THE REPUBLIC OF UGANDA

Ministry of Education and Sports

Final

Core Report

Measuring Learning Achievement 2009:

Language and maths in P3 and P4

UNITY Project, Uganda

With support from USAID

Report prepared by School-to-School International
for Creative Associates International, Inc.

April 2010

Contents

Acknowledgements

The MLA exercise described in this report was conducted by a number of people without whom the quality of the tests and the items, as well as the management of the MLA exercise in general, would not have been possible. In particular, the authors would like to extend their sincerest appreciation to Jon Silverstone, Sandhya Bandrinath, and Roseline Fodouop Tekeu of Creative Associates' home office for providing management continuity, management and moral support; Renuka Pillay, Chief of Party of the UNITY Project, for the vigilant attention she has paid to the effective management of this effort since its inception in 2007; to NCDC for its technical oversight, supervision, and assistance; and to the people listed below, and to the teachers who gave generously of their time to again produce what resulted in a set of tests of extremely high quality.

We would again like to thank the UNITY Project and Creative Associates for organizing a conference presentation at the annual meeting of the Comparative International Education Society (CIES) in March 2009 held in Charleston, South Carolina to which Martin Opolot, M&E Officer, UNITY Project and Albert Byamugisha, Head of Planning, MoES. Our presentation on "Assessment Across Languages," at which Martin, Albert, Richard Bertrand and Mark Lynd presented was well attended and extremely well-received, clearly addressing a topic of growing concern to people around the world.

And again in 2009, a special note of thanks to Martin Opolot and Dickson Turyareeba who continued to be available at all times to ensure the management and monitoring of each step of a large, complex, and time-bound process.

The staff and consultants of School-to-School International
March 2010

Those who assisted with the MLA 2009 exercise

Including P4 item writers and P3 translators

(in alphabetical order)

Title	Institution
Teacher	Katojo P/S, Mbarara
Teacher	Kumi Girls P/S, Kumi
Language Expert	NCDC
Maths Expert	NCDC
Principal Education Officer	UNEB
Research and policy analyst	UNITY Project
Teacher	St. Theresa's P/S, Entebbe
Retired Tutor	
Retired Tutor	
Data Assistant	UNITY Project
Retired Head Teacher	
Statistician	Education Planning Department
Teacher	Kawongo P/S, Mukono
Teacher	Nakasero P/S, Kampala District
Teacher	Kankobe P/S, Mpigi
Teacher	Kidetok P/S, Soroti
Tutor	Kabale-Bukinda PrimaryTeacher's College
Program Manager	Creative Associates International, Inc.
Data assistant	UNITY Project
Education Planner	Education Planning Department
Teacher	SOS Hermann Gmeiner, Kakiri, Wakiso
Senior Education Officer	Teacher Education Department
Teacher	Buloba C/U, Kampala
Monitoring and Evaluation Officer	UNITY Project
Teacher	Namirembe Infants P/S, Kampala District
Teacher	Bat Valley P/S, Kampala
Teacher	Lira P/S, Lira
Retired Tutor	
Retired Tutor	
Program Officer	UNITY Project
Data Assistant	UNITY Project
Chief of Party	UNITY Project
Retired Head Teacher	
Teacher	Opit P/S, Gulu
Program Officer	CAII
Tutor	Gulu Teacher's College

Acronyms

CAII	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
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

Monitoring Learning Achievement (MLA) is an assessment of pupil achievement conducted for the Uganda Initiative for TDMS and PIASCY (UNITY) Project funded by USAID.[1] The purpose of the MLA is to determine the extent to which pupil learning has increased with the implementation of the new primary school curriculum launched in 2007. The first MLA was conducted in 2007, when the UNITY Project, in collaboration with UNEB, NCDC, and the MoES, tested pupils in 4 regions, 2 districts per region, to create a baseline in language and maths at the P2 level - the last year in which the old English medium curriculum was used at that level. Subsequent tests were conducted in the same schools and additional ones in 2008 and 2009. Specifically, the MLA 2009 measured the following dimensions of pupils' achievement:

i) P4 achievement in literacy and numeracy in English – this measure was to serve as a baseline.

ii) P3 achievement in literacy and numeracy in local language, as a follow up to the baseline taken in 2008.

iii) Progress in literacy and numeracy achievement in English for pupils from 2007 (P2) to 2009 (P4) – i.e., a panel design following "cohort 1" during the last years of the use of English as the medium of instruction.

iv) Progress in literacy and numeracy achievement in local language for from 2008 (P2) to 2009 (P3) - i.e., a panel design following "cohort 2," the first group to study under the new curriculum in local language.

[1] Creative Associates International, Inc. is the prime contractor for the UNITY Project. Creative engaged the services of School-to-School International to coordinate out the MLA in 2007, 2008 and 2009.

This report summarizes the results of the 2009 MLA.

Study implementation

In May 2009, P3 tests (developed in English for the 2008 MLA) were translated into 6 local languages, a new P4 test was developed in language and maths (to be administered in English), interview instruments for P4 teachers were developed, and Head Teacher interview instruments revised. The tests were piloted, then revised and administered in September/October 2009. In all, 3,833 P3 pupils in 146 schools and 2,239 P4 pupils in 115 schools took the language and maths tests. Local teams in Uganda scored the tests and entered test and interview data to be sent to North America for analysis by School-to-School International (STS). Data were then cleaned and additional pupil panel information was registered. MLA data analysis and report writing were subsequently conducted by STS.

Summary of findings

P3 test in local languages

For the most part, MLA 2009 results mirrored those of 2007 and 2008 when broken down by geography, school type, language, gender, and other pupil characteristics. The following is a summary of those results.

- ***Geography***: Pupils in the Western and Central regions again scored significantly higher than in the East and North in both language and maths. And again, pupils in the Mbarara District scored higher than all other districts in language and maths, and pupils in Kumi and Gulu scored significantly lower in language and Lira scored the lowest in maths.
- ***Control vs. experimental***: Pupils taking test in English (control) scored significantly higher than those taking it in local language – an expected outcome since control schools are mostly private schools which historically score better.
- ***Language***: And again, pupils taking the test in Runyankole scored higher than pupils in all other language groups. Pupils taking the test in Acoli scored the lowest in language, pupils taking the test in Lango scored the lowest in maths.

- ***Gender:*** As before, no significant differences were found between the performance of girls and boys in language or maths, either as a whole or when disaggregated by geography, age, language, repeater status, books in the home, or a mother who reads.
- ***Other pupil characteristics***: The youngest pupils (age 9) scored significantly better than all others in language and maths. Non-repeaters scored significantly better than repeaters in both language and maths. Pupils with books at home or a mother who reads scored significantly better in both language and maths than their peers.
- ***Strongest correlations***: Higher scores significantly correlated with the number of male teachers in a school, greater qualification of the Head Teacher, having a library in a school, and having exercise books and materials. Surprisingly, pupils whose teachers had less experience scored higher in language and maths.
- ***Perspectives on the curriculum and training***: As was in 2008, the greatest strength of the new P3 curriculum cited by teachers and Head Teachers was the use of local language, resulting in greater comprehension for pupils and ease of communication in the classroom for teachers. The biggest difficulties included translating the curriculum into local language and the lack of materials. The principle concern noted about the training was that it was too short, so not all the material could be covered.

P4 test in English

- ***Geography***: As in P3, pupils in the Western Region scored significantly higher than all other regions in both language and maths. Pupils in the North scored the lowest in both subjects. And as in P3, Mbarara scored higher than all other districts in language and maths. Gulu was also lowest in language and instead of Lira (P3), Gulu lowest in maths
- ***Control vs. experimental***: Pupils in English medium (control) schools scored significantly higher than pupils in experimental schools.
- ***Language***: Pupils who took the test in Runyankole tested significantly higher than all other language groups in both language and maths; pupils testing in Acoli scored the lowest in both language and maths.
- ***Subject***: Language scores were significantly higher than maths scores in all districts except Gulu, Kumi and Mpigi.

- ***Gender***: No significant differences were found between the performance of girls and boys in language or maths, either globally, by region or by language.
- ***Other groups***: The youngest pupils (under 10) scored significantly better than all others in language and maths. Non-repeaters scored significantly better than repeaters in both language and maths, and pupils with books at home or having a mother who reads scored significantly better in both language and maths than their counterparts.
- ***Strongest correlations***: Higher scores were significantly correlated with the number of teachers in a school, male and female combined (i.e., the more teachers in the school, the better the pupils performed), when pupils had their own rulers and exercise books, and where Head Teachers said they liked the old curriculum better.

Overall, factors such as teachers' qualifications, years of service, amount of training received for the new curriculum were not correlated with pupil performance; nor was the quantity or nature of materials in school libraries.

Comparing performance using the old and new curriculum

Key finding: As was the case with P2 pupils in 2008, our cross-sectional analysis (see MLA design, p. xxx) showed that P3 pupils in the experimental group performed significantly better in 2009 under the new curriculum than P3 pupils did in 2008 under the old, in both language *and* maths, whereas pupils in the control schools showed no statistically significant differences between 2008 and the 2009, either in language *or* in maths. This was also true within subgroups: pupils had significantly higher scores in experimental schools *and not in control schools* when disaggregated by age (younger performed better) and by repetition (non-repeaters performed better). Similarly, all boys and girls performed significantly better with the new curriculum, especially in language. Pupils in the control group did not show significant gains in these areas. Perhaps the most important finding, as was the case in 2008, is that *all* pupils in experimental schools performed significantly better with the new curriculum whether their mothers read or not, or whether they had books in the home or not, creating a type of "affirmative action effect" helping more disadvantaged children.

Similarly, our cohort analysis (P2 2008 to P3 2009) showed gains in both language and maths for both control and experimental groups, but greater gains for the experimental group, indicating a cumulative effect of the new curriculum for both groups, but especially for the experimental group.

Finally, our analysis of cohort 1 (old curriculum) from P2 2007 to P4 2009 showed that while pupils in control schools consistently scored higher than those in experimental schools, scores rose substantially from Year 1 to Year 3 for both groups in both language and maths, with the greatest rates of improvement in language for the experimental group in Year 3.

These findings show that:

1. The new curriculum also appears to be having a positive impact on pupils in control schools and pupils who have not yet benefitted from the new curriculum in experimental schools – in effect, "floating all boats."
2. If the new curriculum is floating all boats, some appear to be floating higher, as the curriculum is having a significantly greater and sustained impact on the experimental group (pupils using the new curriculum).
3. Continued support for teachers and more strategic use of materials in libraries could have a positive effect on achievement.
4. Finally, and perhaps most importantly, disadvantaged students are showing significantly higher performance levels under the new curriculum than they did under the old.

Recommendations in this report address the need to ensure continued training in the new curriculum, the provision of materials and assistance in translation of the curriculum, and review of data collection and test administration procedures to reduce the rate of missing data.

Introduction

The Uganda Initiative for TDMS and PIASCY (UNITY) Project is a USAID-funded initiative managed by Creative Associates International, Inc. The goals of the UNITY Project (hereafter called "UNITY") include 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 have occurred as a result of project interventions, UNITY initiated a student testing effort in 2007, called Measuring Learning Achievement (MLA). This report presents the results of the third MLA test administered in September/October 2009 to P3 and P4 pupils in all four geographic regions of Uganda. This document constitutes the first part of the report, or the "Core Report." It begins by describing the design of the exercise, then presents the findings from the 2009 MLA, followed by a discussion of salient findings and recommendations for future MLA exercises and curriculum implementation issues. 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 technical aspects of each part of this report.

