THE REPUBLIC OF UGANDA

Ministry of Education and Sports

Final Core Report

Measuring Learning Achievement 2010:

Language and Maths in P4

UNITY Project, Uganda

With support from USAID

For Creative Associates International, Inc.

By School-to-School International

March 2011

Contents

Acknowledgments 3
Acronyms 4
Executive Summary 5
Introduction 9
Methodology 9
Findings 11
Recommendations 31
Annex 1: Sampling and Testing 34
Annex 2: Panel Analysis 44
References 49

Acknowledgments

The MLA (Measuring Learning Achievement) exercise described in this report was conducted by a number of people without whom the quality of the assessments, as well as the implementation of the MLA exercise in general, would not have been possible. In particular, the authors would like to extend their sincerest appreciation to Renuka Pillay, Chief-of-Party for the UNITY project, for providing overall guidance and in-country support; to Sandhya Bandrinath and Roseline Fodouop Tekeu, Project Manager and Project Associate, respectively, for providing management and logistical support from Creative Associates' home office; to the National Curriculum Development Center (NCDC), especially Baale Remegious, a Curriculum Specialist, for its technical oversight and assistance; and to the teachers, content experts, supervisors, data assistants, and others who gave generously of their time to produce what resulted in assessments of extremely high quality.

And again, in this final phase of the program evaluation, a special note of thanks goes to Dickson Turyareeba, a Research and Policy Analyst with the UNITY project, who continued to be available at all times in Uganda to ensure the smooth functioning of each step of a large, complex, and time-bound process.

Sincerely,

The Staff and Consultants of School-to-School International
For Creative Associates International, Inc.
March 2011

Acronyms

Creative	Creative Associates International, Inc.
CC	Coordinating Center
MLA	Measuring Learning Achievement
MoES	Ministry of Education and Sports
NCDC	National Curriculum Development Center
PIASCY	Presidential Initiative on AIDS Strategy for Communication to Youth
PMP	Project Monitoring Plan
PTC	Primary Teachers' College
STS	School-to-School International
TDMS	Teacher Development Management System
UNEB	Uganda National Examinations Board
UNITY	Uganda Initiative for TDMS and PIASCY
USAID	United States Agency for International Development

Executive Summary

Overview

The Uganda Measuring Learning Achievement (MLA) is an assessment of pupil achievement conducted for the USAID-funded Uganda Initiative for TDMS and PIASCY (UNITY) Project.[1] The purpose of MLA is to determine the extent to which pupil learning has changed with the implementation of the new curriculum launched in 2008. The first UNITY MLA was conducted in 2007, when UNITY, in collaboration with the Uganda National Examinations Board (UNEB), the National Curriculum Development Center (NCDC), and the Ministry of Education and Sport (MoES), tested pupils in 2 districts in each of 4 regions to create a baseline in literacy (language) and numeracy (maths) at the P2 level - the last year in which the old curriculum was used at that level. Subsequent tests were conducted in the same schools and additional ones in 2008, 2009, and 2010.

Design

MLA 2010 measured P4 pupils' achievement in literacy and numeracy in English. This permitted two analyses:

1. Of the first cohort of pupils (Cohort 1) in P4 compared to the second cohort of pupils (Cohort 2) in P4. The Cohort 1 pupils started in P1 in 2006 and then progressed to P4 in 2009; this was the last group with instruction in the old curriculum. The Cohort 2 pupils were in P1 in 2007 and then progressed to P4 in 2010; this was the first group with instruction in the new curriculum.
2. Of Cohort 1 and Cohort 2 pupils in experimental and control groups. In Cohort 1, both the experimental and control groups received instruction in the old curriculum. In Cohort 2, the two groups received instruction in the new curriculum though the experimental group's instruction was mostly in local languages while the control group's instruction was in English. The purpose of Cohort 1 was to set a baseline during the last year of the old curriculum. The purpose of Cohort 2 was to examine the first group with the new curriculum. Since the new curriculum was implemented nationwide, it was not possible to design a study with schools placed into each group at random. The control group was actually a comparison group comprised of private schools, for the

[1] Creative Associates International, Inc. is the prime contractor for the UNITY Project. Creative engaged the services of School-to-School International to implement the MLA from 2007 to 2010. This is the final report.

most part, and a few public schools in which the new curriculum was implemented in English.

Implementation

Starting with MLA 2007, pupil achievement tests were administered to Cohort 1, both the experimental and control groups, in P2 through P4. All of these tests were constructed in English, since these pupils were provided instruction at all grade levels under the old curriculum in English. The project then created forms of the same test in 6 local languages for P2 and P3 for implementation with Cohort 2, along with interview instruments for Head Teachers and P4 teachers. For each test administration, local teams scored the tests and entered test and interview data, all of which was sent to the U.S. for analysis by School-to-School International (STS). MLA report writing was subsequently conducted by STS.

Key Findings

Several key findings from the MLA 2010 emerged from the analysis of the pupil achievement testing results and the interview results:

1. Clear trends in overall performance -- the cross-sectional analysis in P4 of Cohort 1 2009 vs. Cohort 2 2010, as well as the Cohort 2 panel analysis that tracked pupils from P2 in 2008 to P4 in 2010, showed similar results. The gains in the control schools were fairly consistent from P2 to P4. However, in the experimental schools, there were larger gains from P2 to P3 when using the new curriculum and then a large drop-off from P3 to P4 also when using the new curriculum. The latter change was most likely due to difficulties on the part of pupils and their teachers with the transition from local language to English. The encouraging part is that pupils made much progress when receiving instruction in local language with the new curriculum. The discouraging part was that these same pupils gave back most of their gains. The hope is that the pupils and their teachers can build on their positive early experiences with local language as they progress into P5 and P6, so that Uganda can follow the literature showing that pupils often experience a drop-off during the language transition year but that they recoup their gains in the upper primary grades.[2]
2. Smaller drop-off in maths than in language -- the decline in scores in P4 was less in maths than in language for the experimental group. In other words, the

[2] Benson, C. (2004). *The importance of mother tongue-based schooling for educational quality*. Paris: UNESCO. (For the *Education for all global monitoring report, 2005*).

pupils had less difficulty with the transition from local language to English when learning maths as opposed to English. This is likely due to the similarities in maths in local language and English, at least for numbers and symbols. With language, however, almost everything is different in the two languages – words, pronunciation, accents, phonetics, grammar, syntax, etc.

3. Correlations between scores and other factors – disaggregated data revealed that there were disparities in scores based on region, pupil age, repetition status, availability of books, and mother's literacy. There is room for improvement among pupils who scored lower due to these factors, so that compensatory programs may have chances of success.
4. Gender equity in scores – one factor that did not generally correlate with achievement was gender. It is encouraging that this trend is evident in Uganda, as it is in many other countries, and it is likely due to the long-term presence of strong girls' education programs.
5. Enthusiasm for the new curriculum – most of the teachers and Head Teachers said that they prefer the new curriculum, in spite of difficulties they reported with insufficient training and materials (see below). They said that instruction in local languages was more practical in terms of building on children's prior knowledge and was more consistent with the children's environments.
6. Demand for more training and materials – the interviews showed that the training on the new curriculum was too short and that there were not enough instructional materials in local languages. Also mentioned was difficulty in implementing classroom-based continuous assessment due to large class sizes, which is an issue that could also be addressed through training and materials.
7. Need for additional data – the story of the implementation of the new curriculum does not end with P4. It would be very helpful to have additional data at P5 and P6 in order to follow the trends seen in other countries on local language learning and transition into the international language. This trend could be expected in Uganda, particularly given the strong gains shown by the pupils from P2 to P3 and their proven ability to learn if they understand the language. The learning concepts and styles that the pupils benefitted from while receiving instruction in their local language should "kick in" again as they transition through P4 and progress to P5 to P6. However, the only way to know whether the transition successfully takes place in Uganda is through the collection of additional pupil achievement data at the upper primary grades.

Summary of Key Recommendations

Most of the key recommendations from the study emerged in the key findings and they have been mentioned above, but this summary elaborates on them as follows:

1. More teacher and head teacher training -- in light of concerns raised by teachers and Head Teachers about the training being too short in duration, more training would likely strengthen the outcomes of the reform. This is similar to trends in the U.S. and other countries where the training needs to be of a long enough duration and intensity in order to have both immediate and lasting effects. However, the training should be targeted towards issues such as improved teaching in local languages and mitigating difficulties in transitioning to English. Training should also reinforce teachers' understanding and practice in the critical area of classroom-based continuous pupil assessment.
2. More teacher and pupil materials -- strategic procurement and use of materials would likely benefit teachers and pupils alike. It is recommended that the MoES looks seriously at investments in specific materials and books, including additional teachers' guides and better translations into local languages, and that a program be developed to assist teachers and Head Teachers with strategies for effective use of materials.
3. Development of strategies to reduce disparities – as seen by successful girls' education programs, it is possible to design and implement successful strategies to reduce disparities in achievement due to factors such as region, pupil age, repetition, and availability of materials.
4. Research on achievement trends and language of instruction -- it would be useful to conduct additional studies of the reasons behind the large gains in achievement when pupils are learning in local languages, and the pupils' difficulties in their transition from local language to English. Another issue to examine is the existence of multiple local languages in schools and the kinds of support teachers, Head Teachers and parents need in these situations.
5. Additional data in higher grade levels -- it is worth reiterating the need for follow-up studies of the pupils as they progress from P4 to P5 and P6. Clearly, these data would provide a wealth of information that could lead to improved implementation of the new curriculum and additional strategies. This in turn could help pupils with the transition from local language to English as the medium of instruction, in both language and maths.

Introduction

The Uganda Initiative for TDMS and PIASCY (UNITY) Project is a USAID-funded initiative implemented by Creative Associates International, Inc. (Creative). The main goals of the UNITY Project (UNITY) are the improvement of teaching, learning, and health for primary school children throughout Uganda. One aspect of UNITY focuses on pre-service and in-service training in order to support the Ministry of Education and Sports (MoES) in its implementation of the new national curriculum. UNITY's Project Monitoring Plan identifies the measure of the success of these efforts in the following indicator:

> At least 70 percent of surveyed children demonstrate higher levels of learning achievement as a result of pre- and in-service training activities.

In order to demonstrate that higher levels of learning achievement have occurred as a result of project interventions, UNITY initiated a pupil testing effort in 2007, called Measuring Learning Achievement (MLA). This report presents the results of the fourth test administration, or MLA 2010, which was conducted in the field September/October 2010 with P4 pupils in all four geographic regions of Uganda and through Head Teacher and P4 teacher interviews conducted in October 2010. This document constitutes the first part of the report, or the "Core Report." It begins by describing the design of the exercise, and then presents the findings from the 2010 MLA, followed by a discussion of salient findings and recommendations for future MLA exercises and curriculum implementation. In the annexes can be found a more detailed description of the assessment's methodology, including sampling, test development, data collection, scoring, data entry, and analysis. The second part of the report, called the "Technical Report," presents additional tables, charts, and explanations of more technical aspects of the exercise.

