.gov.et/documents/20182/49719/School+Inspection+Framework+2013.pdf/c63f7af5-d2ec-4f80-9733-e49f1b61c698?version=1.0

Nitko, A. J. (1995). Curriculum-based Continuous Assessment: a framework for concepts, procedures and policy. Assessment in Education: Principles, Policy & Practice, 2(3), 321–337.

Orkin, Kate. (2013). "The Effect of Lengthening the School Day on Children's Achievement in Ethiopia." In. Oxford, Great Britain: Young Lives.

Ovando, M.N. (1994), "Constructive Feedback: A Key to Successful Teaching and Learning", International Journal of Educational Management, Vol. 8 No. 6, pp. 19-22.

Overseas Development Institute (ODI). (2011). "Ethiopia's progress on education: A rapid and equitable expansion of access." In. London, Great Britain: Overseas Development Institute (ODI).

Powell, Mary Ann; Taylor, Nicola; Fitzgerald, Robyn; Graham, Ann; Anderson, Donnah (2013). Ethical Research Involving Children, Innocenti Publications. UNICEF Office of Research - Innocenti, Florence.

Pritchett, L., & Beatty, A. (2012). "The Negative Consequences of Overambitious Curricula in Developing Countries." CGD Working Paper 293. Washington, D.C.: Center for Global Development

Quansah, K. B. 2005. Continuous Assessment Handbook. INEE.

Reed, D. (2012). Clearly communicating the learning objective matters! Middle School Journal, 43, 16-24.

Rossiter, J., Azubuike, O. B., & Rolleston, C. (2016) Young Lives School Survey, 2016-17: Evidence from Ethiopia. Oxford: Young Lives.

Solomon, J., & S. Tresman. (1999). A Model for Continued Professional Development: Knowledge, Belief and Action. Journal of In-Service Education 25 (2): 307–319.

Stallings, J. (1977). Learning to Look: A Handbook on Classroom Observation and Teaching Models. Belmont, CA: Wadsworth Publishing.

Uiseb, I. (2009). The Role of Teachers in Continuous Assessment: A Model for Primary Schools in Windhoek. University of South Africa [Master's Dissertation]

UNESCO. (2008). EFA Global Monitoring Report. United Nations Educational, Scientific and Cultural Organisation, Paris.

UNESCO. (2015). Incheon Declaration – Education 2030: Towards Inclusive and Equitable Quality Education and Lifelong Learning for All. World Education Forum 2015.

Unesco Institute for Statistics (UIS). (2019). "UIS Database." In.: Unesco Institute for Statistics.

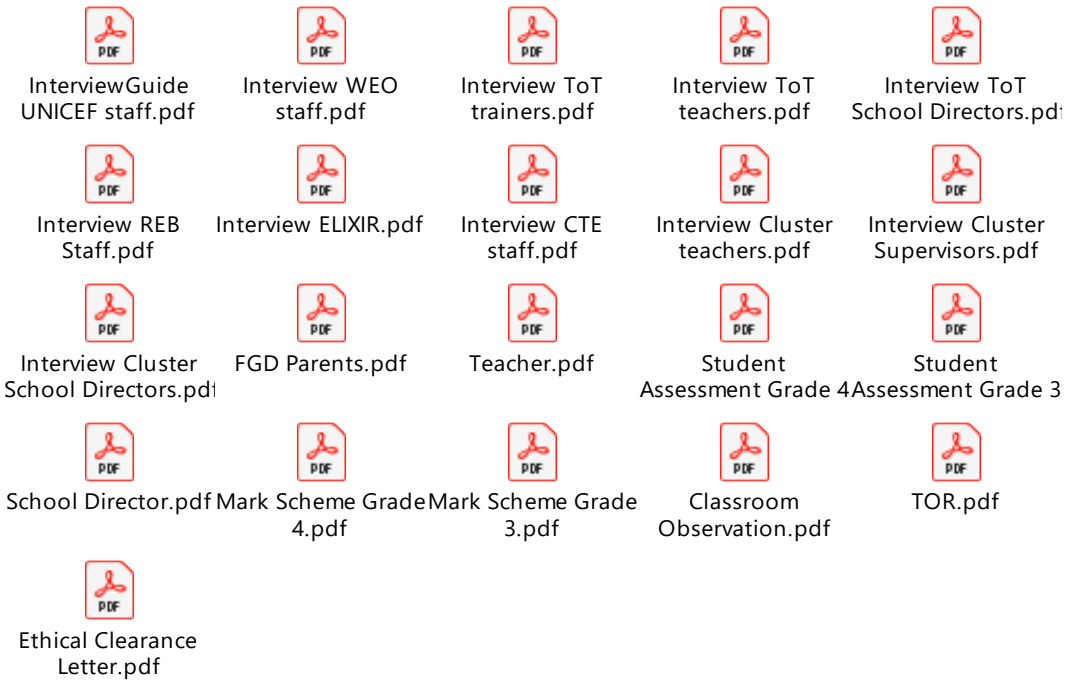
UNICEF. (2019). 'Country profiles: Ethiopia: UNICEF Data: monitoring the situation of children and women', Accessed 18/11/19. https://data.unicef.org/country/eth/.

USAID. (2011). "Ethiopia Early Grade Reading Assessment." In. Addis Ababa, Ethiopia: United States Agency for International Development (USAID).

World Bank. 2019. Ethiopia Economic Update 7: Special Topic - Poverty and Household Welfare in Ethiopia, 2011-16. In.: World Bank Group.

World Bank. 2020. Ethiopia General Education Quality Improvement Program For Equity. [online] Available at: <https://projects.worldbank.org/en/projects-operations/project-detail/P163050> [Accessed 24 June 2020].

ANNEX


InterviewGuide UNICEF staff.pdf
Interview WEO staff.pdf
Interview ToT trainers.pdf
Interview ToT teachers.pdf
Interview ToT School Directors.pdf
Interview REB Staff.pdf
Interview ELIXIR.pdf
Interview CTE staff.pdf
Interview Cluster teachers.pdf
Interview Cluster Supervisors.pdf
Interview Cluster School Directors.pdf
FGD Parents.pdf
Teacher.pdf
Student Assessment Grade 4
Student Assessment Grade 3
School Director.pdf
Mark Scheme Grade 4.pdf
Mark Scheme Grade 3.pdf
Classroom Observation.pdf
TOR.pdf
Ethical Clearance Letter.pdf

Center for Evaluation and Development

C4ED

O7, 3
68161 Mannheim, Germany

Phone: +49 621 9504070

Fax: +49 621 95040710

Email: info@c4ed.org

www.c4ed.org