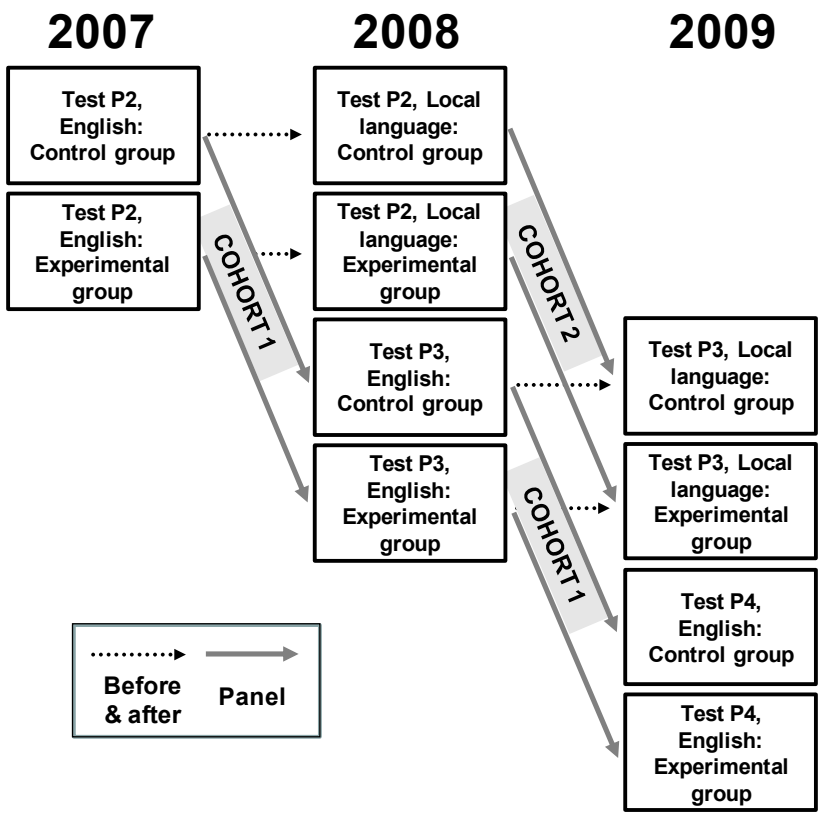
MLA design

The UNITY MLA was designed to consist of three rounds of tests to be conducted from 2007 to 2009 to demonstrate change over time between two cohorts. Cohort 1 consisted of the last group of students to move through the system using the old English-based curriculum. To capture change within this group, a baseline was conducted in 2007 in English (the last year it would be used as the medium of instruction), testing pupils' language and maths skills at the P2 level. The following

year, cohort 1 was again tested in P3 in language and maths, using English as the language of the test, and in 2009 they were tested in P4 in the same subjects, again using in English as the language of the test.

In 2008, the MLA continued to track cohort 1 some pupils being tracked individually in a "panel design." That year, cohort 2 also began participating in the MLA. Being the first group to move through the system under the new curriculum, cohort 2 pupils took the same tests as those in cohort 1, but in local language, the medium of instruction in the new curriculum. Six major local languages were used in the test from 4 regions; in schools where English was still being used as the medium of instruction (mostly private schools), students were tested as a proxy for continued use of the old curriculum. Finally, in 2009, cohort 1 was tested at the P4 level, again in English, while cohort 2 was tested in P3 in local language.[2] The MLA design from 2007 to 2009 is illustrated below:

Figure 1: MLA design, 2007 – 2010



[2] The final round is scheduled for P4 in 2010 in English, the year cohort 2 transitions from local language to English under the new curriculum.

The 2009 MLA P3 tests were administered in September/October 2009 in 6 local languages for the experimental schools and in English for the control schools; the P4 tests were administered in English only. In all, the 2009 MLA reached 3,833 P3 pupils in 143 schools and 2,239 P4 pupils in 115 schools as follows:

Table 1: MLA 2009: Number of schools and P3 pupils

	Schools			Pupils		
Region	**Experimental**	**Control**	**Total**	**Experimental**	**Control**	**Total**
Central	21	9	30	514	180	694
East	20	8	28	481	157	638
North	33	10	43	954	300	1254
West	35	7	42	1031	216	1247
Total	**109**	**34**	**143**	**2,980**	**853**	**3,833**

Table 2: MLA 2009: Number of schools and P4 pupils

	Schools			Pupils		
Region	**Experimental**	**Control**	**Total**	**Experimental**	**Control**	**Total**
Central	21	9	30	403	179	582
East	20	8	28	379	158	537
North	20	9	29	396	180	576
West	21	7	28	404	140	544
Total	**82**	**33**	**115**	**1,582**	**657**	**2,239**

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

Findings

Four types of analyses were conducted for the MLA 2009:

- A summary of P3 and P4 test results for 2009,
- A comparison of P3 results from 2008 to 2009,
- A measure of the progress of cohort 1 - the last group using the old curriculum , and
- A measure of results for cohort 2 – the first group to use the new curriculum.

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

P3 pupil performance

Geograghy. An analysis by region shows that P3 pupils in the West and Central regions had significantly higher scores in language and maths than their counterparts in the East and North. Note that the North and West regions have double the number of pupils, owing to the necessity to have a minimum number of pupils in each language group in order to be able to conduct reliable statistical analyses (see Figures 2 and 3).

Figure 2: P3 mean language scores by region

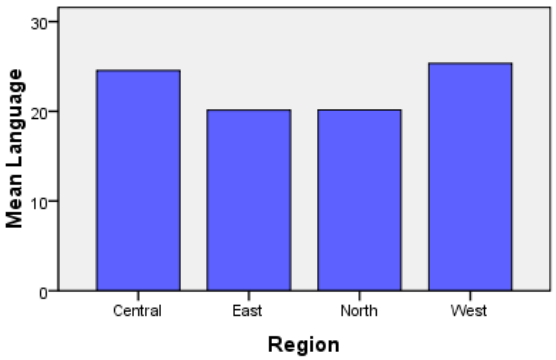
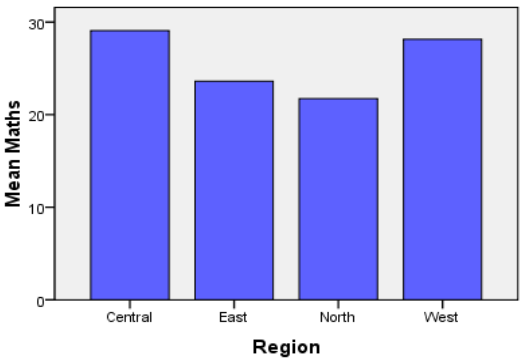


Figure 3: P3 mean maths scores by region



Region	Mean	Std. Deviation	N
Central	24.54	9.577	694
East	20.13	10.310	638
North	20.15	10.657	1254
West	25.34	9.864	1247
Total	22.63	10.446	3833

Region	Mean	Std. Deviation	N
Central	29.08	8.190	694
East	23.61	9.047	638
North	21.73	10.665	1254
West	28.15	8.816	1247
Total	25.46	9.901	3833

Amongst districts, Mbarara pupils scored significantly higher than their counterparts in all other districts, both in language and in maths (except for Mukono), with the Kumi and Gulu districts reporting the scores significantly lower than all other districts in language: in maths, Lira district got significantly lower scores than all other districts except Gulu.

Figure 4: P3 mean language scores by district

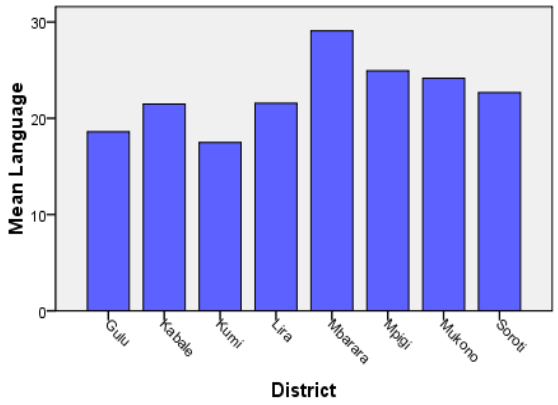
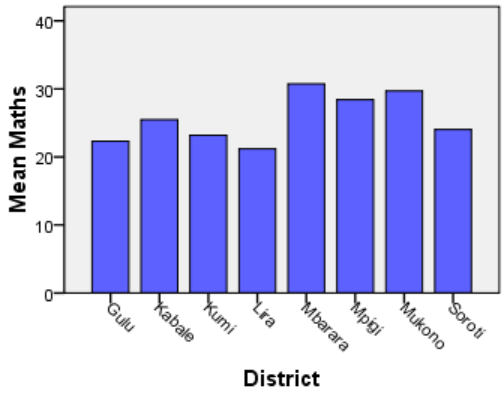


Figure 5: P3 mean maths scores by district



District	Mean	Std. Deviation	N
Gulu	18.59	10.364	594
Kabale	21.48	10.284	615
Kumi	17.49	9.513	313
Lira	21.56	10.728	660
Mbarara	29.09	7.787	632
Mpigi	24.94	8.861	340
Mukono	24.16	10.215	354
Soroti	22.67	10.423	325
Total	22.63	10.446	3833

District	Mean	Std. Deviation	N
Gulu	22.32	10.450	594
Kabale	25.49	9.684	615
Kumi	23.18	8.694	313
Lira	21.20	10.835	660
Mbarara	30.73	6.976	632
Mpigi	28.43	7.956	340
Mukono	29.70	8.372	354
Soroti	24.02	9.369	325
Total	25.46	9.901	3833

Language. Because the language of instruction in the classroom is sometimes different from the language some pupils speak at home, analysis of pupil performance was based on the language of the test booklet used by the pupils – a choice based on the recommendation of the teacher when the administrator entered the classroom. Figures 6 and 7 show that pupils taking the English test (control schools) scored significantly higher than pupils in all other local language groups in both language and maths. Except for English, pupils taking the test in Runyankole scored significantly better than all other pupils in both language and in maths (except for Luganda). Pupils taking the test in Acoli scored significantly lower than all other language groups (except for Ateso pupils) in language achievement; pupils taking the test in Lango scored significantly lower in maths than all other languages.

Figure 6: P3 mean language scores by language of the booklet

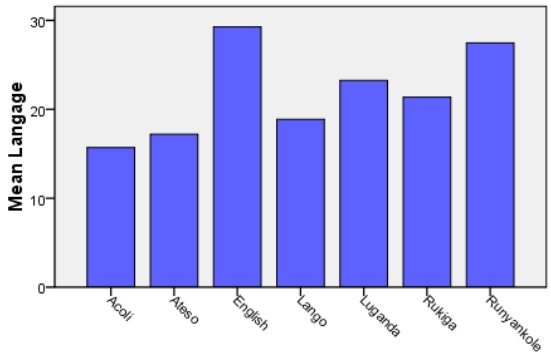
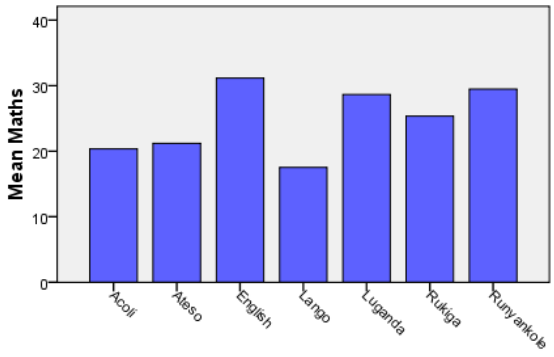


Figure 7: P3 mean maths scores by language of the booklet



Language	Mean	Std. Deviation	N
Acoli	15.69	9.256	444
Ateso	17.19	9.085	481
English	29.27	7.954	853
Lango	18.87	10.142	510
Luganda	23.25	9.977	514
Rukiga	21.36	10.493	549
Runyankole	27.46	7.925	482
Total	22.63	10.446	3833

Language	Mean	Std. Deviation	N
Acoli	20.35	10.264	444
Ateso	21.19	8.262	481
English	31.16	8.058	853
Lango	17.53	8.695	510
Luganda	28.65	8.184	514
Rukiga	25.34	9.777	549
Runyankole	29.47	6.842	482
Total	25.46	9.901	3833

Three other variables were examined in reference to pupil performance and the following differences were found:

- **Age:** The youngest pupils – those under 9 years old[3] – performed significantly better than all other pupils both in language and in maths.
- **Status as repeaters:** Non-repeating pupils (about 80% of the total) performed significantly better than repeaters in both language and maths.
- **Home environment:** In order to have a picture of pupils' home background, P3 pupils were asked to respond yes or no to two statements: "There are books in my home" and "My mother reads at home." Pupils who said yes to having books in the home (68%) and to having a mother who reads (62%) performed significantly better in both language and maths than those who said no to these questions.