Methodology

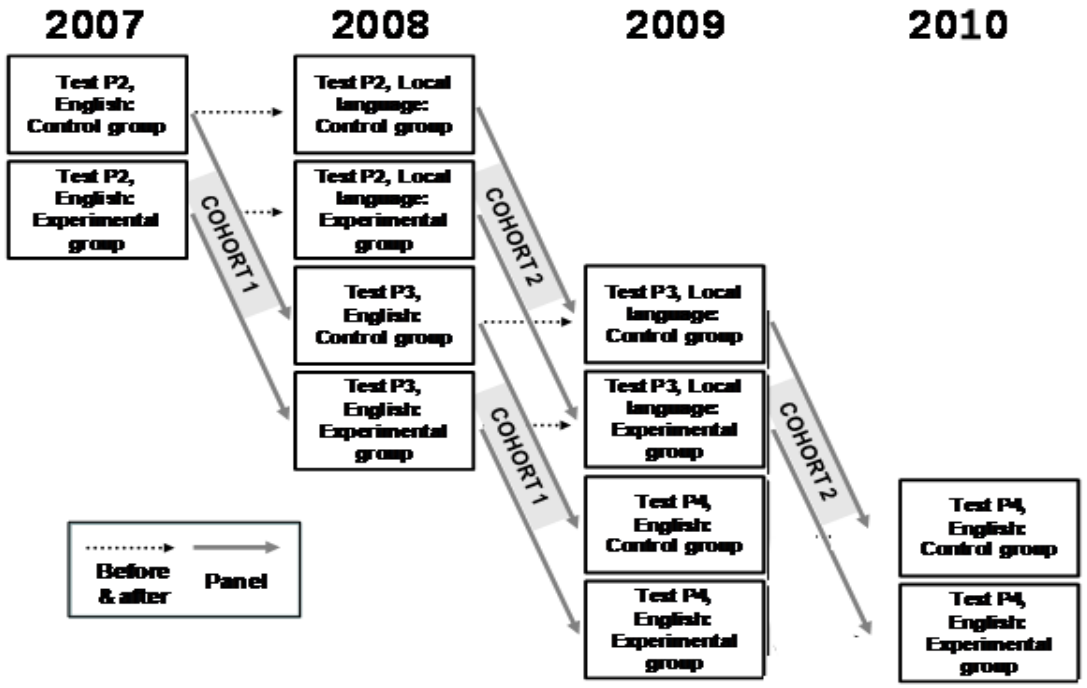
The UNITY MLA was designed to consist of four rounds of tests to be conducted from 2007 to 2010 to demonstrate change over time for two cohorts of pupils. In terms of instruction, the difference between the two cohorts was that Cohort 1, which was assessed from 2007 to 2009, was the last group of pupils to use the old curriculum, from P2 (2007) to P4 (2009). In contrast, Cohort 2, which was assessed

from 2008 to 2010, used the new curriculum, from P2 (2008) to P4 (2010). The old curriculum used English as the main medium of instruction at all grade levels. The new curriculum was designed to use local languages as the main medium of instruction up to P3, with a transition to English as the main medium of instruction starting in P4.

For Cohort 1, all testing took place in English, from P2 to P4, in language and maths, for both the experimental and control groups. The control group was comprised of pupils in schools (mostly private) that also used the old curriculum.

For Cohort 2, testing for the experimental group took place in local languages in P2 and P3, and in English in P4. The tests were the same as the English-based tests used in Cohort 1 except for the translation. Six local languages were used in the testing in the four regions. Again, a control group used pupils in schools (mostly private) where English (under the new curriculum) was still being used as the medium of instruction.

Figure 1: MLA design, 2007 – 2010



As in previous years, the tests were administered in September and October, reaching 2,276 P4 pupils in 116 schools, with a sample of approximately 20 pupils per school (Table 1).

Table 1: Number of schools and P4 pupils tested in 2010

	Schools			Pupils		
Region	Experimental	Control	Total	Experimental	Control	Total
Central	21	9	30	419	180	599
East	21	8	29	420	132	552
North	20	9	29	399	181	580
West	21	7	28	405	140	545
Total	83	33	116	1,643	633	2,276

To provide greater insight on the implementation of the new curriculum, interviews were conducted with Head Teachers and P4 teachers in October 2010. The interviews covered key domains such as the teacher training and the content of the new curriculum. From the 116 schools, 100 Head Teachers and 107 P4 teachers participated in the interviews.

For information concerning sampling, test construction, administration, scoring, data entry and analysis, see Annex 1.

Findings

Five types of analyses were conducted for the MLA 2010:

- A summary of P4 pupil performance for 2010;
- A summary of the Head Teacher and teacher interviews for 2010;
- A comparative analysis of P4 pupil performance from 2009 to 2010 in experimental and control schools;
- An analysis of progress for Cohort 2 from P3 2009 to P4 2010; and
- An analysis of progress for a subsample of Cohort 2 (panel design) – from P2 2008 to P3 2009 to P4 2010.

The results of these analyses are presented in the following sections.

P4 Pupil Performance for 2010

This summary refers to pupils in both experimental and control schools (combined). The results are disaggregated by demographic and other variables.

Geography

By region, pupil scores in both language and maths were significantly higher in the Western Region than in the other three regions. Results in the East Region were

significantly lower in language and maths. Note that only two districts per region were sampled (Figures 2 and 3).

By district, maths and language scores in Mbarara were significantly higher than all other districts, with Kumi district reporting the lowest scores both in language and in maths (Tables 4 and 5).

Figure 2: P4 mean language scores by region

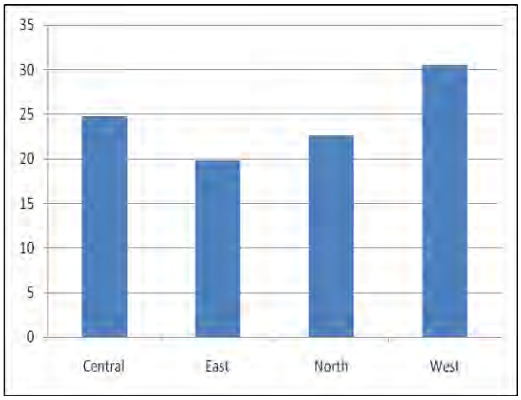


Figure 3: P4 mean maths scores by region

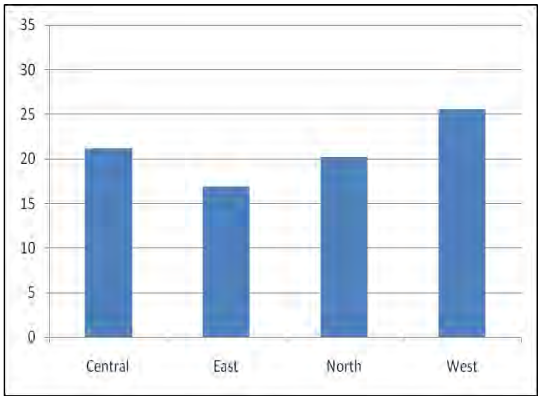


Table 2: P4 mean language scores by region

Region	Mean	Std. Deviation	N
Central	24.79	11.20	599
East	19.83	14.21	552
North	22.63	15.41	580
West	30.54	12.54	545
Total	24.42	13.96	2276

Table 3: P4 mean maths scores by region

Region	Mean	Std. Deviation	N
Central	21.18	10.23	599
East	16.91	11.28	552
North	20.20	13.36	580
West	25.58	11.45	545
Total	20.95	12.01	2276

Table 4: P4 mean language scores by district

District	Mean	Std. Deviation	N
Gulu	20.02	14.33	280
Kabale	26.24	13.41	260
Kumi	15.86	12.14	257
Lira	25.07	15.99	300
Mbarara	34.47	10.24	285
Mpigi	24.08	10.26	299
Mukono	25.50	12.04	300
Soroti	23.28	14.97	295
Total	24.42	13.96	2276

Table 5: P4 mean maths scores by district

Region	Mean	Std. Deviation	N
Gulu	17.24	12.13	280
Kabale	22.72	11.77	260
Kumi	13.95	10.81	257
Lira	22.96	13.87	300
Mbarara	28.18	10.51	285
Mpigi	21.15	9.93	299
Mukono	21.21	10.54	300
Soroti	19.48	11.06	295
Total	20.95	12.01	2276

Language at Home

Pupils who said their first language was Runyankole scored significantly higher than other groups in both language and maths while the pupils from the Acoli and

the Ateso language groups scored the lowest (Tables 6 and 7).

Table 6: P4 mean language scores by language at home

Language at home	Mean	SD	N
Acoli	19.97	14.27	257
Ateso	19.96	14.08	518
Lango	24.01	15.97	303
Luganda	24.68	11.15	586
Rukiga	25.88	12.78	293
Runyankole	36.01	9.79	227
Other	28.29	15.40	92
Total	24.42	13.96	2276

Table 7: P4 mean maths scores by language at home

Language at home	Mean	SD	N
Acoli	16.96	11.88	257
Ateso	17.06	11.34	518
Lango	22.12	13.78	303
Luganda	20.93	10.04	586
Rukiga	22.35	11.55	293
Runyankole	29.63	10.40	227
Other	24.29	12.78	92
Total	20.95	12.01	2276

Age

Pupils under 10 years old performed significantly better than pupils 10 and older in both language and maths (Tables 8 and 9).

Table 8: P4 mean language scores by age group

Age	Mean	Std. Deviation	N
Under 10 Years Old	34.29	12.44	286
10 Years Old	26.88	13.97	717
11 Years Old	23.04	13.47	483
12 Years Old	20.45	13.07	434
12+ Years Old	18.97	11.47	305
Total	24.66	13.97	2225

Table 9: P4 mean maths scores by age group

Age	Mean	Std. Deviation	N
Under 10 Years Old	26.87	11.25	286
10 Years Old	22.63	12.37	717
11 Years Old	19.95	11.61	483
12 Years Old	18.76	11.54	434
12+ Years Old	17.78	10.73	305
Total	21.17	12.01	2225

Repetition

Non-repeating pupils performed significantly better in language and maths than repeating ones (Table 10).

Table 10: P4 mean language and maths scores by repeater status

Subject	Repeater	N	Mean	Std. Deviation
Language	Yes	289	18.89	11.72
	No	1935	25.59	14.02
Maths	Yes	289	17.26	10.71
	No	1935	21.80	12.05

Home Environment

Pupils with books at home (Table 11) and with mothers who read (Table 12) performed significantly better in language and maths than those without books or whose mothers do not read.