[3] P3 pupils were divided into 4age groups for analysis: under 9 years of age, 9 years old, 10 years old and 11 and older.

Gender. Several analyses were conducted to determine how gender was associated with performance according to other independent variables. As Table 3 shows, 49.5% of all P2 pupils tested were boys and 50.5% girls. As a group, no significant differences were found between girls' and boys' language or math scores.

Table 3: P3 mean language and maths scores by pupil's sex

Subject	Sex	N	Mean	Std. Deviation
Language	Girl	1928	22.86	10.240
	Boy	1890	22.38	10.658
Maths	Girl	1928	25.32	9.754
	Boy	1890	25.59	10.067

A closer look at variations in girls' performance was taken by region, age, language of the booklet in which the girls were tested, repeater status, and home profiles. The following are the results of those analyses.

- **Region**: Girls and boys perform at the same level both in language and maths for each of the four administrative regions.
- **Age**: No statistically significant differences were found between girls and boys either in language or maths for each of the four age groups.
- **Language of the booklet**: Girls' and boys' performance was comparable (not statistically different) across languages.
- **Repeater, books in home, mother who reads**: No statistically significant differences in performance were found between girls and boys whether or not they were repeaters, had books in the home or had a mother who reads.

P3 correlations and perspectives

In order to understand key conditions about learning in MLA schools, teachers and Head Teachers participating in the MLA were asked about school conditions as well as their experience, qualifications, and views on the new curriculum and the training they had received in its use. This section presents correlations between school conditions, personnel attributes and student performance.

Correlations

- **Number of male vs. female teachers:** The number of male teachers in a school had a significant impact on P3 pupil achievement in both language and maths – that is, the greater the number of male teachers, the better the performance.
- **Existence of a library and materials:** Both maths and language scores were higher in schools with libraries, which reportedly existed in two-thirds of the sample. Language and maths scores were higher when teachers reported having access to dictionaries and other materials. Bizarrely, the greater the number of science books in school libraries, the better pupils' language scores. Library materials were not abundant, but teachers reported that on average, each school had between 35 and 50 books in each subject area by grade (see Table 5).

Table 4: Well-supplied teachers

Materials	P3	P4
Flash cards	68%	61%
Word cards	70%	54%
Wall charts	63%	78%
Work cards	68%	55%
Stationery	82%	85%
Dictionaries	73%	85%

Table 5: Average number of books in school libraries

Subject	P3	P4
English	43	49
Maths	38	47
Science	32	41
Social Studies	35	35

- **Pupil ownership of materials:** Language scores were higher when pupils had their own exercise books and rulers; maths scores were higher when pupils had their own rulers. Importantly, two-thirds of teachers reported that none of their pupils had pencils.
- **Experience of the P3 teacher:** Surprisingly, P3 pupils whose teachers had less experience scored higher in language and maths.
- **Highest academic qualification of the Head Teacher:** Language and maths scores of P3 pupils were higher when the Head Teacher of their school had a higher academic qualification (31% reported having a certificate of Grade V or above).

No significant correlations were found between pupils' scores and the:

- Highest academic qualification attained by the teacher (69% were Grade III or below, but this did not influence pupil performance),
- Sex of pupils' teacher in a given class (though the total number of male teachers in a school made a difference, as noted above),
- Number of boys or girls in school: Language and math scores did not vary in correlation to the proportion of boys or girls in school,
- Sex of the Head Teacher: Male and female Head Teachers' pupils scored roughly the same (75% of Head Teachers reporting were male),

- Head Teachers' experience: length of career was not associated with higher or lower scores (65% of Head Teachers reported having over 5 years experience as a Head Teacher,
- Number of days of training received by teacher or Head Teacher in the new curriculum (see below),
- Quality of training in the new curriculum as rated by the teacher or Head Teacher (see below),
- Ease of interpretation of the new curriculum as rated by the teacher or Head Teacher, (most Head Teachers and P3 teachers said the new curriculum was easy to interpret, yet this did not translate into improved pupil performance), or
- Attitudes toward the new curriculum.

Perspectives

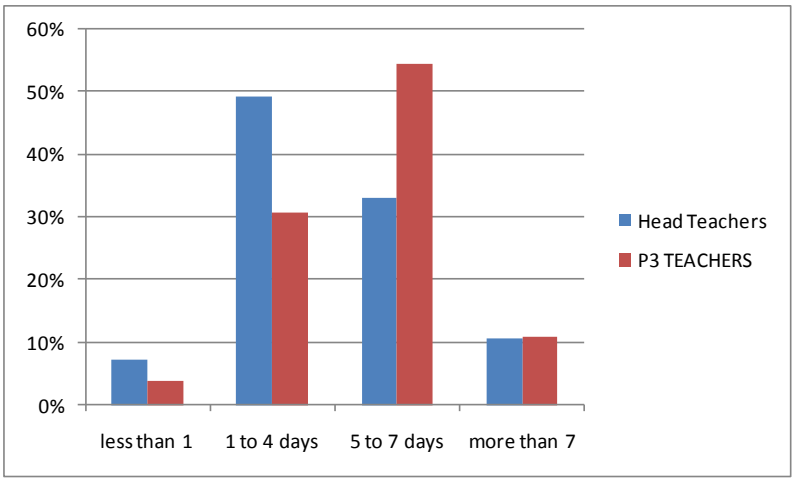
Summary of views concerning the new curriculum

The new curriculum consists of a number of important elements that would normally require considerable training if teachers and Head Teachers are to implement it well, including the thematic organization of content, use of local language as the medium of instruction, and the incorporation of continuous assessment as a central teaching/learning approach. A minimum of five days of training would therefore seem important if teachers and Head Teachers are to understand these important topics.

The following information is presented in order to provide a context for understanding teachers' and Head Teachers' perspectives on the length of their training, its quality, and the curriculum itself.

As Figure 8 illustrates, most P3 teachers received 5 or more days of training; however, more than one-third did not, and over half of Head Teachers received less than 5 days of training:

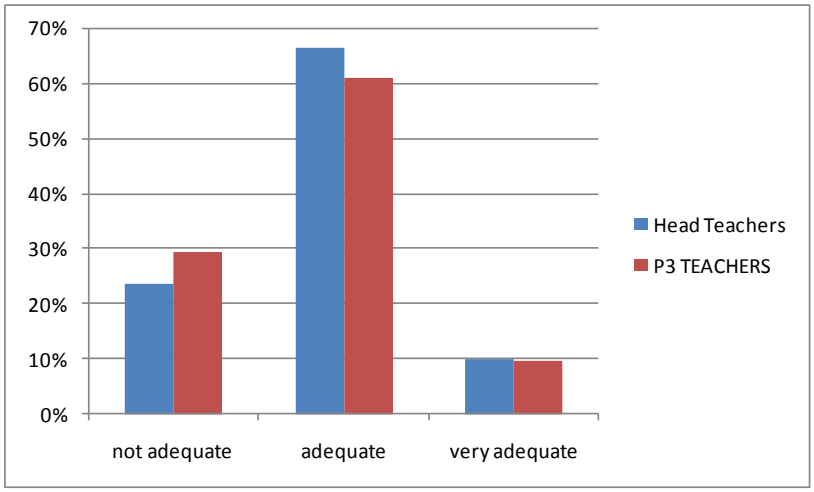
Figure 8: Number of days of training in the new curriculum
P3 teachers and Head Teachers



Number of training days	P3 teachers		Head Teachers	
	Number	%	Number	%
less than 1	4	4%	8	7%
1 to 4	31	31%	55	49%
5 to 7	55	54%	37	33%
more than 7	11	11%	12	11%
Total	101	100%	112	100%

For teachers and Head Teachers who received training, Figure 9 shows general satisfaction with the quality of the training, with three-quarters of each group reporting “adequate” or “very adequate” quality:

Figure 9: Quality of training received by P3 teachers & Head Teachers



Quality of training	P3 teachers		Head Teachers	
	Number	%	Number	%
Very adequate	9	9%	11	10%
Adequate	58	61%	73	66%
Not adequate	28	29%	26	24%
Total	95	100%	110	100%

Views on the training

When asked whether the curriculum training was sufficient, the main concern expressed by teachers and Head Teachers was that the time was too short and that as a consequence, not all material was covered. Numerous teachers and Head Teachers also commented on the lack of training materials, the need to train all teachers, the need to translate content (which some noted to be difficult), and the need to provide support for teaching in contexts where multiple languages are spoken. In order to improve the training, Teachers and Head Teachers made two key recommendations: that additional training be organized, and that materials such as reference books, assessment material and textbooks be provided. Teachers and Head Teachers also stressed the importance of including all teachers in training and increasing the level of remuneration during the training.

Views of the curriculum

Teachers and Head Teachers were asked: What is the biggest strength of the curriculum? The overwhelming response from both teachers and Head Teachers

was its emphasis on learning in the local language and how this helps children become literate and numerate while building a greater understanding of concepts. Teachers and Head Teachers also said the new curriculum facilitates the acquisition of practical knowledge, increases student participation, improves teaching skills, and benefits weaker learners. Some respondents also noted that the new curriculum is detailed and clear, child-centered, improves the learning environment (friendly learning atmosphere, learning is more interesting), that it builds student confidence, and that it puts an emphasis on continuous assessment.

When asked whether the curriculum was easy to interpret, the biggest problem cited by both teachers and Head Teachers was the difficulty translating the curriculum into local languages. Teachers and Head Teachers also cited the problem of insufficient teaching materials reference books, especially in local languages, problems associated with large class sizes and multiple languages in the classroom, and problems teaching subjects that are integrated. One teacher noted that the curriculum should be written in local languages.

When asked: What is the biggest weakness of the new curriculum? the biggest single response (almost half of teachers and Head Teachers) was "insufficient instructional materials." Also cited as a major weakness was the difficulty translating the curriculum. Other weaknesses cited include the amount of work placed on the teacher, the lack of relevance for students with difficulties (including disabled students), the reluctance of school communities to adopt the new curriculum, problems choosing a language in multi-lingual communities, and assessment and examination issues – e.g., lack of assessment booklets, problems aligning local language instruction with exams written in English, etc. Some teachers and Head Teachers feared that the new curriculum weakens students' English, which will place an additional burden on them in later years of schooling. It is instructive to note that though only 25% of respondents indicated that they preferred the old curriculum, those who did said that they preferred it because reference books are available, what is taught is examined, and the old curriculum is easy to interpret.

When asked for recommendations for improving the new curriculum, teachers and Head Teachers most frequently provided the following responses:

- Provide more instructional materials (more than half of respondents said this),
- Provide more training and refresher courses,
- Provide monitoring and supervision of the curriculum,
- Train more teachers,
- Sensitize parents and community members, and
- Keep transfers at a minimum for teachers trained in the new curriculum.

P4 pupil performance

Geography. Pupil scores in both language and maths were significantly higher in the Western Region than in the other three regions. Results in the North Region were significantly lower in language and maths.

Figure 10: P4 mean language scores by region

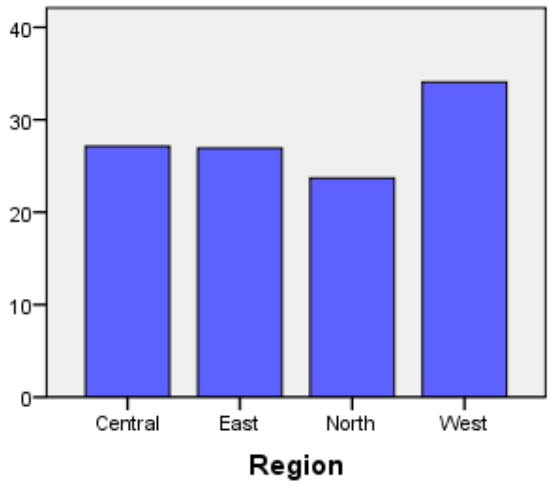
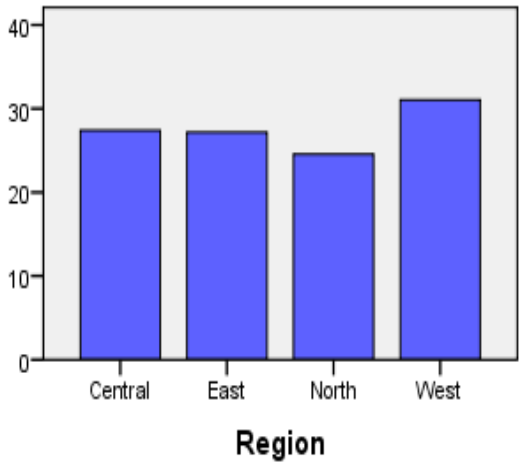


Figure 11: P4 mean maths scores by region

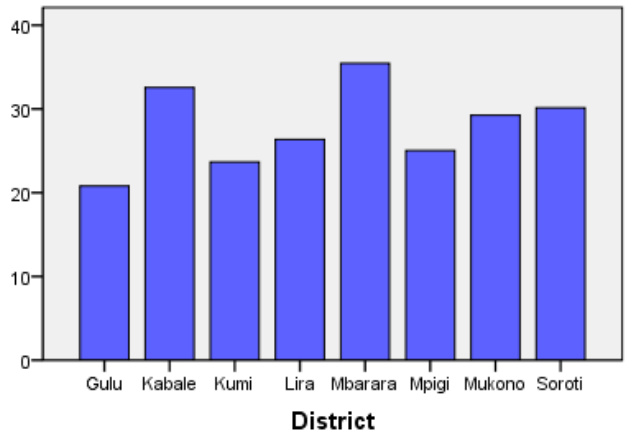


Region	Mean	Std. Deviation	N
Central	27.13	11.190	582
East	26.92	13.821	537
North	23.69	15.554	576
West	34.07	11.262	544
Total	27.88	13.616	2239

Region	Mean	Std. Deviation	N
Central	27.41	9.321	582
East	27.15	12.215	537
North	24.54	13.371	576
West	31.04	9.667	544
Total	27.49	11.499	2239

Maths and language scores in Mbarara were significantly higher than all other districts (except for Kabale in language), with the Gulu district reporting the lowest scores in language than all other districts except for Kumi. In maths, Gulu scores were significantly lower than all other discricts except for Kumi, Lira and Mpigi.

Figure 12: P4 mean language scores by district



District	Mean	Std. Deviation	N
Gulu	20.79	14.295	276
Kabale	32.55	12.439	260
Kumi	23.66	13.200	267
Lira	26.36	16.197	300
Mbarara	35.46	9.885	284
Mpigi	25.03	11.152	292
Mukono	29.24	10.843	290
Soroti	30.14	13.690	270
Total	27.88	13.616	2239

Language. Pupils who said their first language was Runyankole scored significantly higher than all other language groups (except for a few English and "other" language pupils) in both language and maths. Pupils from the Acoli language groups scored the lowest, both in maths (except for the Lango, English and "other" groups) and language tests.

Table 6: P4 mean language scores by language at home

Language at home	Mean	Std. Deviation	N
Acoli	20.87	14.344	277
Ateso	26.74	13.803	529
English	36.80	6.380	5
Lango	26.26	16.205	296
Luganda	27.55	11.177	578
Rukiga	32.04	12.410	270
Runyankole	35.90	9.810	247
Other	29.64	13.626	36
Total	27.88	13.619	2238

Table 7: P4 mean maths scores by language at home

Language at home	Mean	Std. Deviation	N
Acoli	23.40	12.484	277
Ateso	27.03	12.209	529
English	29.80	8.319	5
Lango	25.52	14.077	296
Luganda	27.63	9.259	578
Rukiga	28.15	10.029	270
Runyankole	34.14	8.544	247
Other	28.89	10.642	36
Total	27.49	11.502	2238

Experimental groups. Scores in P4 pupils in control schools (following the old curriculum from P2 and P3) were significantly higher than those in experimental schools (new curriculum for P2 and P3), both in language and maths, as shown in Table 14:

Table 8: P4 mean language and maths scores by experimental group

Subject	Ownership	N	Mean	Std. Deviation
Language	Control	657	39.65	8.830
	Experimental	1582	22.99	12.191
Maths	Control	657	35.68	9.040
	Experimental	1582	24.09	10.669

Age. Pupils under 10 years old performed significantly better than pupils 10 and older in both language and maths:

Table 9: P4 mean language scores by age group

Age	Mean	Std. Deviation	N
Under10YearOld	35.27	12.164	348
10YearOld	29.61	13.463	665
11YearOld	27.62	13.652	417
12YearOld	23.78	12.847	449
12+YearOld	22.81	12.226	349
Total	27.88	13.613	2228

Table 10: P4 mean maths scores by age group

Age	Mean	Std. Deviation	N
Under10YearOld	30.87	10.555	348
10YearOld	28.32	11.152	665
11YearOld	28.21	12.187	417
12YearOld	25.16	11.338	449
12+YearOld	24.64	11.316	349
Total	27.48	11.512	2228

Repeater status. Non-repeating pupils performed significantly better in language and maths than repeating ones.

Table 11: P4 mean language and maths scores by repeater status

Subject	repeat	N	Mean	Std. Deviation
Language	Yes	435	23.53	12.449
	No	1777	29.13	13.591
Maths	Yes	435	24.76	11.128
	No	1777	28.33	11.413

Home environment. Pupils with books at home (Table 18) and with mothers who read (Table 19) performed significantly better in language and maths than those without books or whose mothers who do not read.

Table 12: P4 mean language and maths scores by books at home

Subject	Books at home	N	Mean	Std. Deviation
Language	Yes	1716	29.34	13.512
	No	502	23.44	12.754
Maths	Yes	1716	28.51	11.344
	No	502	24.50	11.292

Table 13: P4 mean language and maths scores by mother who reads

Subject	Mother reads	N	Mean	Std. Deviation
Language	Yes	1455	29.77	13.525
	No	751	24.80	12.962
Maths	Yes	1455	28.73	11.375
	No	751	25.55	11.265

Gender. As in P3, differences in mean maths and language scores between girls and boys were very small; in fact, they were found to be not statistically significant. Analyses disaggregated whether by region or by first language show that no other statistically significant differences were found.

Table 14: P4 mean language and maths scores by sex of the pupil

Subject	Sex	N	Mean	Std. Deviation
Language	Girl	1086	28.60	13.311
	Boy	1152	27.21	13.873
Maths	Girl	1086	27.23	11.350
	Boy	1152	27.75	11.642

Language and maths comparison. Except for Gulu, Kumi and Mpigi districts, language scores were significantly higher than maths scores in all districts.

Table 15: P4 mean language and maths scores by district (in percent)

District/Subject		Mean (%)	N	Std. Deviation
Gulu	Language	42.42	276	29.174
	Maths	44.18	276	23.679
Kabale	Language	66.43	260	25.385
	Maths	53.70	260	19.134
Kumi	Language	48.28	267	26.938
	Maths	48.74	267	23.672
Lira	Language	53.80	300	33.055
	Maths	48.25	300	26.465
Mbarara	Language	72.37	284	20.173
	Maths	63.03	284	16.168
Mpigi	Language	51.08	292	22.760
	Maths	49.31	292	17.458
Mukono	Language	59.68	290	22.128
	Maths	54.16	290	17.408
Soroti	Language	61.50	270	27.938
	Maths	53.70	270	22.181

P4 correlations

86 P4 teachers were interviewed their years of experience, qualifications, class sizes, and the availability of instructional materials. Based on the information obtained in those interviews, as well as Head Teacher interviews discussed above, a number of correlations were found between P4 pupil performance and school characteristics:

- **Number of male or female teachers in school:** A statistically significant and positive relation was found between the number of teachers in a given *school*, whether male or female teachers, 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:** P4 scores tend to be higher when there is a library in the school, though the difference is not significant (see right). Interestingly, no correlation was found between P4 maths or language scores and the number of P4 books in the school library.

Table 16: Mean achievement by library in the school

Subject	Existence of library	N	Mean	Std. Deviation
LANGUAGE	Yes	48	30.8714	9.85563
	No	34	25.8413	11.66274
MATHS	Yes	48	29.7501	6.84263
	No	34	27.0611	10.06678

- **Materials owned by pupils:** Pupils with their own rulers scored significantly higher in language and maths.
- **Attitude toward the new curriculum:** In schools where Head Teachers said they "like the old curriculum better," P4 pupils scored significantly higher in maths and in language than pupils whose Head Teachers said the new curriculum was better or that the new curriculum was not different from the old one. (NB: the new curriculum had not yet been initiated in these schools; Head Teachers were asked this question based on their familiarity with the new curriculum as implemented in P1-3.)

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 like 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 as Head Teacher or highest qualification earned.

Comparative analyses

Because the overall purpose of this assessment is to determine whether pupils learn more under the new curriculum than the old, we will now turn to how pupils performed in 2009 using the new curriculum compared to how they performed in 2008 using the old. This section presents the results of the analyses of the assessments conducted in these two years, first by discussing global comparisons, then by discussing comparisons of control and experimental groups in greater detail.