Table 11: P4 mean language and maths scores by books at home

Subject	Books at home	N	Mean	Std. Deviation
Language	Yes	1662	27.65	13.58
	No	564	16.04	11.05
Maths	Yes	1662	23.37	11.86
	No	564	14.80	9.89

Table 12: P4 mean language and maths scores by mother who reads

Subject	Mother reads	N	Mean	Std. Deviation
Language	Yes	1485	27.75	13.96
	No	732	18.76	11.70
Maths	Yes	1485	23.46	12.17
	No	732	16.74	10.25

Gender

The difference in mean language scores between girls and boys was very small and not statistically significant. The difference in mean maths scores was also small, though it did show statistically significant results in favor of the boys (Table 13). Analyses disaggregated by region show that the difference between boys and girls in maths was statistically significant (in favor of boys) only in the Northern region. When disaggregated by language at home, only Acoli boys had significantly better results than Acoli girls in maths (not shown).

Table 13: P4 mean language and maths scores by sex of the pupil

Subject	Sex	N	Mean	Std. Deviation
Language	Girl	1136	24.43	14.06
	Boy	1140	24.40	13.87
Maths	Girl	1136	20.09	11.77
	Boy	1140	21.80	12.20

Experimental vs. Control Groups

Scores in P4 pupils in control schools (following the new curriculum in English) were significantly higher than those in experimental schools (following the new curriculum but primarily in local languages up to P3 and then transitioning to English in P4), both in language and maths (Table 14).

Table 14: P4 mean language and maths scores by group

Subject	Group	N	Mean	Std. Deviation
Language	Control	633	38.87	8.01
	Experimental	1643	18.85	11.56
Maths	Control	633	32.42	9.30
	Experimental	1643	16.53	9.82

Correlations

In all, 107 P4 teachers were interviewed and provided information on their years of experience, qualifications, class sizes, and the availability of instructional materials. Based on this information, as well as 100 Head Teacher interviews, correlations were found between P4 pupil performance and these characteristics:

- Number of male teachers in school: A statistically significant and positive relation was found between the number of male teachers in a given *school* and P4 language and maths achievement; however, no significant relation was found between pupil scores and the teacher's sex for a given *class*.
- Library in a school: There were no significant differences in P4 scores in language or maths when there is a library in the school (Table 15).

Table 15: Mean achievement by library in the school

Subject	Existence of library	N	Mean	Std. Deviation
Language	Yes	66	26.93	12.22
	No	38	20.80	11.64
Maths	Yes	66	22.76	9.92
	No	38	18.02	9.54

- Pupil materials: Pupils with their own rulers and pencils scored significantly higher in language and maths.
- Attitude towards the new curriculum: In schools where P4 teachers preferred the old curriculum, P4 pupils scored significantly higher in maths and in language than pupils whose teachers preferred the new curriculum was or who said that the new curriculum was not different from the old one.
- Significant relationships were not found between P4 pupils' scores and the proportion of boys to girls in a school, the number of boys or girls in a school, teachers' access to materials (e.g., stationery, flash cards, or dictionaries), the teacher's number of years teaching or highest qualification earned, the number of books in their libraries, or the Head Teachers' number of years of experience or highest qualification earned.

Head Teacher and Teacher Interviews for 2010

Head Teachers and teachers were interviewed on the new P3 and P4 curriculums and training programs.

P3 Curriculum and Training

When the Head Teachers were asked whether the training on the new P3 curriculum was sufficient, the main concern was that the trainings were too short. When they were asked how the training could be improved, half of the respondents said that the training should be longer and a quarter said that relevant materials should be provided, especially relevant texts for pupils and reference books for teachers. To prevent issues that arise when teachers are transferred, Head Teachers recommended that more teachers from each school receive training in the new curriculum and that the MoES considers whether teachers have been trained when organizing transfers.

When the Head Teachers were asked the greatest weakness of the new P3 curriculum, the foremost answer was a lack of teaching materials, such as relevant text and reference books. Others said that the local language vocabulary was too limited for teaching the various subject areas. Of the teachers who found it difficult to interpret the new P3 curriculum, over a third reported that they had difficulties translating English into the local language, especially in science. In relation to

continuous assessment, several Head Teachers remarked that it was not realistic to introduce this approach in large classrooms.

Head Teachers reported that the greatest strengths of the new P3 curriculum were enhanced pupil understanding, enthusiasm for learning, and class participation. They took particular notice of increased pupil literacy. A few Head teachers also said the new curriculum was pupil-centered and context specific.

The majority of Head Teachers who preferred the new P3 curriculum said that this was because local language enhanced pupil understanding. Others mentioned that the practical themes allowed pupils to learn from their local environment. Of the Head Teachers who preferred the old curriculum, some said they did so because they could access support materials relevant to the old curriculum and because certain teachers were not trained to teach the new curriculum.

To enhance the new P3 curriculum, almost half of the Head Teachers recommended providing sufficient teaching and learning materials. Most stated that schools required relevant textbooks, especially those in the local languages. They also highlighted the need for reference materials, teaching aids, and thematic curriculum guides. Many Head Teachers also said that there was a need for increased training in the form of continuous professional development. Some suggested that all teachers needed to be trained due to the high frequency of teacher transfers.

P4 Curriculum and Training

Of the Head Teachers and teachers who found the new P4 curriculum training inadequate, most said that the time allocated for training was too short. Some Head Teachers and teachers said that the large amount of travel required to reach the training centers was costly and difficult. Others felt that the program suffered from a lack of materials and reference books.

When the Head Teachers and teachers were asked how the training could be improved, a common response was by providing more materials, especially texts for pupils and reference books for teachers. Many also said the program should have refresher trainings and continuous professional development opportunities for

teachers. It was also noted that this continuous training would provide an opportunity for monitoring the implementation of the new curriculum. A few Head Teachers suggested scheduling the training over the holidays to avoid withdrawing teachers from their schools.

Of the teachers who found it difficult to interpret the new P4 curriculum, many said that this was due to the lack of materials, such as text and reference books. Half of the Head Teachers, who also reported having difficulty, attributed their challenges to not receiving training in the new curriculum. A difficulty that both Head Teachers and teachers cited was their pupils' inability to understand English.

For teachers, the greatest strength of the new P4 curriculum was the fact that it integrated the English language. Over a third of Head Teachers agreed that it was important for pupils to start using English at least by P4. In addition, Head Teachers and teachers thought that a strong point of the new curriculum was the use of simplified content. It was reported that as a result of this new curriculum pupils' understanding and participation increased. As with the new P3 curriculum, teachers took particular notice of improvements in pupils' literacy.

When asked the greatest weakness of the new P4 curriculum, almost half of the teachers said there was a lack of relevant text and reference books. They also noted that pupils struggled with the English language. Many Head Teachers agreed with these critiques. A few Head Teachers and teachers also said that large class sizes and the absence of guides made continuous assessment extremely difficult.

Most Head Teachers and teachers preferred the new P4 curriculum to the old P4 curriculum, generally because it was simpler, more user-friendly, and it was better at improving pupil understanding. A minority of teachers favored the old curriculum because supplementary materials, such as textbooks and reference books, were readily available. Some also preferred the old curriculum because they thought that the new curriculum was confusing and required more work.

When teachers and Head Teachers were asked for their suggestions on enhancing the new P4 curriculum, the overwhelming response was the need to increase materials such as reference books, textbooks, training materials, and teaching aids.

The next most common suggestion was the need to train all teachers, increase the time allotted for training, and provide regular refresher courses.

Teachers and Head Teachers also provided the following recommendations:

- Monitor the implementation of the new curriculum;
- Sensitize parents and community members to the new curriculum;
- Train pre-service teachers on the use of the new curriculum; and
- Align the new curriculum with the examinations.

P4 Pupil Performance from 2009 to 2010

Because the overall purpose of this assessment was to determine whether similar pupils learn better under the new curriculum than the old, this section examined how P4 intervention pupils performed in 2010 using the new curriculum compared to how P4 pupils performed in 2009 using the old. This section presents the results of the analyses conducted in these two years, first through global comparisons, then through comparisons of control and experimental groups in greater detail.

To make this comparison of two cohorts of pupils (Cohort 1 = 2009 with the old curriculum vs. Cohort 2 = 2010 with the new curriculum) in the same school year (P4), the analysis used a procedure called test equating. To ensure that scores from year to year are comparable, scores from tests must be equated – otherwise, it is not possible to know if the tests were of equal difficulty. Equating is a process that links the two tests and produces equivalent scores (called "expected true scores") so we can measure pupils' performance *as if they had taken the same tests*. This is done by linking the tests through the use of "anchor items" – a subset of items common to each test. These anchor items serve as a reference against which the difficulty of the two tests can be measured, and then the scores are adjusted to so they can be compared on the same scale. This equating procedure was used here to put the scores (called expected true scores) of the two cohorts on the same scale.

A look at the expected true scores of P4 pupils (Table 16) shows that the 2009 cohort (old curriculum) received higher scores than the 2010 cohort (new curriculum) both in language and maths. This measure combines the scores of experimental and control groups. This difference in performance between the two

cohorts was statistically significant both subjects (Table 17).

Table 16: Mean comparison between the 2009 and the 2010 P4 cohorts, experimental and control groups combined

Subject	Group Year	N	Mean	Std. Deviation
True_Language	C1 2009	2239	61.70	29.91
	C2 2010	2276	50.90	35.11
True_Maths	C1 2009	2239	55.10	19.39
	C2 2010	2276	41.95	22.67

Table 17: T-tests comparing the true scores of the 2009 and the 2010 P4 cohorts, experimental and control groups combined *(equal variances not assumed)*

Subject	Levene's Test for Equality of Variances		t-test for Equality of Means		
	F	Sig.	t	df	Sig. (2-tailed)
True_Language	1.91	.000	11.14	4.42	.000
True_Maths	1.09	.000	20.96	4.42	.000

General comparisons such as these, however, do not capture the different types of changes registered by pupils studying in English (control group) compared to pupils studying in local language (experimental group). For the Cohort 1 (2009) pupils, both the experimental and control groups used the old curriculum through P4. For Cohort 2 (2010), the experimental pupils used the new curriculum; they received instruction for the most part in any of the six local languages (Acoli, Ateso, Lango, Luganda, Rukigafor, Runyankole) through P3, and then transitioned to English in P4. Both groups took the P4 tests in English, for language and maths.

For the control groups, it was expected that there would be no significant differences in the 2009 and 2010 scores. In language, this was the case in that the scores in 2009 and 2010 were virtually the same. In maths, however, the 2009 scores were higher; the reason for this difference is not clear (Table 18).

Table 18: Mean achievement of the two cohorts of P4 pupils for the control and the experimental groups

Group/Subject		Group Year	N	Mean	Std. Deviation
Control	True_Language	C1 2009	657	86.45	16.70
		C2 2010	633	86.68	18.19
	True_Maths	C1 2009	657	68.82	14.18
		C2 2010	633	63.84	16.65
Experimental	True_Language	C1 2009	1582	51.42	28.12
		C2 2010	1643	37.11	29.94
	True_Maths	C1 2009	1582	49.40	18.39
		C2 2010	1643	33.52	18.69

The main comparison in Table 18 involved the experimental groups in 2009 and 2010, since they had different curriculums. In both language and maths, the 2009 experimental group (Cohort 1 or old curriculum) scored significantly higher than the 2010 experimental group (Cohort 2 or new curriculum).

Geography

By region, the difference in performance between the two cohorts in the control group was generally not significant in language but it was significant in maths. Differences in experimental group were statistically significant for all regions, ranging from 9 to 23 points for language and 12 to 15 points for maths (Table 19).

Table 19: Mean achievement by region of the two cohorts of pupils for the control and the experimental groups

Group	Region/Subject		Group Year	N	Mean	Std. Deviation
Control	Central	True_Language	C1 2009	179	80.50	16.72
			C2 2010	180	79.29	18.15
		True_Maths	C1 2009	179	63.58	13.56
			C2 2010	180	57.69	15.37
	East	True_Language	C1 2009	158	86.99	18.52
			C2 2010	132	86.22	20.15
		True_Maths	C1 2009	158	71.49	13.72
			C2 2010	132	62.47	15.63
	North	True_Language	C1 2009	180	89.60	15.23
			C2 2010	181	92.10	12.80
		True_Maths	C1 2009	180	71.23	15.16
			C2 2010	181	68.97	16.50
	West	True_Language	C1 2009	140	89.39	14.34
			C2 2010	140	89.59	19.17
		True_Maths	C1 2009	140	69.41	12.42
			C2 2010	140	66.41	16.80
Experimental	Central	True_Language	C1 2009	403	52.85	23.46
			C2 2010	419	41.39	26.12
		True_Maths	C1 2009	403	51.79	15.09
			C2 2010	419	36.20	17.49
	East	True_Language	C1 2009	379	47.63	26.37
			C2 2010	420	24.59	24.48
		True_Maths	C1 2009	379	46.88	18.70
			C2 2010	420	25.70	14.12
	North	True_Language	C1 2009	396	33.97	25.52
			C2 2010	399	24.82	25.22
		True_Maths	C1 2009	396	39.90	18.62
			C2 2010	399	27.18	16.15
	West	True_Language	C1 2009	404	70.67	24.02
			C2 2010	405	57.78	30.51
		True_Maths	C1 2009	404	58.70	15.62
			C2 2010	405	45.10	19.86

Pupils' Age

The difference in performance of pupils for control schools between P4 2009 and P4 2010 was statistically significant only for the 10 and 11 year-olds in maths. For experimental schools, all comparisons between the two cohorts were found to be statistically significant in language and maths for most age groups (Table 20).

Table 20: Mean achievement by age of the two cohorts of pupils (control and the experimental groups)

Group	Age/Subject		Group Year	N	Mean	Std. Deviation
Control	Under 10 Years Old	True_Language	C1 2009	185	89.22	14.35
			C2 2010	155	91.67	12.48
		True_Maths	C1 2009	185	68.16	13.55
			C2 2010	155	64.20	15.21
	10 Years Old	True_Language	C1 2009	238	85.90	18.30
			C2 2010	255	85.89	19.42
		True_Maths	C1 2009	238	68.29	14.92
			C2 2010	255	63.55	17.06
	11 Years Old	True_Language	C1 2009	122	85.75	16.77
			C2 2010	111	85.19	17.82
		True_Maths	C1 2009	122	71.49	13.78
			C2 2010	111	62.40	16.76
	12 Years Old	True_Language	C1 2009	70	82.79	16.55
			C2 2010	67	87.42	15.16
		True_Maths	C1 2009	70	67.67	13.87
			C2 2010	67	66.24	16.31
	Over 12 Years Old	True_Language	C1 2009	39	84.67	16.22
			C2 2010	37	82.11	22.85
		True_Maths	C1 2009	39	69.08	14.37
			C2 2010	37	68.92	17.32
Experimental	Under 10 Years Old	True_Language	C1 2009	163	63.63	27.79
			C2 2010	131	55.98	33.22
		True_Maths	C1 2009	163	52.75	17.61
			C2 2010	131	40.36	20.10
	10 Years Old	True_Language	C1 2009	427	54.07	28.00
			C2 2010	462	40.82	31.37
		True_Maths	C1 2009	427	50.11	17.39
			C2 2010	462	35.03	19.70
	11 Years Old	True_Language	C1 2009	295	50.91	28.35
			C2 2010	372	36.52	29.16
		True_Maths	C1 2009	295	49.59	19.23
			C2 2010	372	33.39	18.48
	12 Years Old	True_Language	C1 2009	379	47.50	27.89
			C2 2010	367	32.67	28.44
		True_Maths	C1 2009	379	48.00	18.79
			C2 2010	367	32.41	18.36
	Over 12 Years Old	True_Language	C1 2009	310	46.69	26.35
			C2 2010	268	31.49	25.17
		True_Maths	C1 2009	310	48.05	18.69
			C2 2010	268	31.50	16.35

Pupils by Repeater Status

In experimental schools, both repeaters and non-repeaters showed significant differences between 2009 and 2010 both in language and maths. Differences in the control schools were not significant for repeaters (Table 21).

Table 21: Mean achievement by repeating year of the two cohorts of pupils for the control and the experimental groups

Group	Repeat/Subject		Group Year	N	Mean	Std. Deviation
Control	Yes	True_Language	C1 2009	62	83.69	17.71
			C2 2010	50	76.06	24.62
		True_Maths	C1 2009	62	70.13	12.78
			C2 2010	50	60.20	19.17
	No	True_Language	C1 2009	591	86.76	16.60
			C2 2010	582	87.61	17.26
		True_Maths	C1 2009	591	68.72	14.31
			C2 2010	582	64.15	16.41
Experimental	Yes	True_Language	C1 2009	373	47.27	26.61
			C2 2010	239	28.92	23.90
		True_Maths	C1 2009	373	47.06	18.24
			C2 2010	239	29.44	15.92
	No	True_Language	C1 2009	1186	53.23	28.32
			C2 2010	1353	39.27	30.74
		True_Maths	C1 2009	1186	50.50	18.23
			C2 2010	1353	34.78	19.07

Pupils with Books at Home and Mothers Who Read

In experimental schools P4 pupils got significantly lower scores in 2010 than in 2009 both in language and maths (Tables 22 and 23).

Table 22: Mean achievement by books at home of the two cohorts of pupils for the control and the experimental groups

Group	Books in home/Subject		Group Year	N	Mean	Std. Deviation
Control	Yes	True_Language	C1 2009	568	87.16	16.03
			C2 2010	574	88.25	15.99
		True_Maths	C1 2009	568	68.93	14.29
			C2 2010	574	64.31	16.26
	No	True_Language	C1 2009	87	81.63	20.17
			C2 2010	58	71.28	28.77
		True_Maths	C1 2009	87	68.54	13.33
			C2 2010	58	59.10	19.82
Experimental	Yes	True_Language	C1 2009	1148	53.88	28.20
			C2 2010	1088	43.52	30.93
		True_Maths	C1 2009	1148	50.87	18.24
			C2 2010	1088	37.25	19.16
	No	True_Language	C1 2009	415	45.76	26.76
			C2 2010	506	25.22	23.65
		True_Maths	C1 2009	415	46.10	18.01
			C2 2010	506	26.83	15.57

Table 23: Mean achievement by mother who reads of the two cohorts of pupils for the control and the experimental groups

Group	Mother who reads/Subject		Group Year	N	Mean	Std. Deviation
Control	Yes	True_Language	C1 2009	490	87.87	15.71
			C2 2010	513	89.27	15.71
		True_Maths	C1 2009	490	69.74	13.85
			C2 2010	513	65.31	16.14
	No	True_Language	C1 2009	163	82.21	18.77
			C2 2010	119	75.74	23.36
		True_Maths	C1 2009	163	66.25	14.83
			C2 2010	119	57.62	17.50
Experimental	Yes	True_Language	C1 2009	965	54.48	28.18
			C2 2010	972	43.23	31.63
		True_Maths	C1 2009	965	50.81	18.17
			C2 2010	972	36.98	19.65
	No	True_Language	C1 2009	588	47.74	27.24
			C2 2010	613	29.37	25.13
		True_Maths	C1 2009	588	47.87	18.31
			C2 2010	613	29.28	16.18

Pupil's Sex

The difference in performance of P4 girls and boys in control schools between 2009 and 2010 was minor, except in maths where there were significant differences for both. In experimental schools, both P4 girls and boys had significant differences in language and maths scores between 2009 and 2010 (Table 24).

Table 24: Mean achievement by sex of the two cohorts of pupils for the control and the experimental group

Group	Sex/Subject		Group Year	Mean	Std. Dev.	Std. Error
Control	Girl	True_Language	C1 2009	87.33	15.48	0.87
			C2 2010	87.36	17.60	1.03
		True_Maths	C1 2009	67.96	15.03	0.84
			C2 2010	62.09	17.26	1.01
	Boy	True_Language	C1 2009	85.62	17.75	0.96
			C2 2010	86.10	18.67	1.01
		True_Maths	C1 2009	69.63	13.32	0.72
			C2 2010	65.32	16.00	0.86
Experimental	Girl	True_Language	C1 2009	53.51	27.56	0.99
			C2 2010	38.43	30.98	1.07
		True_Maths	C1 2009	49.23	17.99	0.65
			C2 2010	33.02	18.75	0.64
	Boy	True_Language	C1 2009	49.46	28.54	1.00
			C2 2010	35.72	28.76	1.02
		True_Maths	C1 2009	49.56	18.77	0.66
			C2 2010	34.05	18.63	0.66

Progress for Cohort 2 from P3 (2009) to P4 (2010)

This part of the analysis concerns the comparison of scores between control and experimental groups in Cohort 2 as the pupils progressed from 2009 to 2010. The MLA design called for an assessment of the progress made by Cohort 2 in P3 (2009) using the new curriculum as they move to P4 (2010). To make this comparison, the test equating or linking procedure described in the previous section was used, so that the P4 2010 scores are linked to the P3 2009 score scale.

Table 25 below shows the "expected true" or equated scores – the ones used to compare P3 2009 scores with the P4 2010.[3] All scores are reported as percentages. Using these expected true scores, the following mean scores (Table 25) show that the pupils from the control schools where the language of instruction was English had significantly higher scores in P4 (2010) than in P3 (2009) on both English language and maths. On the other hand, the pupils from the experimental schools where the language of instruction (and testing) was one of the six local languages in P3 and then English (in the new curriculum) in P4 had significantly lower scores in P4 (2010) than in P3 (2009) in both language and maths. In other words, the pupils who used local language for instruction in P3 had low mean scores in P4 when the language of instruction was English.