Measuring progress in P3 from 2008 to 2009

A general look at the progress of P3 pupils from 2008 to 2009 shows significant improvement in performance in both language and in maths of all pupils. This measure combines the scores of experimental and control groups as follows:

Table 17: Mean comparison between the 2008 and the 2009 P3 cohorts, experimental and control groups combined

Subject	GroupYear	N	Mean	Std. Deviation
Language	2008	2294	18.4804	11.78060
	2009	3833	22.6306	10.44621
Maths	2008	2294	23.6212	10.50009
	2009	3833	25.4600	9.90147

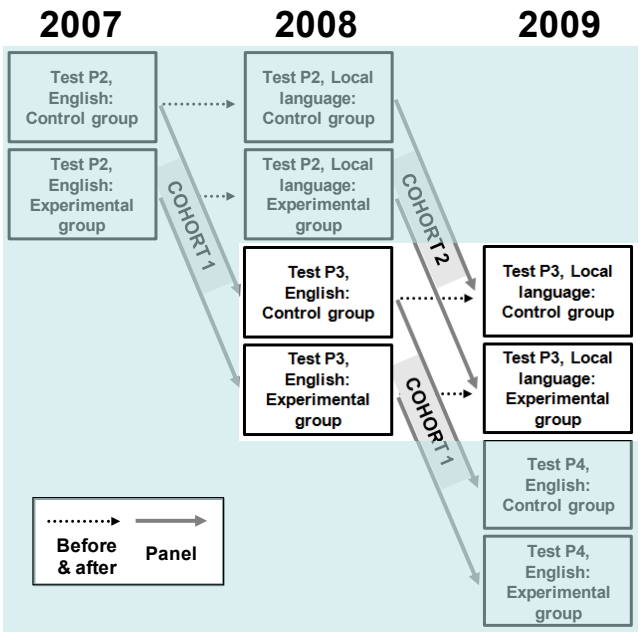
Table 18: T-tests comparing the 2008 and the 2009 P3 cohorts, experimental and control groups combined
(equal variances not assumed)

Subject	t	df	Sig. (2-tailed)
Language	-13.914	4378.660	.000
Maths	-6.776	4602.967	.000

Global comparisons such as these do not capture the different types of changes registered by pupils studying under the old curriculum (control group) compared to pupils studying under the new (experimental group). The statistical analyses presented in this section focus on these two groups, the first being represented by

the upper "before and after" dotted line in Figure 13 – i.e., comparing how P3 pupils in control schools performed between 2008 and 2009 – and the other being represented by the lower dotted line– i.e., how P3 pupils in experimental schools performed over the two years.

Figure 13: Comparing P3 2008 and P3 2009 cohorts



The experimental group is defined as the P3 pupils attending schools for which the 2009 language and maths tests were administered in any of the six local languages: Acoli, Ateso, Lango, Luganda, Rukiga, Runyankole. The control group includes pupils in schools where the 2009 language and maths tests were administered in English. As a reminder, this distinction assumes that the experimental schools, by definition, use local languages as the medium of instruction (since local language instruction is a feature of the new curriculum) whereas the control schools use English as the medium of instruction – suggesting that these schools have not adopted the new curriculum.

As was the case with P2 pupils in 2008, P3 pupils in the experimental group performed significantly better in 2009 under the new curriculum than in 2008 under the old in both language *and* maths, whereas their counterparts in the control schools showed no statistically significant differences between 2008 and the 2009, either in language *or* in maths.

Table 19: Mean achievement of the two cohorts of P3 pupils for the control and the experimental group

Group/Subject		GroupYear	N	Mean	Std. Deviation
Control	LANGUAGE	2008	676	28.7500	8.74738
		2009	853	29.2696	7.95365
	MATHS	2008	676	30.7012	8.48734
		2009	853	31.1618	8.05830
Experimental	LANGUAGE	2008	1618	14.1897	10.11678
		2009	2980	20.7302	10.29773
	MATHS	2008	1618	20.6632	9.82662
		2009	2980	23.8279	9.77578

The results discussed in the remainder of this section correlate pupils' performance by geography and pupil attributes: sex, age, repeater status, books in the home and having a mother who reads.

Geography. In 2009, the performance of P3 pupils in control schools in each of the 4 regions improved slightly in most cases except in the West where results were significantly better for the 2008 cohort both in language and in maths. The 2009 performance of P3 pupils in experimental schools, on the other hand, was significantly better in language in all 4 regions, but only significantly better in maths in the East and the Central regions.

Table 20: Mean achievement by region of the two cohorts of pupils for the control and the experimental group

Group	Region/Subject		GroupYear	N	Mean	Std. Deviation
Control	Central	LANGUAGE	2008	180	27.2500	6.96911
			2009	180	28.2444	7.15481
		MATHS	2008	180	29.4667	6.90025
			2009	180	30.2889	8.10854
	East	LANGUAGE	2008	153	28.5817	9.10221
			2009	157	29.1401	8.47278
		MATHS	2008	153	30.6471	8.10496
			2009	157	31.0255	7.12431
	North	LANGUAGE	2008	203	26.8276	10.44765
			2009	300	28.9267	8.02096
		MATHS	2008	203	28.7537	10.33874
			2009	300	30.9167	8.61455
	West	LANGUAGE	2008	140	33.6500	5.22298
			2009	216	30.6944	7.96324
		MATHS	2008	140	35.1714	5.80367
			2009	216	32.3287	7.77523
Experimental	Central	LANGUAGE	2008	411	15.6058	8.13936
			2009	514	23.2490	9.97718
		MATHS	2008	411	21.4696	8.18809
			2009	514	28.6518	8.18365
	East	LANGUAGE	2008	420	10.3262	8.78004
			2009	481	17.1913	9.08483
		MATHS	2008	420	17.2476	9.84149
			2009	481	21.1871	8.26201
	North	LANGUAGE	2008	384	8.8203	7.64769
			2009	954	17.3941	9.86317
		MATHS	2008	384	17.7161	9.39836
			2009	954	18.8407	9.55669
	West	LANGUAGE	2008	403	21.8883	10.17092
			2009	1031	24.2124	9.85854
		MATHS	2008	403	26.2084	9.10777
			2009	1031	27.2696	8.77303

Pupils' age. The difference in performance of pupils in each age group in control schools from 2008 to 2009 was minor, whereas for pupils all age groups in the experimental schools, the increase in performance on the language or the maths test from 2008 to 2009 was significant:

Table 21: Mean achievement by age of the two cohorts of pupils for the control and the experimental group

Group	Age/Subject		GroupYear	N	Mean	Std. Deviation
Control	Under9 YearOld	LANGUAGE	2008	243	30.4568	7.85675
			2009	216	30.2639	7.18654
		MATHS	2008	243	30.1564	8.20251
			2009	216	30.3102	7.29836
	9YearOld	LANGUAGE	2008	197	28.3096	9.14512
			2009	337	29.5697	8.09516
		MATHS	2008	197	29.9645	8.81810
			2009	337	31.4866	8.16135
	10YearOld	LANGUAGE	2008	153	27.5490	9.41579
			2009	173	28.0809	8.52522
		MATHS	2008	153	31.3268	8.55772
			2009	173	30.6127	8.52120
	11+YearOld	LANGUAGE	2008	83	27.0120	8.28236
			2009	123	28.4634	7.77690
		MATHS	2008	83	32.8916	8.06379
			2009	123	32.6260	8.25158
Experimental	Under9 YearOld	LANGUAGE	2008	234	17.0342	11.32238
			2009	249	23.2851	10.76336
		MATHS	2008	234	20.9103	9.86369
			2009	249	25.9116	10.09313
	9YearOld	LANGUAGE	2008	332	14.5753	10.48479
			2009	605	22.0132	10.07586
		MATHS	2008	332	20.3825	9.76312
			2009	605	23.8017	9.86231
	10YearOld	LANGUAGE	2008	521	13.4376	9.60951
			2009	963	20.5265	10.25466
		MATHS	2008	521	20.0345	9.83258
			2009	963	23.7508	9.83788
	11+YearOld	LANGUAGE	2008	531	13.4331	9.58403
			2009	1151	19.6811	10.19173
		MATHS	2008	531	21.3465	9.82433
			2009	1151	23.4544	9.56094

Pupils by repeater status. Pupils who reported having repeated P3 in control schools in 2009 scored slightly higher than repeaters in 2008, though no differences were found to be significant. In experimental schools, on the other hand, both repeaters and non-repeaters showed significant increases both in language and maths (level of significance = .01)

Table 22: Mean achievement by repeating year of the two cohorts of pupils for the control and the experimental group

Group	Repeat/Subject		GroupYear	N	Mean	Std. Deviation
Control	Yes	LANGUAGE	2008	95	26.0105	8.50969
			2009	100	26.1700	6.98347
		MATHS	2008	95	29.6211	8.76062
			2009	100	27.9300	7.70315
	No	LANGUAGE	2008	574	29.2666	8.66239
			2009	748	29.7420	7.95126
		MATHS	2008	574	30.9495	8.37572
			2009	748	31.6471	7.99689
Experimental	Yes	LANGUAGE	2008	403	12.8908	8.56831
			2009	644	20.0807	10.18459
		MATHS	2008	403	20.6129	9.21598
			2009	644	24.5528	9.84066
	No	LANGUAGE	2008	1175	14.8894	10.53386
			2009	2098	21.5486	10.26301
		MATHS	2008	1175	20.9821	9.99543
			2009	2098	24.2269	9.65065

Pupils with books at home and mothers who read. P3 pupils in control schools reporting having books in the home or a mother who reads in 2009 performed slightly better than P2 pupils in 2008, but again, no significant differences were found. However, pupils in experimental schools performed significantly better both in language and maths whether they had books at home or not, or a mother who reads or not:

Table 23: Mean achievement by books at home of the two cohorts of pupils for the control and the experimental group

Group	Books in home/Subject		GroupYear	N	Mean	Std. Deviation
Control	Yes	LANGUAGE	2008	560	29.1804	8.68107
			2009	721	30.1567	7.43073
		MATHS	2008	560	30.9821	8.42666
			2009	721	31.6574	7.90098
	No	LANGUAGE	2008	112	26.8304	8.77639
			2009	129	24.5039	8.90236
		MATHS	2008	112	29.4196	8.73291
			2009	129	28.5814	8.32794
Experimental	Yes	LANGUAGE	2008	1006	16.0586	10.30517
			2009	1678	22.0858	9.97593
		MATHS	2008	1006	21.9225	10.09596
			2009	1678	25.1508	9.40365
	No	LANGUAGE	2008	572	11.3462	9.08140
			2009	1004	19.9193	10.41552
		MATHS	2008	572	18.9388	9.05798
			2009	1004	23.0149	10.02454

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

Group	Mother who reads/Subject		GroupYear	N	Mean	Std. Deviation
Control	Yes	LANGUAGE	2008	506	29.7767	8.48199
			2009	697	29.8637	7.44308
		MATHS	2008	506	31.0277	8.47309
			2009	697	31.5409	7.89197
	No	LANGUAGE	2008	164	26.0122	8.41412
			2009	152	26.8618	9.43086
		MATHS	2008	164	30.0488	8.14124
			2009	152	29.6776	8.50290
Experimental	Yes	LANGUAGE	2008	917	16.0098	10.32236
			2009	1399	22.2723	10.18257
		MATHS	2008	917	21.7296	9.68200
			2009	1399	25.1129	9.65641
	No	LANGUAGE	2008	650	12.1815	9.34550
			2009	1255	20.0757	10.23822
		MATHS	2008	650	19.8323	9.79109
			2009	1255	23.4749	9.77681

Pupil's sex. The difference in performance of both girls and boys in control schools from 2008 to 2009 was minor, whereas both girls and boys in the experimental schools saw a significant improvement in scores, especially in language, from 2008 to 2009:

Table 25: Mean achievement by sex of the two cohorts of pupils for the control and the experimental group

Group	Sex/Subject		GroupYear	N	Mean	Std. Deviation
Control	Girl	LANGUAGE	2008	323	29.6687	7.91601
			2009	400	29.4150	7.57339
		MATHS	2008	323	30.6718	8.61265
			2009	400	30.8325	8.04744
	Boy	LANGUAGE	2008	352	27.9659	9.32961
			2009	450	29.1067	8.29668
		MATHS	2008	352	30.8068	8.26327
			2009	450	31.4489	8.07472
Experimental	Girl	LANGUAGE	2008	824	15.0595	10.19304
			2009	1528	21.1466	10.15612
		MATHS	2008	824	20.7694	9.81197
			2009	1528	23.8737	9.64758
	Boy	LANGUAGE	2008	790	13.3418	9.95756
			2009	1440	20.2715	10.44250
		MATHS	2008	790	20.6127	9.82815
			2009	1440	23.7569	9.92911

Measuring progress in cohort 2, P2 to P3, 2008 to 2009

This part of our analysis concerns the comparison of scores between control and experimental groups in cohort 2 from 2008 to 2009, as represented by the solid lines in the figure at right. This part of the MLA, called a panel design, calls for an assessment of the progress made by the P2 2008 cohort as they move to P3 in 2009 using the new curriculum. To make this comparison, we used a procedure called test equating or linking. In order to ensure that scores from year to year are comparable, scores from tests over the two years must be "equated" – otherwise, we don't know if the tests were of equal difficulty and of course, if we don't know this, we can't compare the scores. 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 equivalent 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, then the scores adjusted to so they can be compared on the same scale.

Figure 14: Measuring progress in cohort 2

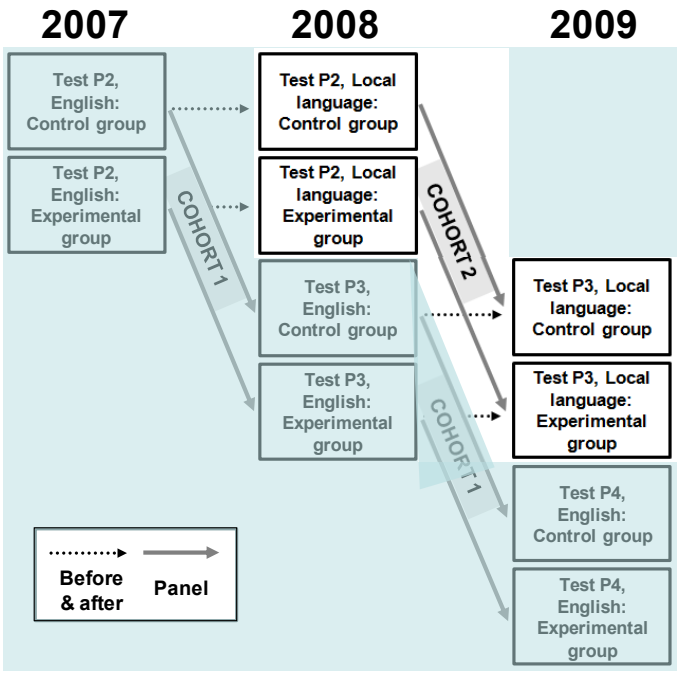


Table 32 below shows the raw scores of the pupils, anchor item (or testlet) scores, and "expected true" or equated scores – the ones used to compare P2 2008 baseline with the P3 2009.[4] All scores are reported as percentages (the standard practice when equating scores)[5].

[4] 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 (2008 P2 and 2009 P3) and each test (language and maths) was calibrated using Samejima's graded model (using «random» option). Then POLYST software was used to equate the 2009 calibrated testlets («new» test) to the 2008 scale («old» test). The theta scores (MAP) were also obtained (using MULTILOG «score» option) for each cohort and each test: the 2009 theta scores were then equated to the 2008 theta scale. Now that the theta scores from the two cohorts were on

Table 26: Equated mean ability in language and maths for P2 2008 and P3 2009

	Language		Maths	
Score	**P208**	**P309**	**P208**	**P309**
Total				
Raw	58.21	56.58	49.70	53.86
Anchor	54.22	70.63	42.88	63.42
True	64.25	74.86	49.88	70.98
Control				
Raw	79.18	73.17	62.69	65.78
Anchor	77.61	88.05	53.55	71.61
True	84.90	89.04	60.99	80.80
Experimental				
Raw	52.03	51.83	45.88	50.45
Anchor	47.34	65.64	39.74	61.08
True	58.17	70.79	46.60	68.17

Using these expected true scores, the following mean scores (Table 27) show that the pupils taking the P3 test had significantly higher scores in 2009 than by taking the P2 test in 2008 on both language and maths. Importantly, differences for the experimental group were greater than for the control group for both language and maths, indicating a *cumulative effect* for the new curriculum as well, especially for the experimental group. All four comparisons shown in Table 34 were found to be statistically significant.

Table 27: Comparing the P3_2009 expected true scores equated to the P2_2008 expected true scores

Group	Score/subject	Group & year	N	Mean	Std. Deviation
Control	True_language	P2_2008	859	84.90	19.932
		P3_2009	853	89.04	14.516
	True_maths	P2_2008	859	60.99	17.575
		P3_2009	853	80.80	13.406
Experimental	True_language	P2_2008	2917	58.17	27.943
		P3_2009	2980	70.79	26.401
	True_maths	P2_2008	2917	46.60	19.438
		P3_2009	2980	68.17	18.814

the same scale they were transformed to expected true scores using the test characteristic curve of the P2 2008 cohort: software GAUSS-IRT was used for that purpose.

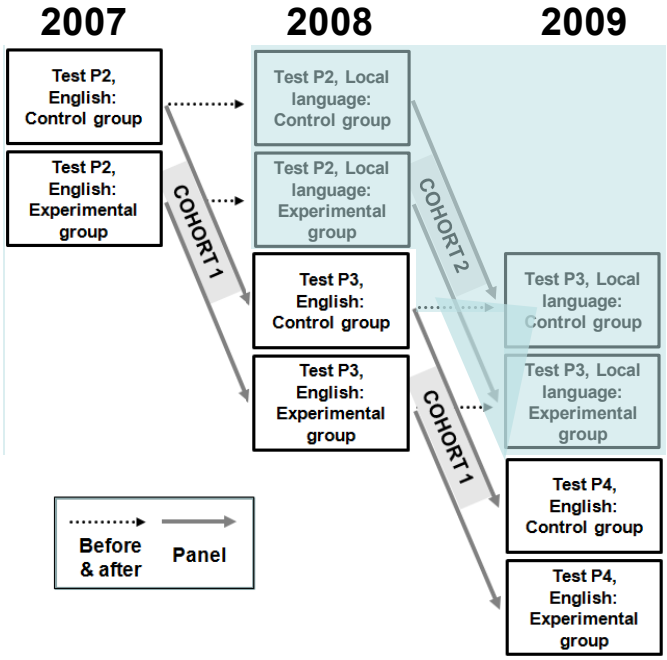
[5] In order to follow the same procedure as last year for P2 2007 – P3 2008 (all pupils had a booklet in English), item language 6.3 had to be removed from testlet 6. Also because language testlet 9 and maths 8 were DIF for P2 2008 and so excluded from last year's comparisons, they were also excluded from the P2 2008 database for the following analyses. Finally, because maths testlet 8 was also an anchor P208-P309 (m8t in P2 = m7t in P3) it was deleted from the anchor testlets for the maths test and excluded from the P3 2009 database for the following analyses.

Table 28: T-test for the comparisons of the P3_2009 expected true scores equated to the P2_2008 expected true scores (equal variances not assumed)

Group	Score/subject	t	df	Sig. (2-tailed)
Control	True_language	-4.919	1568.671	.000
	True_maths	-26.230	1603.591	.000
Experimental	True_language	-17.828	5859.295	.000
	True_maths	-43.285	5877.854	.000

Measuring the progress of cohort 1, P2 to P4, 2007 to 2009

Figure 15: Measuring progress in cohort 1



The last part of our analyses concerns the panel design: to establish a baseline, four hundred and twenty[6] pupils from the last group to use the old curriculum were followed from 2007 (P2) to 2009 (P4): 148 pupils from the control schools and 272 pupils from the experimental schools[7]. For the purposes of this study, results from this group only provide a baseline measure of how pupils performed from year to year under the old curriculum. It does not provide a measure of how pupils perform from year to year under the new curriculum; this comparison will be conducted in the MLA 2010.

It is nevertheless interesting to note three tendencies illustrated by the graphs on the following pages. The first has already been seen before and should be expected: that pupils in control (mostly private) schools consistently scored higher than those in experimental schools. The second tendency is that from Year 1 to Year 3, scores

[6] In fact, of the initial 2,325 P2 pupils, only 420 could be followed individually and tested from 2007 to 2009 for a variety of reasons, including significantly high pupil transfer rates reported by test administrators. This small sample means that findings should be interpreted with caution – see Annex qqq for additional information.
[7] Note here that this cohort of pupils, whether from the control or the experimental schools, all took the test based on the **old curriculum**. Next year, we will analyse the cohort of pupils submitted to the **new curriculum** during three years: 2008 (P2), 2009 (P3) and 2010 (P4).

went up substantially for both control and experimental groups in both language and maths over time (though were flat in language from 2007 to 2008). This suggests reinforces what we have found elsewhere – that the new curriculum is having a "spillover effect" in which pupils benefit even if they are not part of the direct beneficiary group (i.e., using the new curriculum). The third tendency is that language appears to be affecting pupils in the experimental group differentially, showing dramatically greater gains in 2009 than their peers in control schools with an approximately 15 point increase compared to roughly 7 point gain in the control group. Again, this reinforces the finding discussed elsewhere that language gains appear to be the most pronounced with the introduction of the new curriculum – in this case, even with pupils who are exposed to the new curriculum in their schools but who are not yet directly benefitting from it.

For a discussion of how these measures were taken, see Annex 2:

Figure 16: Plot of the mean equated/linked language true scores for pupils in the control and the experimental schools in P2 2007, P3 2008 and P4 2009