Table 25: Comparing the P4_2010 expected true scores equated to the P3_2009 expected true scores

Group	Score/subject	Grade Year	N	Mean	Std. Deviation
Control	True_language	P3_2009	853	76.80	17.66
		P4_2010	633	89.87	14.72
	True_maths	P3_2009	853	69.76	14.67
		P4_2010	633	80.83	13.16
Experimental	True_language	P3_2009	2980	56.47	27.09
		P4_2010	1643	42.26	33.46
	True_maths	P3_2009	2980	55.99	18.58
		P4_2010	1643	52.68	19.57

[3] This expected true score was obtained through the test characteristic curve using Samejima's graded model found in MULTILOG and GAUSS-IRT software. First, each cohort (2009 P3 and 2010 P4) and each test (language and maths) was calibrated using Samejima's graded model. Then the calibrated tests were used to link/equate the 2010 tests to the 2009 scale (using MULTILOG), with the 2010 theta scores equated to the 2009 theta scale. The theta scores from the two cohorts were transformed to expected true scores using the test characteristic curve of the P3 2009 cohort using the GAUSS-IRT software.

Also note that all four comparisons (shown in Table 26) were found to be statistically significant.

Table 26: T-tests for the comparisons of the P4_2010 expected true scores equated to the P3_2009 expected true scores scale (equal variances not assumed)

Independent Samples Test

			Levene's Test for Equality of Variances		t-test for Equality of Means		
Group/Subject			**F**	**Sig.**	**t**	**df**	**Sig. (2-tailed)**
Control	True language	Equal variances not assumed	62.160	.000	-1.553E1	1464.207	.000
	True maths	Equal variances not assumed	15.533	.000	-1.526E1	1431.909	.000
Experimental	True language	Equal variances not assumed	255.989	.000	1.475E1	2838.936	.000
	True maths	Equal variances not assumed	10.634	.001	5.602	3238.655	.000

Progress of Cohort 2 (Panel) from P2 (2008) to P4 (2010)

The last part of the analysis concerns the panel design for Cohort 2. A total of 673 pupils were followed from 2008 (P2) to 2010 (P4): 219 pupils from the control schools and 454 pupils from the experimental schools.[4] The pupils from the control schools were taught in English with the new curriculum while pupils from the experimental schools were taught in local languages through P3 with new curriculum. All P4 pupils, whether from the control or the experimental schools, were taught with the new curriculum in English.

The basic procedure used to analyse the panel data is called repeated measures analysis and is described in Annex 2.

[4] In fact, of the initial 3,776 P2 pupils, only 673 could be followed individually and tested from 2008 to 2010 for a variety of reasons, including significantly high pupil transfer rates to other schools and high dropout rates. This small proportion (.18) of pupils that could be followed means that findings should be interpreted with caution – see Annex 2 for additional information. Note that the same small proportion of pupils (.18) was found in the panel of cohort 1 (from 2007 to 2009) analysed last year.

Table 27 shows the evolution of the mean scores both in language and maths for Cohort 2 pupils from P2 (2008) to P3 (2009) to P4 (2010). The mean scores are represented graphically in Figure 4 for language and Figure 5 for the maths.

Pupils in control (mostly private) schools consistently scored higher than those in experimental schools both in language and maths, most likely due to a selection effect since there was such a large gap at the baseline. The trends show that the experimental pupils made progress on closing the gap between P2 and P3, only to see the gap widen in P4, as the instruction and testing changed to English. The data show the same tendency observed in the previous parts of this report: P4 experimental pupils had low scores in 2010, the year of language transition in the new curriculum. The statistical details of the repeated measures analyses are shown in Annex 2.

Table 27: Panel results for Cohort 2 pupils: comparing the mean scores from 2008 to 2010 for the control group and the experimental group

Cohort 2	P2 2008	P3 2009	P4 2010
Control group			
True language	95.25	94.77	98.32
True maths	74.82	86.33	91.23
Experimental group			
True language	77.33	86.17	71.53
True maths	61.99	77.99	69.52

Figure 4: Plot of the mean equated/linked language true scores for pupils in the control and the experimental groups (Cohort 2) in P2 2008, P3 2009, and P4 2010

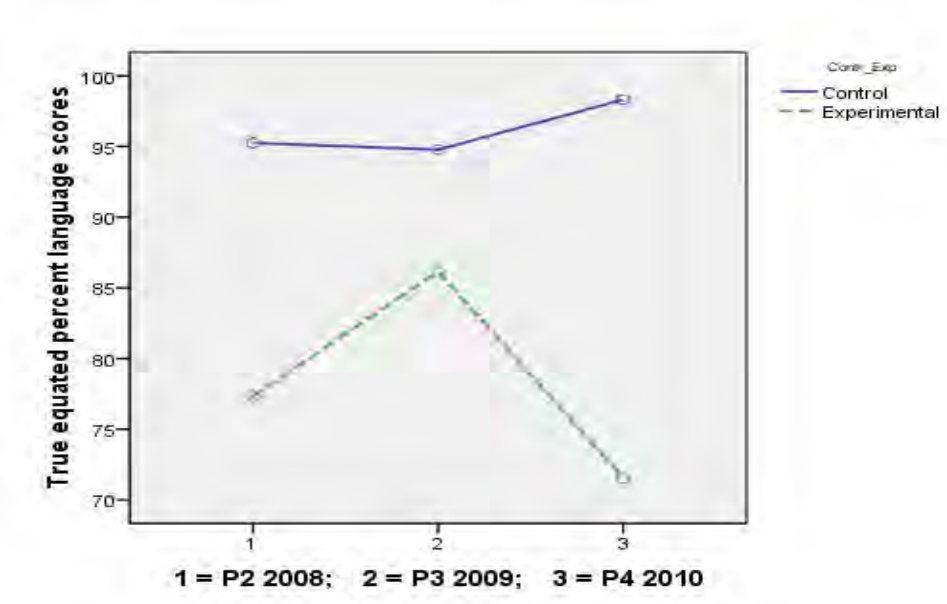
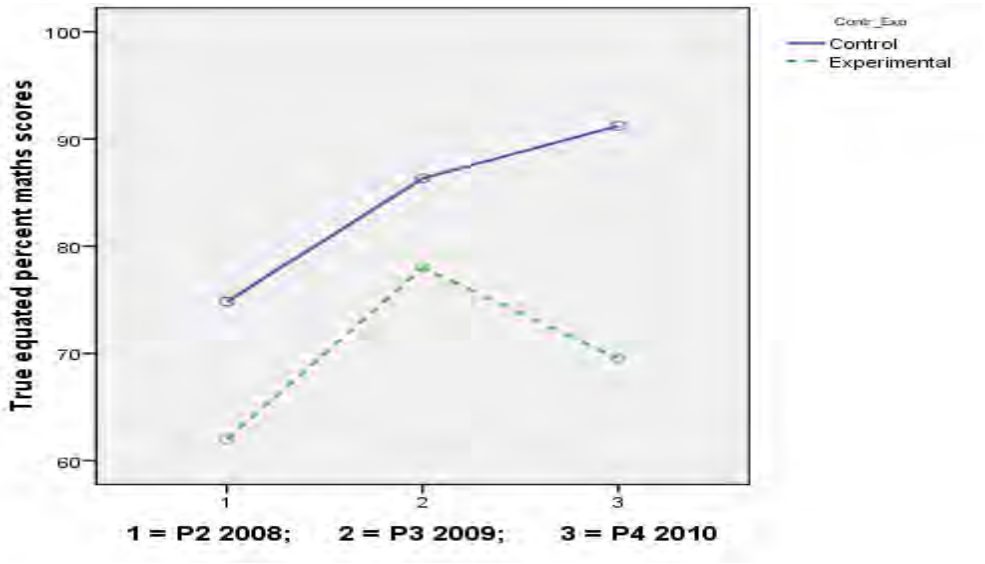


Figure 9: Plot of the mean equated/linked maths true scores for pupils in the control and the experimental groups (Cohort 2) in P2 2008, P3 2009, and P4 2010



Note that the after strong increases by the experimental group in both subjects from P2 (2008) to P3 (2009), the drop-off from P3 (2009) to P4 (2010) for the experimental group was more pronounced in language (~15 percentage points) than in maths (~10 percentage points).

Discussion

This year's MLA revealed a number of tendencies already seen in 2008 and 2009, and several new ones. A discussion of these tendencies is as follows:

Gains from P2 to P3 but Drop-off from P3 to P4

The cross-sectional analysis (P4 in 2009 vs. P4 in 2010) and the panel analysis (P2 to P4 experimental vs. control from 2008 to 2010) discussed in the previous sections show a trend often observed in local language transition programs. The pupils in the experimental group made large gains as they received most of their instruction in the local language that was familiar to them. In fact, these gains were larger than those seen in the control group that received instruction in English throughout the period of observation. This finding is very encouraging and shows that the pupils benefitted from instruction in local languages. However, as the experimental pupils made their transition to English in P4, their scores showed a significant decline. So the new curriculum helped the experimental pupils prior to P4 but then appeared to cause significant problems in P4. The reason for this

decline could be explained in several ways. First, the new language was difficult for the pupils. Second, the pupils did not know how to adapt their learning styles when the language was relatively unfamiliar after having studied in a familiar language for three years. Third, the teachers themselves may have had difficulty with the English language. Fourth, the teachers may not have had the pedagogical skills to provide instruction in the P4 curriculum in English to pupils whose English language skills were relatively low. Fifth, there may not have been enough instructional materials for the pupils in the experimental schools in P4. Sixth, there could have been a problem with a lack of instructional leadership and supervision for the teachers as the pupils (and teachers) made the transition from local language to English in P4.

Smaller Drop-Off in Maths than in Language

Again, as shown in other similar situations, the decline in scores was less in maths than in language. In other words, the pupils had less difficulty with the transition from local language to English when learning maths as opposed to English. This is likely due to the similarities in maths, at least for the numbers and symbols, in local language and English. However, with language, almost everything is different in the two languages – vocabulary, pronunciation, grammar, syntax, etc.

Trends in Achievement Scores and Geography, Age, Repetition Status, Books at Home, Literacy of Mother, and Instructional Materials

Looking at specific factors other than language of instruction is useful for policy-makers. One of these factors is geography. Scores on both language and maths were at least 10 points higher in some regions (e.g., West) than in others (e.g., East). Another is pupils' age. In both language and maths, the scores consistently declines as pupil age increased. The decline was more pronounced in language than in maths. Repeaters also had lower scores in both subject areas. Pupils with instructional materials tended to score higher. These are common trends, but the empirical evidence gives the MoES more data to support programs that encourage parents to send their children to school at an appropriate age, that provide extra instructional support to prevent repetition, and that encourage the availability of reading materials at home for the pupils. In addition, with the transition from local language to English, the trend seen in books at home becomes more nuanced. Pupils in the experimental group with books at home tended to show a lower drop-

off from P3 to P4 than those pupils with no books at home; this shows that the presence of books can help mitigate the pupils' difficulties as they make the transition from local language to English. The same trend was seen with literacy of the mother.

Gender Equity in Achievement Scores

As is increasingly shown in achievement studies, there were no significant differences between the scores of girls and boys in either language or maths. This was generally the case in the control and experimental groups, and at the different grade levels. It is encouraging that this trend is evident in Uganda as it is in many other countries.