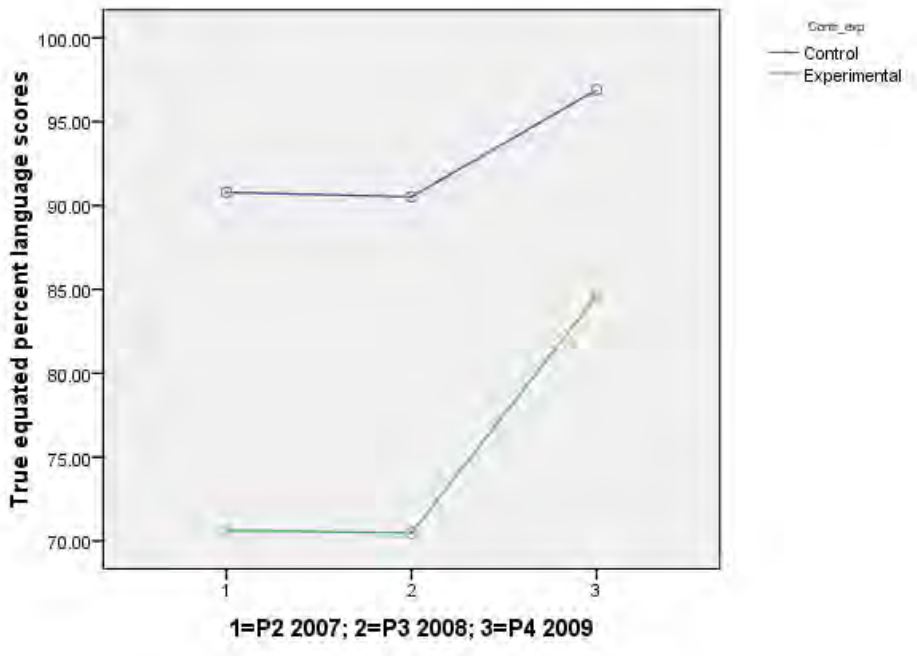
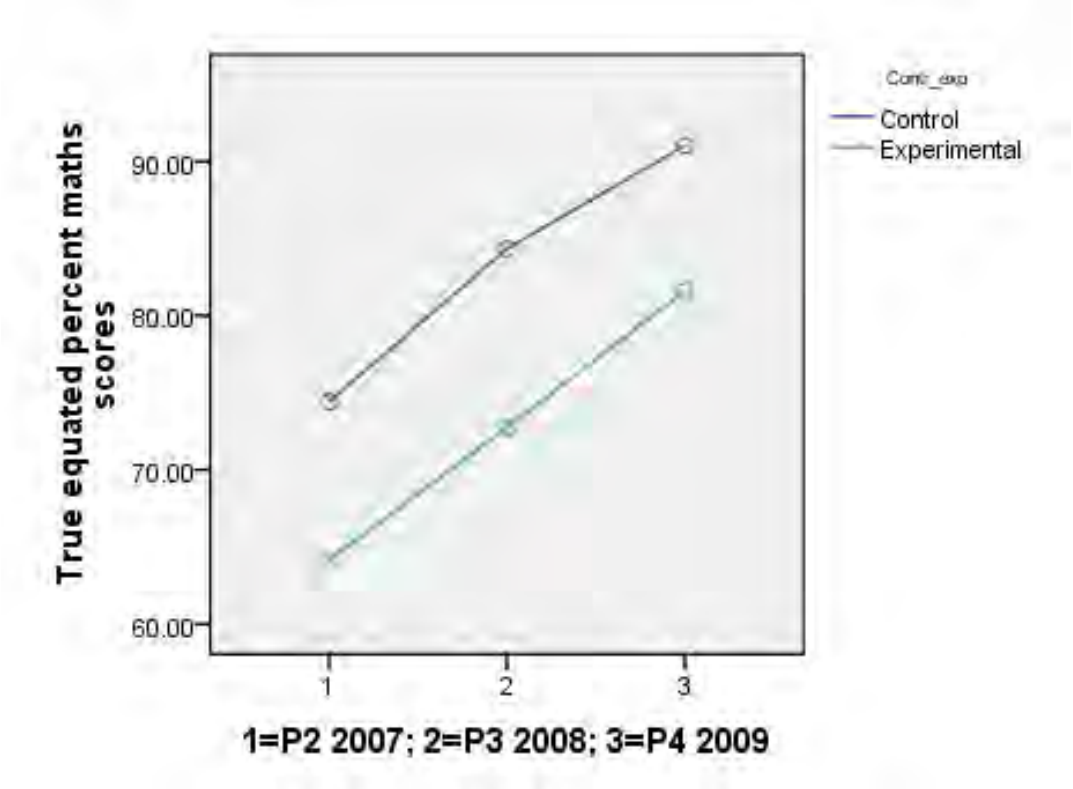


Figure 17: Plot of the mean equated/linked maths true scores for pupils in the control and the experimental schools in P2 2007, P3 2008 and P4 2009



Discussion

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

Continued effect of "floating all boats." The panel analysis we've just discussed illustrates a pattern of rising scores for pupils in both experimental *and* control schools over the last three years – in effect, the new curriculum appears to be "floating all boats." This pattern was first identified in the 2008 MLA with the finding that the introduction of the new curriculum in a school seems to benefit all pupils, even ones who have not yet received it. For example, P3 pupils were still learning under the old curriculum in 2008, yet their scores were higher than P2 pupils' scores were the year prior, also under the old curriculum. Over the same period, control school scores rose as well, where the new curriculum is presumably not being implemented. Since this pattern is now being detected for a second year, it is less likely due to the Hawthorne effect (where spikes occur in the beginning of

an intervention due to the excitement generated by the new activity), and perhaps more likely due to something else. It is not clear what that "something else" might be, though all large-scale interventions tend to introduce a new language, a new rhythm, and a new way of doing business – effects that are felt by all, not just target populations – i.e., a "spill-over effect." Still, this spill-over effect seems to be affecting having a stronger impact on the intervention group – in essence, floating all boats, but some boats higher. A comparison of the *improvement* between control and experimental groups over time shows that pupils moving from P2 (2008) to P3 (2009) using the new curriculum showed greater gains (+12% in language and +22% in maths) than pupils from P2 (2007) to P3 (2008) using the old curriculum (+6% in language, +15% in maths). Future assessments, both the MLA and comparable ones, will provide an opportunity to test whether Hawthorne, spill-over or other effects are in play. Still, current results are encouraging.

Abiding effect of the new curriculum: For the first time, MLA 2009 was able to track a complete cohort studying under the new curriculum from one year to the next – from P2 in 2008 to P3 in 2009. Preliminary evidence shows that gains for this experimental group in both language and maths were significant, whereas gains for the control group were not. Importantly, this shows that significant increases are being sustained as these pupils move through the system.

"Affirmative action" for disadvantaged pupils: As was the case in 2008, disadvantaged pupils (ones whose mothers do not read or had no books in the home) performed significantly better in 2009 with the new curriculum. Repeaters also made significant improvements with the new curriculum and did not with the old. This finding holds significant promise both for Uganda and for other countries seeking ways to close the gap between traditionally high-performing children and ones in need of more assistance.

Correlations as insight into systemic patterns: Some correlations identified in MLA 2009 made intuitive sense. For example, P3 language and maths scores were higher when teachers reported having access to dictionaries and other materials. Pupil performance was significantly higher when they had access to exercise books and rulers. The existence of a library was positively correlated with student performance, sometimes significantly.

Perhaps equally revealing was what was *not* found to be correlated. For example, teachers' qualifications were not correlated with positive pupil performance, and in the case of P3, pupils' whose teachers had less experience scored *higher* in language and maths. No correlation was found between the length of teachers' or Head Teachers' careers and student performance. And though the existence of libraries was consistently associated with stronger performance, the number of books, or types of books, in the libraries did not correspond to pupil performance in any consistent way. For example, no correlation was found between P4 performance and the number of P4 books in the school library, and in one bizarre example, the greater the number of science books in school libraries, the better pupils' language scores. In some cases, the absence of a correlation was a good thing, such as the fact that performance was not tied to the proportion of boys or girls in a school, or the sex of the Head Teacher. Perhaps a disappointing finding, also found in 2008, was the absence of a correlation between number of days of training received by teachers or Head Teachers in the new curriculum and student performance.

Then there's the pesky case of inconsistent correlations. For example, the number of male teachers in a school had a significant impact on P3 performance, but not on P4. Head Teachers' qualifications correlated with stronger performance in P3 but not in P4. Teachers' access to dictionaries improved pupil performance in P3 but not in P4.

What can be learned from these patterns? One pattern seems apparent: things like qualifications and access to materials were more likely to correlate with performance in P3 (where the new curriculum had been implemented) than in P4 (where it had not), suggesting that simply implementing the new curriculum activates other school resources. Of course, one year's measure is insufficient to draw any conclusions, but based on correlations found this year, the following conclusions can be drawn:

1. Several positive signs were found in Ugandan schools in the 2009 MLA, including comparable performance of boys and girls, results that were achieved regardless of the sex of their teacher or Head Teacher, and a strong relationship between performance and the use of some instructional

materials such as exercise books, rulers, indicating that these materials are probably being used well.

2. While having a library and access to materials increases the chances of improved student performance, how materials are used is probably a stronger predictor, suggesting that in some instances, materials are available but not being used to maximum effect.
3. Personnel qualifications or years of service do not predict student achievement – a pattern found throughout the world that reminds us that initial qualification or experience does not suffice for targeted support that results in empirical improvements in learning.
4. The amount of training received by teachers under the curriculum could not be correlated with impact on student performance, meaning that 5 or more days produced effectively the same results as 1 or 2.

Difficulties with the panel design: Finally, as noted earlier in the report, substantial efforts were made to track the same pupils from 2007 to 2008 to 2009. In 2008 and 2009, specific instructions were given to administrators to select students who had been tested the since 2007 when they selected pupils to participate in the test. Administrators read children's names from lists and if the children were not present, asked teachers how they could be found. Administrators then filled in forms indicating how many children had been tracked and for those not identified, reasons were given for why the child might be absent. The reason most often cited in these forms for pupils' absence was transfers, with comments like "majority of pupils in school transferred back to their original sites from the camps" and "school enrolment has drastically reduced due to transfer of pupils to other schools." Anomalies in tracking rates across MLA schools also suggest that some administrators performed well while others simply did not. Whatever the reason, the panel group in 2009 consisted of approximately 20% of the original group (420 of the 2,325 pupils in the initial 2007 cohort) – too small a number to generalize to the entire population, but still useful for analyses that illustrate types of gains observed.

Recommendations

Based on the findings from the 2009 MLA, the following recommendations are made:

1. **Training**: In light of concerns raised by teachers and Head Teachers about the training being to brief, and in light of the finding that the amount of training received by teachers was not correlated with on pupil performance, it seems clear that additional training would strengthen outcomes of the reform. 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.

2. **Materials**: In light of findings that types and quantities of materials in libraries were not necessarily correlated with outcomes, more strategies use of these materials would probably benefit teachers and pupils alike. It is therefore recommended 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 the transition to the new curriculum.

   NB: the need for additional training and materials were also the biggest concerns expressed by teachers and Head Teachers in last year's MLA. The fact that the same pattern occurred this year raises a concern about the sustainability of the reform. It is our opinion that if these two key aspects of the reform are not corrected, the impressive gains made so far in the could be eroded, leading to a "backwash effect" (the inverse of the Hawthorne effect) in which all evidence of improvement resulting from the new curriculum disappears. Attending to teachers' and Head Teachers' repeated pleas for support is imperative if such an outcome is to be avoided.

3. **Language issues**: Examine the issue of local language instruction in schools where multiple languages are spoken, and what kinds of support teachers,

Head Teachers and parents need to accommodate local language instruction in these contexts.

4. **Let the world know!** Results from the 2008 and 2009 MLA are impressive. Ugandan decision makers and stakeholders (and the world) should know what is possible with such a reform. Therefore, publicize the results of MLA 2008 and 2009. Points to highlight should include increased student participation and mastery of concepts and basic skills at a young age, sustained improved outcomes, and the differential effect the new curriculum is having on disadvantaged children – that with the old curriculum, the "normal" children in experimental schools showed greater gains than their peers in private schools; with the new curriculum, *all* children in experimental schools benefit.

5. **Data quality:** Examine procedures for monitoring the quality of test administration to ensure uniform implementation of tests and interviews and selection of pupils. This recommendation aims to avoid the problem of missing data in the 2010 MLA.

Annex 1: Methodology

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 and control schools continuing to use the old. In order to make this distinction, several assumptions were made:

1. In most cases, private schools would continue using the old curriculum and government schools would use the new one – this because many parents opt to send their children to private schools because English is the medium of instruction.

2. The use of English as the medium of instruction was taken as a proxy for use of the old curriculum; similarly, the use of local language as the medium of instruction served as a proxy for adoption of the new curriculum.