Large Percentage of Pupils Showing Learning Gains

As the experimental pupils progressed from P2 to P3, the vast majority (more than 70 percent) demonstrated higher levels of learning achievement as a result of pre- and in-service training activities, as measured by the panel study and the achievement test scores. The average gains of approximately 10 percentage points in language and 15 percent in maths were high, and in fact much higher than the gains by pupils in the control group (though there was most likely a ceiling effect with the language scores for the control group). The fact that all of the gains were "given back" by the experimental pupils in language, as well as most of the gains by the experimental pupils in maths, was predictable from the literature.

Enthusiasm for the New Curriculum and Instruction in Local Languages

Most of the teachers and Head Teachers interviewed said that they prefer the new curriculum, in spite of some of the difficulties with training and materials. They said that instruction in local languages was practical and more consistent with the children's environments.

Demand for Additional Training and Materials

Clear trends from the interviews of teachers and Head Teachers were that the training on the new curriculum was too short and that there were not enough instructional materials in the local languages. This was particularly apparent in two areas: the teachers had difficulty with finding vocabulary that would correspond to some of the curriculum content, and teacher had difficulty with the

transition from local language to English at P4. One other area mentioned by some of the teachers was the difficulty of implementing classroom-based continuous assessment with a lack of training and materials, especially given large class sizes. Additional training and materials would play a large role in helping to alleviate some of these difficulties.

Need for Additional Data

Studies in other countries have shown that pupils usually start showing gains again in P5 and P6 as they become used to learning in the international language. This could be expected in Uganda, particularly given the strong gains shown by the pupils from P2 to P3. The pupils and the teachers, as well as the supervisors, should become more familiar with the English language and teaching methods after they struggle through the first year of the transition. In fact, the learning concepts and styles that the pupils benefitted from while learning in their local language should "kick in" as they make it through P4 and move on to P5 to P6. It would be highly advisable for the Ugandan MoES to track the pupils through two more grade levels to examine the trends in their learning. The data over the next two years are as important as those from the preceding four years.

Recommendations

Based on the findings from the 2010 MLA, recommendations are similar to those from the previous studies, with a few modifications:

Training

In light of concerns raised by teachers and Head Teachers about the training being too short in duration, it seems clear that additional training would strengthen the outcomes of the reform. This is similar to trends in the U.S. and other countries where the training needs to be of a long enough duration and intensity in order to have both immediate and lasting effects. More training is particularly critical given the drop-offs in performance in the experimental group from P3 to P4 as the teachers and pupils transition to English. Lessons should be learned from the large gains realized when the pupils are learning in a familiar language from P2 to P3. It is therefore recommended that a program be developed to provide ongoing training and support of teachers and Head Teachers implementing the new curriculum,

with a special focus on teachers who have not as yet been able to participate in training. Training should also reinforce teachers' understanding and practice of continuous assessment.

Materials

The positive effects of instructional materials and of books at home, and in light of the findings that types and quantities of materials in libraries were not necessarily correlated with outcomes, more strategic procurement and use of materials would likely benefit teachers and pupils alike. It is therefore recommended that the MoES looks seriously at investments in materials and books, and that a program be developed to assist teachers and Head Teachers with strategies for effective use of materials in their libraries. Materials such as teachers' guides and translations of the curriculum should also be provided to teachers to facilitate use of the new curriculum, as well as the critical transition from local languages to English. In fact, the need for additional training and materials were the biggest concerns expressed by teachers and Head Teachers in the MLA from the previous two years.

Disparities

Encouragingly, the disparities in gender in the two subject areas at the different grade levels were very small and most often not statistically significant. However, there were differences when disaggregating by factors such as region, pupils' age at a grade level, repetition, existence of books at home, and the literacy of the mother. The MoES can examine these issues further and perhaps figure out ways of addressing the disparities, in the same way that they have had long-term programs to promote gender equity in schools. There are other factors, such as pupil dropout, that would be worth investigating as well, particularly if there are correlations with language of instruction.

Language Issues

It would be useful to conduct additional studies of the reasons behind the large gains in achievement when pupils are learning in local languages, and the difficulties in their transition from local language to English. Another issue to examine is the existence of multiple local languages in schools and the kinds of support teachers, Head Teachers and parents need to accommodate local language instruction in these contexts.

Additional Data

It is worth reiterating the need for follow-up studies of the pupils' achievement in language and maths as they progress from P4 to P5 and P6. Clearly, these data would provide a wealth of information that would lead to improved implementation of the new curriculum and additional strategies that can help pupils with the transition from local language to English as the medium of instruction, in both language and maths.

The implementation of these recommendations – in the broad areas of training, materials, disparities, language, and data – has the potential to substantially improve the implementation of the new curriculum as Uganda moves forward in the continued improvement of its education system.

Annex 1: Sampling and Testing

Sample

The selection of schools and pupils for this assessment was based on a *stratified 2-stage cluster sampling design*. Two-stage cluster sampling involves two levels of selection – in this case, the identification of schools called clusters, then pupils in those schools. Schools were selected based on geographic representation, namely region and district: the Eastern, Northern, Western and South Central (called Central in this report) regions were selected in order to represent the largest populations and language groups, and to represent the geographic diversity of the country. Within each region, one urban and one rural district were selected using purposive sampling in order to represent those two settings, the most remote districts being excluded due to time constraints. Within each district, the selection of schools was made according to language criteria, school criteria, and required sample sizes (see following sections). Once these criteria were established, government (public) schools were selected randomly within the categories specified below. A subset of private schools was also selected to serve as control schools; selection was achieved through convenience sampling. Finally, once the administrators were in the schools, they selected pupils randomly.

Language Considerations

As stated above, the purpose of this assessment is to determine the extent to which pupil learning has increased with the new curriculum. This requires a comparison of pupil performance using the old curriculum in English and the new one which, among other things (e.g., thematic instruction, pupil-centered methodology, etc.), calls for instruction to be conducted in local languages in P1-P3. This distinction required a definition of experimental schools as ones adopting the new curriculum in local language and control schools adopting the new curriculum in English. In order to make this distinction, the assumption was made that, in most cases, private schools would change to the new curriculum but continue to use English, while the government schools would use the new curriculum with local languages – many parents opt to send their children to private schools because English is the medium of instruction.

In 2008, in 2009 and in 2010 it was found that not all private schools continued to use English as the medium of instruction; nor did all government schools switch to local language instruction. Thus, the distinction between control and experimental schools changed from private vs. public to English medium vs. local language medium. This shift required a re-categorization of data from the baseline MLA 2007 in order to be able to compare schools according to these new definitions. The Ugandan technical team, consisting of UNITY staff and members of NCDC, UNEB and the MoES, took the decision to include 6 of the most commonly-spoken languages for the experimental group: Acoli (North), Ateso (East), Lango (North), Luganda (Central), Rukiga (West) and Runyankole (West).

School Attributes

Once the languages were selected, the question became which schools in each language area to include. Since the MLA in 2010 for P4 was a replication of the MLA in 2009 for P4 (using equivalent tests) the schools used in the P4 2010 MLA were the same as those retained for the P4 2009 MLA. We should keep in mind that, in each district, schools had been selected by:

- Location: Urban, peri-urban and rural schools,
- Size: Large and small schools,
- Ownership: Government and private schools,
- Distance: Larger and smaller distances from the district center,
- Boarding type: Some of the day schools, some partly boarding and some full boarding, and
- Gender: Co-educational, boys only and girls only.

Once these parameters were established and exclusions were made, remaining schools in the data base were provisionally selected on a random basis by Ministry and UNITY staff.

Sample Size

Once a list of eligible schools was generated, the final question concerned sample size: the number of schools and the number of pupils in each school was to be determined in order to minimize sampling error. The main criterion for school and pupil selection in the sample was to ensure that results could be reported with a 95% confidence interval and a 5% margin of error – standards typically used for

pupil achievement testing. The goal was to be able to generalize the findings of the MLA to the entire population of pupils in each area in which the MLA was conducted. A total of 13 experimental and 7 control schools from each district with a sub-sample of 20 pupils per school was the minimum necessary to obtain an acceptable margin of error. Accordingly, 2,325 P2 pupils from 117 schools participated in the 2007 MLA; 2,294 P3 pupils from the same[5] 117 schools and 3,776 P2 pupils from 146 schools[6] (including the 117 P3 schools) participated in the 2008 MLA.[7] The 2009 MLA used roughly the same numbers of pupils and the same schools as the 2008 MLA: 2,239 P4 pupils in 115 schools and 3,833 P3 pupils in 146 schools. For the 2010 MLA a total of 2,276 P4 pupils in 116 schools[8] participated in the exercise. The total numbers of schools and P4 pupils in the 2010 MLA were as follows (Table 28):

Table 28: MLA 2010: Number of schools and P4 pupils

	Schools			Pupils		
Region	**Experimental**	**Control**	**Total**	**Experimental**	**Control**	**Total**
Central	21	9	30	419	180	599
East	21	8	29	420	132	552
North	20	9	29	399	181	580
West	21	7	28	405	140	545
Total	**83**	**33**	**116**	**1,643**	**633**	**2,276**

Test Development

Since the objective of the 2010 MLA test development was to construct an equivalent version of the P4 test, the same P4 2009 specifications table was used in the 2010 MLA. Then, two item writing workshops were conducted simultaneously for P4, one for the language test and one for the maths test. The coordinator of the workshops was Dickson Turyareeba, a Research and Policy Analyst with the UNITY project, in partnership with Baale Remegious, a Curriculum Specialist with the NCDC. Each workshop was attended by five people: three primary level teachers, a language expert or a maths expert from the NCDC and the UNITY coordinator. Teams wrote and selected items and ordered them, and graphics and illustrations were developed. In order to be able to equate (and therefore compare) P3 2009 to P4

[5] Mengo school replaced Ryamihanda school in 2008 sample.
[6] A few schools from the 6 districts were added to the sample to insure the representativity of the control group.
[7] The 146 schools selected for the P2 test in 2008 includes the 117 P3 schools and some 29 "new" schools chosen to ensure a minimum number of English pupils for the control group.
[8] Grace Christian Academy school in the Gulu district (North) was closed in 2010.

2010 outcomes and also to equate P4 2009 and P4 2010, a subset of items common to each test, called *anchor items*, was selected from the P3 2009 tests to be included in the P4 2009 and P4 2010 tests (equating is discussed in more detail below). All items were then analyzed by the international consultants in Uganda and in North America for their pedagogic and psychometric properties, and then organized into two versions for pilot testing. Also developed were teachers' and Head Teachers' interview instruments, test administrators' guides, and guides for the training of administrators.

Eight administrators were trained in the use of all materials to administer 2 versions of each test on a pilot basis in the four target regions. Data were subsequently entered and analyzed (see the section on Data Quality below) and recommendations were made by test administrators and project advisors. The best items were selected for the actual or *operational tests* which consisted of one language test (12 groups of items, called *testlets*) and one maths test (18 testlets). (See also the Technical Report for more discussion of the results of the pilot.)

Test Administration

For the operational test, 80 Coordinating Center Tutors (CCTs) were recruited to serve as test administrators – 10 for each District by team of 2. Sixteen people, 2 by District, worked as supervisors. These supervisors were responsible for training the CCTs in test administration, distributing all testing and administration materials, monitoring test administration, and collecting administration reports. In each of the schools, pupils who had taken the test the previous year were asked to sit for the test this year (for the panel portion of the analysis). If all 20 pupils could not be found, additional pupils were chosen randomly by test administrators according to procedures detailed in the administrators' guides. Once pupils were selected, pupils who were *not* selected were asked to join pupils in other classrooms. The administrator distributed the English booklets to the pupils and instructed them as to the rules of test taking – e.g., no verbal responses, no looking at other pupils' answers, etc. The administrator then led the pupils through the language test item by item, followed by a 15 minute break, then continued with the maths test. Following the administration of pupil tests, the CTTs interviewed the P4 teachers whose pupils had taken the test, as well as the Head Teacher of that school.

Scoring and Data Entry

After administering the tests, the CCTs returned to their regional centers to submit the booklets and administration reports according to procedures outlined in their administrator's guides. The technical team members then took the test booklets back to Kampala for sorting, tracking each booklet with its own code. Tests were then scored by project staff and Ministry officials working in groups using a common scoring sheet. Next, data were entered by project staff and contractors using Excel templates developed by the international consultants. Data entry quality control was assured by selecting at random 5 test booklets per district and checking the entered data against the original. Finally, data sets were sent in electronic format to the international consultants in the US and Canada for cleaning, analysis, and report compilation.

Data Analysis

Quantitative data analysis consisted of three steps: verification of item quality and test reliability (described in Data quality below), basic descriptive analyses and more advanced procedures, including T-tests, Levene's homogeneity of variance test, analysis of variance (ANOVA), Pearson correlations, post-hoc procedures (Tukey's b, Dunnett's C) and equating procedure. Quantitative analyses were conducted using SPSS, MULTILOG, and GAUSS-IRT software. It should be noted that a significance level of .01 was used instead of the standard .05 for t-tests and correlations; this is common practice when running multiple analyses with the same data sets.

Qualitative data analysis consisted of organizing responses given by teachers and Head Teachers into categories, then tallying their responses to identify the most frequent responses to interview questions. All qualitative analyses were conducted in Excel and Word.

Data Quality

In any test of pupil achievement, two major sources of error can compromise the quality of the data:

- sampling errors resulting from the sampling design,

- measurement error due to the lack of reliability of the tests or insufficient item discrimination, and

This section describes measures taken in these two categories in order to assess item and test quality.

Sampling Error

Sampling error is a measure of the error caused by observing a sample instead of a whole population. The larger the sampling error, the less faith one should have that a study's reported results are close to the "true" figures - that is, the figures for the whole population.

In the 2010 MLA, the size of the population and the sampling design yielded the following statistics for P4 (see Part 2: "Other results," Section 1 of the Technical Report for more details):

- for the language test, the 95% margin of error of the mean was 0.361
- for the maths test, the 95% margin of error of the mean was 0.302.

Overall, these margins of error were found to be relatively small, showing very good precision for the estimated means of the two tests.

Measurement Error

P4 test reliability and item characteristics were measured using three classical indices: Cronbach alpha, item difficulty, and item-to-test correlations. A fourth index was also used: item characteristic curves, based on item response modeling. All procedures were conducted with SPSS and MULTILOG software. Summaries of these tests appear below; additional statistical information is provided in the Technical Report. [9]

It is important to note that test items were not analyzed independently, but in group called "testlets" in which pupils followed the same instructions to answer all items in that group. For example, a task might ask pupils to draw a line connecting pictures to words. The test administrator would start by giving an example, and the pupils would write the answer to that example in their test booklets. The pupils

[9] See also Bertrand & Blais (2004) for fuller descriptions of these procedures.

would then answer the remaining items in that group – usually 2 to 4 – following the same instructions. The advantage of this approach is that it reduces the number of different types of instructions pupils must follow in order to complete each item – a strategy often used in contexts where pupils are unfamiliar with testing procedures. While this is an effective format for standardized tests, items cannot be analyzed separately; in a sense, they are the same item with different parts, and are thus considered "locally dependent." [10] To address this problem, items were grouped and analyzed as "testlets" - a technique described in Thissen & Wainer[11] (2001). A description of the analyses used to assess the psychometrical properties of the items and the tests follows.

Cronbach's Alpha

The first analysis of test reliability is Cronbach's alpha coefficient, which concerns how strongly items are correlated with one another. The more correlated the items are, the greater the reliability of the test – that is, the more the items are seen to be measuring the same general construct (e.g. math ability). A Cronbach's alpha coefficient of 0.7 or higher (the maximum possible is 1) is generally considered acceptable in pupil achievement testing. As can be seen below, both P4 language (12 testlets) and maths (18 testlets) tests obtained extremely high Cronbach's alpha measures, indicating a very high level of internal consistency (Table 29).

Table 29: Cronbach's alpha values, P4

Class	Language	Maths
P4	.944	.911

Testlet Difficulty Index

The next analysis provides a "testlet difficulty index," or the mean score of all pupils for each testlet. The higher the value, the easier the test was for the pupils. For the P4 language tests (Table 30), some language testlets were found to be very difficult (e.g., testlets 5 and 11) while others were rather easy (e.g., testlets 2 and 4) for the P4 pupils in 2010. The same was true in maths, where, for example testlet 5 was difficult while testlet 8 was easy (Table 31).

[10] A procedure called Yen's Q3 is used to verify the degree of local dependency.
[11] Thissen, D., & Wainer, H. (Eds). (2001). *Test scoring*. New Jersey: Lawrence Erlbaum Associates.

Table 30: P4 testlet means for language

Testlet	Mean	Maximum score	Std. Deviation	N
1	1.64	5	1.53	2276
2	3.16	6	1.81	2276
3	2.19	5	1.47	2276
4	3.88	6	1.35	2276
5	1.02	4	1.20	2276
6	1.97	5	1.71	2276
7	1.25	4	1.32	2276
8	1.58	5	1.52	2276
9	2.59	5	1.24	2276
10	1.82	5	1.47	2276
11	1.14	4	1.23	2276
12	2.18	5	1.74	2276

Table 31: P4 testlet means for maths

Testlet	Mean	Maximum score	Std. Deviation	N
1	1.07	3	.92	2276
2	1.36	4	1.05	2276
3	1.63	3	.59	2276
4	1.29	3	.93	2276
5	.55	3	.75	2276
6	.81	3	.82	2276
7	1.33	4	1.09	2276
8	1.74	4	1.00	2276
9	.72	2	.45	2276
10	1.05	3	.79	2276
11	1.37	4	1.36	2276
12	1.16	4	1.32	2276
13	1.04	4	1.15	2276
14	1.62	5	1.57	2276
15	1.18	5	.96	2276
16	.98	4	1.21	2276
17	.68	6	1.17	2276
18	1.36	5	1.28	2276

Item-Total Correlation

The third of these classical indices is the "item-total correlation" – a measure of how well an item discriminates between low and high achievers. An item is said to have good discrimination when pupils with high exams scores get an item correct, and pupils with low exam scores get the item incorrect. The item-total correlation is a measure of this relationship – i.e., how well each pupil performed on each item relative to his/her total exam score. In other words, correlations of closer to 1 (the maximum) have greater discrimination.

Most of the item-total correlations were found to be very high on both the language and maths tests, meaning that they provide a high level of discrimination between pupils of different abilities. Tables 32 and 33 present the item-total correlations (see Technical Report for more complete tables). Note also that the strong item-total correlations for most of the items in these tables are consistent with the high value of Cronbach's alpha coefficient presented above.

Table 32: Item-total correlations for P4 language test

Testlet	Corrected Item-Total Correlation
1	.780
2	.734
3	.68
4	.64
5	.76
6	.85
7	.84
8	.75
9	.59
10	.69
11	.82
12	.82

Table 33: Item-total correlations for P4 maths test

Testlet	Corrected Item-Total Correlation
1	.64
2	.66
3	.43
4	.53
5	.59
6	.61
7	.67
8	.52
9	.44
10	.44
11	.75
12	.58
13	.76
14	.66
15	.49
16	.45
17	.56
18	.68

Threats to Validity

Finally, four threats to validity must be considered in their interpretation:

1. **Multiple measures:** The 2010 MLA represents the fourth and final year of measure of a large phenomenon – i.e., the effect of Uganda's national curriculum reform on learning in its schools. To obtain a reliable measure of such a phenomenon, multiple measures (e.g., a minimum of 3-5 years) are needed in order to make valid claims. Indeed 5 measures would be better than 4 measures and 4 measures would be better than only 3.
2. **Hawthorne effect**: Any new intervention such as this reform is likely to create a "spike" or change in behavior in the short term, due to initial excitement,

changed expectations, or other factors. This phenomenon, called the Hawthorne Effect, is less likely to be a factor in this fourth year of testing than it might have been in Years 1, 2 and 3.

3. **Ceiling effect:** Members of a subgroup who are near the "ceiling," or upper range of a measurement scale (e.g., pupils with higher scores) are less likely to make the same size gains as those who are closer to the bottom. This phenomenon, called the Ceiling Effect, might account for the smaller differences noted in control schools than in the experimental schools. Again, multiple measures will reveal the extent to which pupils' scores might be attributable to the new curriculum or to other factors.
4. **Problems with test quality or administration:** Every measure has been taken, and described in this report, to ensure the highest quality possible of test construction, test administration, scoring, and data entry. In the 2010 MLA, some missing data values were observed, especially return rates of interview instruments in certain districts and low percentages of pupils followed in the panel design. These data gaps can impact the interpretation of data and generalizability of findings; these concerns are noted in the interpretation of this report.

Annex 2: Panel Analysis

Comparing the Total Sample Scores to the Subsample Scores

Table 34 shows that the subsample of 673 pupils out of the sample of the 3776 pupils in 2008 P2 Cohort 2 is proportionally equal to the subsample of 420 pupils found in 2007 P2 Cohort 1 where the sample counted 2325 pupils. This very low proportion (.18) of pupils in the panel constitutes a major drawback for the analysis and the interpretation of the panel data. Besides, as can be seen in Table 35, the subsample scores in language and maths are found to be much higher than the sample scores.

Table 34: Mean language and maths scores for the Cohort 2 sample (3776 pupils) and the subsample (673 pupils) in P2 (year 2008)

Results of the subsample

	N	Minimum	Maximum	Mean	Std. Deviation
LANGUAGE	673	3.00	38.00	27.01	9.50
MATHS	673	2.00	37.00	21.77	7.99
Valid N (listwise)	673				

Results of the global sample

	N	Minimum	Maximum	Mean	Std. Deviation
LANGUAGE	3776	.00	38.00	22.12	10.64
MATHS	3776	.00	37.00	18.39	8.53
Valid N (listwise)	3776				

Repeated Measures Analyses

According to what has been done for the analyses in the precedent parts of this report, where P3 2009 pupils were compared to P2 2008 pupils, all «true» achievement scores in P3 and P4 were linked to the baseline P2 scale. Then a repeated measures analysis (also known as «split-plot design») was performed with the linked scores: the within-subject factor being the three-year achievement test (either in language or maths) and the between-subject factor being the control-experimental grouping.

Table 35 and Table 36 show the true percent scores in language and maths for the 673 pupils while they were in P2 (2008), P3 (2009) and P4 (2010). The more obvious result stemming from those 2 tables is that the pupils in the control schools, whether in P2 2007, P3 2008 or P4 2009, got much higher language and maths mean true scores than the pupils in the experimental schools. It can also be observed that, except for the language test between P2 2008 to P3 2009, there is a general increase in mean scores from P2 2007 to P3 2008 and then to P4 2009 for the pupils from the control schools while pupils from the experimental schools increased their mean scores between P2 2008 and P3 2009 but their mean scores decreased dramatically between P3 2009 and P4 2010.

Table 35: Mean language true scores for the pupils of the control and the experimental schools in P2 2008, P3 2009 and P4 2010

		Mean	Std. Deviation	N
True language P2 2008	Control	95.25	4.95	219
	Experimental	77.33	21.80	454
	Total	83.16	19.97	673
True language P3 2009	Control	94.77	5.18	219
	Experimental	86.17	15.95	454
	Total	88.96	14.02	673
True language P4 2010	Control	98.32	4.50	219
	Experimental	71.53	26.47	454
	Total	80.25	25.23	673

Table 36: Mean maths true scores for the pupils of the control and the experimental schools in P2 2008, P3 2009 and P4 2010

		Mean	Std. Deviation	N
True maths P2 2008	Control	74.82	12.10	219
	Experimental	61.99	17.03	454
	Total	66.16	16.71	673
True maths P3 2009	Control	86.33	9.50	219
	Experimental	77.99	13.32	454
	Total	80.70	12.81	673
True maths P4 2010	Control	91.23	7.8321	219
	Experimental	69.52	17.29	454
	Total	76.58	18.03	673

Table 37 shows part of the result of the repeated measures analysis for the language test: the within-subject factor (3-year measure of achievement: P2 2008, P3 2009, P4 2010) is statistically significant[12] (F=33.421; p<.001) but also the interaction «factor*Control» between this factor and the between-subject variable (control-experimental) is statistically significant (F=82.844; p<.001): in fact, as already observed in Figure 8 (see the section untitled « Measuring the progress of Cohort 2, P2 to P4, 2008 to 2010»), the control and the experimental lines are not parallel for the language test, a graphical manifestation of this interaction.

Table 37: Repeated measures analysis of the control and the experimental schools (between-subject factor) in P2 2008, P3 2009 and P4 2010 (within-subject factor) for the language scores

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Within-subject factor	Lower-bound	9855.898	1.000	9855.898	33.421	.000
factor * Control	Lower-bound	24431.183	1.000	24431.183	82.844	.000
Error(factor1)	Lower-bound	197881.321	671.000	294.905		

This statistically significant interaction implies that «simple main effect» analyses must be implemented to compare the true mean language scores between the pupils from the control schools and the pupils from the experimental schools for each of the 3-year within-subject factor, or P2 2008, P3 2009 and P4 2010. Table 38 presents the results of this simple main effect analysis. The tests are shown to be statistically significant for P2 2008 (F=144.249; p<.001), for P3 2009 (F=60.56; p<.001) and for P4 2009 (F=221.095; p<.001). These results agree with the above observations (see Table 33) while comparing the mean language scores of the control and the experimental schools: 95.25 (control) vs. 77.33 (exp) for P2 2008; 94.77 (control) vs. 86.17 (exp) for P3 2009; 98.32 vs. 71.53 for P4 2010.

Table 38: Simple main effect analyses comparing the mean language true scores for the pupils of the control and the experimental schools in each of the three group-years P2 2008, P3 2009 and P4 2010

```
EFFECT .. Control_Experimental

Univariate F-tests with (1,418) D. F.

Variable          Hypoth. SS     Error SS       Hypoth. MS     Error MS      F       Sig

Language P2 2008   47422.395    220592.297      47422.395       328.751    144.249   .000
Language P3 2009   10931.410    121117.733      10931.410       180.503     60.560   .000
Language P4 2010   106002.158   321704.388     106002.158       479.440    221.095   .000
```

[12] Mauchly's test of sphericity was found to be statistically significant (25.566; p<.001). Lower bound degrees of freedom correction was used.

The simple main effect procedure can also be used to contrast any two-year results (P2 2008 vs. P3 2009, P3 2009 vs. P4 2010, P2 2008 vs. P4 2010) for each of the control group and the experimental group: Table 39 presents these simple effects results for the language test. These results confirm what has been seen in Figure 8 since only the P2 2008 vs. P3 2009 contrast for the control group was found not statistically significant. All other two-by-two comparisons must be considered statistically significant.

Table 39: Simple main effect analyses comparing the mean language true scores for the pupils between P2 2008 and P3 2009, between P3 2009 and P4 2010 and between P2 2008 and P4 2010 for each of the control group and the experimental group

Control	*MS*	*F*	*Signif.*
P2 2008 vs P3 2009	25.65	.21	.648
P3 2009 vs P4 2010	1378.38	8.01	.005
P2 2008 vs P4 2010	1027.95	6.96	.009
Experimental			
P2 2008 vs P3 2009	17700.53	144.37	.000
P3 2009 vs P4 2010	48629.98	282.70	.000
P2 2008 vs P4 2010	7652.53	51.80	.000

Table 40 shows part of the result of the repeated measures analysis for the maths test: the within-subject factor (3-year measure of achievement: P2 2008, P3 2009, P4 2010) is statistically significant[13] (F=358.480; p<.001) but the interaction «factor*Control» between this factor and the between-subject variable (control-experimental) is also statistically significant (F=74.128; p<.001):): in fact, as already observed, the control and the experimental lines are not parallel for the maths test, a graphical manifestation of this interaction.

Table 40: Repeated measures analysis of the control and the experimental schools (between-subject factor) in P2 2008, P3 2009 and P4 2010 (within-subject factor) for the maths scores

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Within-subject factor	Lower-bound	66121.139	1.000	66121.139	358.480	.000
factor * Control	Lower-bound	13672.804	1.000	13672.804	74.128	.000
Error(factor1)	Lower-bound	123764.865	671.000	184.448		

[13] Mauchly's test of sphericity was found to be statistically significant (15.073; p<.001). Lower bound degrees of freedom correction was used.

This statistically significant interaction implies that «simple main effect» analyses must be implemented to compare the true mean maths scores between the pupils from the control schools and the pupils from the experimental schools for each of the 3-year within-subject factor, that is P2 2008, P3 2009 and P4 2010. Table 41 presents the results of this simple main effect analysis. The tests are shown to be statistically significant for P2 2008 (F=99.891; p<.001), for P3 2009 (F=68.977; p<.001) and for P4 2009 (F=313.922; p<.001). These results are actually in accord with the above observations while comparing the mean maths scores of the control and the experimental schools: all maths mean score are much higher for the pupils of control schools than for the pupils of experimental schools for each of the three group-year P2 2008, P3 2009 and P4 2010.

Table 41: Simple main effect analyses comparing the mean maths true scores for the pupils of the control and the experimental schools in each of the three group-year P2 2008, P3 2009 and P4 2010

EFFECT .. Control_Experimental

Univariate F-tests with (1,418) D. F.

Variable	Hypoth. SS	Error SS	Hypoth. MS	Error MS	F	Sig
Maths P2 2008	24312.381	163313.639	24312.381	243.388	99.891	.000
Maths P3 2009	10280.719	100008.249	10280.719	149.043	68.977	.000
Maths P4 2010	69621.600	148813.906	69621.600	221.779	313.922	.000

As was done for the language test, the simple main effect procedure can also be used to contrast any two-year results (P2 2008 vs. P3 2009, P3 2009 vs. P4 2010, P2 2008 vs. P4 2010) for each of the control group and the experimental group for the maths test: Table 42 presents these simple effects results. These results confirm what has been seen earlier: all these two-by-two comparisons are statistically significant.

Table 42: Simple main effect analyses comparing the mean maths true scores for the pupils between P2 2008 and P3 2009, between P3 2009 and P4 2010 and between P2 2008 and P4 2010 for each of the control group and the experimental group

Control	*MS*	*F*	*Signif.*
P2 2008 vs. P3 2009	14510.14	148.62	.000
P3 2009 vs. P4 2010	2628.61	33.46	.000
P2 2008 vs. P4 2010	29490.49	293.52	.000
Experimental			
P2 2008 vs. P3 2009	58096.00	595.04	.000
P3 2009 vs. P4 2010	16273.50	207.13	.000
P2 2008 vs. P4 2010	12873.97	128.13	.000

References

Benson, C. (2004). *The importance of mother tongue-based schooling for educational quality*. Paris: UNESCO. (For the *Education for all global monitoring report, 2005*).

Bertrand, R. and Valiquette, C. (1986) *Pratique de l'analyse statistique des données*. Québec: Presses de l'Université du Québec.

Bertrand, R. and Blais, J-G. (2004). *Modèles de mesure : l'apport de la théorie des réponses aux items*. Québec: Presses de l'Université du Québec.

Cochran, W. G. (1977). *Sampling techniques*. New York: John Wiley.

Glass, G. V., & Hopkins, K. D. (1996) *Statistical methods in education and psychology*. Boston: Allyn & Bacon.

Lohr, S. L. (1999) *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press.

Lynd, M. and Bertrand, M. (2008) *Measuring Learning Achievement 2007: Language and maths in P2, UNITY Project, Uganda.* San Francisco, CA: School-to-School International and Creative Associates International, Inc.

Lynd, M. and Bertrand, M. (2009) *Measuring Learning Achievement 2008: Language and maths in P2 and P3, UNITY Project, Uganda.* San Francisco, CA: School-to-School International and Creative Associates International, Inc.

Lynd, M. and Bertrand, M. (2010) *Measuring Learning Achievement 2009: Language and maths in P3 and P4, UNITY Project, Uganda.* San Francisco, CA: School-to-School International and Creative Associates International, Inc.

Raju, N.S, van der Linden, W.J., & Fleer, P.F. (1995) IRT-based internal measure of differential functioning of items and tests. *Applied Psychological Measurement*, 19, 4, 353-368.

Rosier, M. (1982). *Sampling and administration manual*. Second IEA Science Study: International Association for the Evaluation of Educational Achievement (IEA).

Samejima, F. (1997). Graded response model. Dans : W.J. Van der Linden et R.K. Hambleton, *Handbook of modern item response theory* . New York : Springer.

Thissen, D. (1991). *MULTILOG user's guide: Multiple categorical item analysis and test scoring using item response theory*. Chigago: Scientific Software International.

Thissen, D. and Wainer. H. (eds). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H., Bradlow, E.T., and Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.