In 2008 and again in 2009, 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. For the 2009 MLA, the schools used in the P2 and P3 2008 MLA were retained respectively for the P3 and P4, – 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 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 student achievement testing. The goal was to be able to generalize the findings of the MLA to the entire P2 and P3 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 calculated as 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[8] 117 schools and 3,776 P2 pupils from 146 schools[9] (including the 117 P3 schools) participated in the 2008 MLA.[10] 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. The total numbers of schools and P3 and P4 pupils in the 2009 MLA were as follows:

Table 29: MLA 2009: Number of schools and P3 pupils

	Schools			Pupils		
Region	**Experimental**	**Control**	**Total**	**Experimental**	**Control**	**Total**
Central	21	9	30	514	180	694
East	20	8	28	481	157	638
North	33	10	43	954	300	1254
West	35	7	42	1031	216	1247
Total	**109**	**34**	**143**	**2,980**	**853**	**3,833**

Table 30: MLA 2009: Number of schools and P4 pupils

	Schools			Pupils		
Region	**Experimental**	**Control**	**Total**	**Experimental**	**Control**	**Total**
Central	21	9	30	403	179	582
East	20	8	28	379	158	537
North	20	9	29	396	180	576
West	21	7	28	404	140	544
Total	**82**	**33**	**115**	**1,582**	**657**	**2,239**

[8] Mengo school replaced Ryamihanda school in 2008 sample.
[9] A few schools from the 6 districts were added to the sample to insure the representativity of the control group.
[10] 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.

Test development

P3

The same P3 test used in 2008 was used again with P3 pupils in 2009. For pupils in schools where the medium of instruction was English, it was administered in identical form. However, for pupils using a local language as the medium of instruction (under the new curriculum), the test had to be translated into the six local languages. To accomplish this, six translators, mostly retired teachers, were hired. Each translator translated the P3 test, item by item, from English into one of the six local languages, then reviewers were recruited to translate the items back into English in order to verify the accuracy of the translation. This procedure, called *back-to-back translation*, safeguards against bias that can occur as a result of items in one language being more difficult than in another due to faulty translation. The end product was six local language copies of the P3 baseline English test.

P4

A specifications table or *test blueprint* was developed in order to ensure that different types of thinking skills were being tested across the different competencies. Then, two item writing workshops were conducted simultaneously for P4, one for the language test and one for the maths test. Each workshop was attended by eight people: five primary level teachers, a language expert or a maths expert from the NCDC, the UNITY Monitoring and Evaluation Specialist, and an outside consultant. 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 to P4 outcomes, a subset of items common to each test, called *anchor items*, was selected from the P3 tests to be included in the P4 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, then organized into two versions for pilot testing. Also developed were teacher's 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 **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 (10 groups of items, called *testlets*) and one maths test (17 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. For P3, the administrator then asked the teacher in which language he/she conducted instruction; each administrator was given sets of tests in the dominant language of his/her district as well as English. For P4, only English tests were administered. The administrator distributed the tests 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 P3 and 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 and analysis.

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 Differential Item Functioning (DIF) analysis (also described below). 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 because in those cases, multiple analyses were run with the same data sets – a practice that, when the .05 significance level is used, can raise the error rate above 5%.

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 student achievement, three 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
- item bias that favors one type of learner over another (e.g., boys over girls, English pupils over Acoli pupils).

This section describes measures taken in these three 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 2009 MLA, the size of the population and the sampling design yielded the following statistics for P3 in 2009 (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 0.186,
- for the maths test, the 95% margin of error of the mean was 0.171.

In the case of P4 2009:

- for the language test, the 95% margin of error of the mean was 0.337
- for the maths test, the 95% margin of error of the mean was 0.256.

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

P3 and 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. [11]

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 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." [12] To address this problem, items were grouped and analyzed as "testlets" - a technique described in Thissen & Wainer[13] (2001). A description of the analyses used to assess the psychometrical properties of the items and the tests follows.

Cronbach alpha

The first analysis of test reliability is Cronbach 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 thing, or general construct (e.g. math ability in P2). A Cronbach alpha coefficient of 0.7 or higher (the maximum possible is 1) is generally considered acceptable in student achievement testing.

[11] See also Bertrand & Blais (2004) for fuller descriptions of these procedures.
[12] A procedure called Yen's Q3 is used to verify the degree of local dependency.
[13] Thissen, D., & Wainer, H. (eds) (2001) Test Scoring. Lawrence Erlbaum Associates.

Cronbach alpha is usually presented as a single figure for an entire test. As can be seen in Table 3 below, both the P3 and P4 language tests obtained extremely high Cronbach's alpha measures, indicating a very high level of internal consistency:

Table 31: Cronbach alpha scores, P3 and P4

Class	Language	Maths
P3	.875	.879
P4	.937	.885

The 17 testlets in the P4 maths Cronbach's alpha measure of 0.885, indicating a very high level of internal consistency. Cronbach alpha was also run by language, yielding very high results as well:

Table 32: Cronbach's alpha for the P3 language and maths test, analyzed by testlets

	Language test		Maths test	
Language of booklet	Cronbach's Alpha	N of testlets	Cronbach's Alpha	N of testlets
Acoli	.858	10	.882	17
Ateso	.845	10	.825	17
English	.868	10	.849	17
Lango	.865	10	.838	17
Luganda	.896	10	.851	17
Rukiga	.890	10	.881	17
Runyankole	.855	10	.760	17

Testlet difficulty index

The next analysis provides a "testlet difficulty index," or the mean score of all students for each testlet. The higher the value, the easier the test was for the pupils. For the P3 language tests, the values of the difficulty index were calculated in each of the 6 local languages in which the test was taken plus English. These values fell with an acceptable range, though varied from test to test and language to language. For example, in Luganda, testlet 1 was found to be much more difficulty than testlet 4. Similarly, for the P4 language tests, some language testlets were found to be very difficult (e.g., testlet 4) while others were rather easy (e.g., testlet 7) for the P4 pupils in 2009 - see Tables 5 and 6:

Table 33: P3 testlet means, Luganda booklet, language test

Testlet	Mean	Maximum score	Std. Deviation	N
1	1.28	4	1.378	514
2	1.33	4	1.323	514
3	3.80	5	1.709	514
4	3.30	4	1.156	514
5	1.58	4	1.387	514
6	3.05	4	1.474	514
7	3.23	4	1.263	514
8	2.62	4	1.543	514
9	1.42	4	1.360	514
10	1.64	3	1.185	514

Table 34: P4 testlet means, English

Testlet	Mean	Maximum score	Std. Deviation	N
1	1.81	4	1.532	2239
2	3.14	5	1.783	2239
3	2.28	4	1.430	2239
4	1.77	4	1.521	2239
5	1.48	3	1.232	2239
6	2.31	4	1.632	2239
7	2.46	4	1.153	2239
8	1.73	4	1.439	2239
9	1.93	4	1.482	2239
10	3.10	4	1.279	2239
11	3.57	5	1.667	2239
12	2.28	4	1.470	2239

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 students with high exams scores get an item correct, and students with low exam scores get the item incorrect. The item-total correlation is a measure of this relationship – i.e., how well each student performed on each item relative to his/her total exam score. The closer to 1 (the maximum), the greater the discrimination.

Most of the item-total correlations were found very high in P3 and P4 tests in both language and maths, meaning that they provide a high level of discrimination between pupils of different abilities. Table 7 presents the item-total correlations for the P3 Luganda pupils in language as an example, and Table 8 for the P4 maths

test (see Technical Report for 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 alpha coefficient presented above.

Table 35: P3 Luganda language test item-total correlations

Testlet	Corrected Item-Total Correlation
1	.611
2	.678
3	.738
4	.526
5	.667
6	.731
7	.598
8	.737
9	.416
10	.738

Table 36: Item-total correlations for P4 maths test

Testlet	Corrected Item-Total Correlation
1.1	.548
1.4	.453
2.1	.516
2.3	.562
2.5	.645
3.1	.612
4.1	.430
5.1	.384
5.2	.449
6.1	.662
7.1	.544
8.1	.724
9.1	.592
10.1	.619
10.5	.452
11.1	.344
12.1	.652

DIF analysis

Our final evaluation of test quality focused on item bias – in this case, item bias related to cultural or translation problems. The key question is: did any testlets show bias for or against pupils as a result of taking the test in any of the six local languages in the P3 2009 cohort? To measure this, a statistical procedure called Differential Item Functioning (DIF) was performed on all testlets in language and maths – specifically, a technique called "Raju's NCDIF statistic for polytomous items" (Bertrand & Blais, 2004). This procedure identified a number of testlets presenting cultural or translation DIF as shown in Tables 9 and 10 below (note that "x" indicates that DIF – i.e., an unacceptable difference in scores, probably attributable to language or culture - was found for this language group). For these analyses, the performance of each of the 6 local language sub-cohorts (the experimental group) is being compared to the reference, or English-speaking group (control group). For our purposes, if an item was found to be "DIF" across all

language groups, it would be eliminated from the analysis due to probable language bias. The DIF analysis found **no** testlets that were "**DIF**" across all language groups. Table 9 shows DIF analysis findings for the language test.[14] Detailed DIF analyses and results can be found in the Technical Report.

Table 37: DIF analysis results, language P3

Testlets	Acoli	Ateso	Lango	Luganda	Rukiga	Runyankole
1	X	X	X		X	X
2	X		X	X		
3	X	X	X	X	X	
4					X	X
5	X			X	X	
6	X	X	X	X	X	
7	X	X	X	X	X	
8	X	X	X	X		
9				X		
10			X	X	X	X

Table 38: DIF analysis results, maths P3

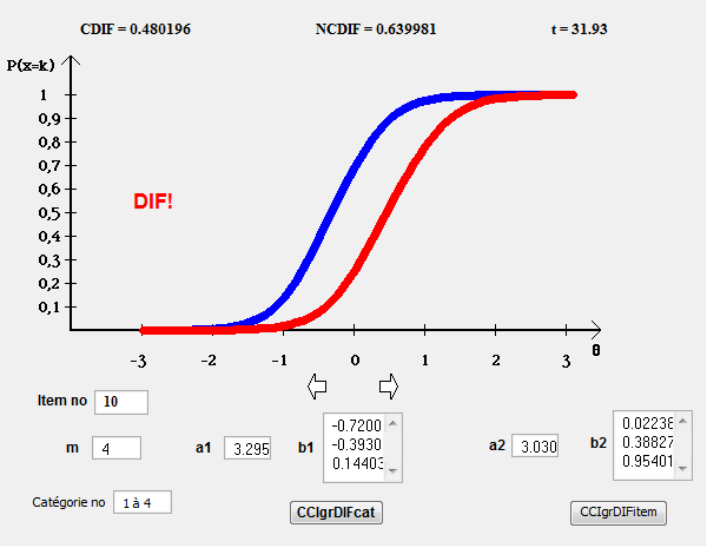
Testlets	Acoli	Ateso	Lango	Luganda	Rukiga	Runyankole
1.1.3	X		X		X	
1.4.5					X	
2.1.2	X		X	X		
2.3.4.7	X	X	X	X	X	
2.5.6		X	X	X		
3.1.2				X	X	
3.3.4		X	X			
3.5.7	X	X		X	X	X
4.1.3				X	X	X
5.1.4		X				
6.1.3	X	X	X	X		X
7.1.3	X	X	X		X	X
8.1.4	X	X		X		X
8.5.7			X			
9.1.3	X	X	X		X	
10.1.2	X		X	X		X
10.3.4	X	X	X	X		

As noted above, DIF analyses were conducted with GAUSS-IRT software, which produces not only statistics but also graphic representations of item and testlet behavior. The following are examples of these representations; note that all curves were produced using Samejima's graded model.

[14] A DIF analysis of the P3 language testlets revealed that languages had varying degrees of language bias problems. For example, only three testlets were found to be DIF in Runyankole whereas 8 testlets were DIF in Luganda. There is no standardized rule for determining when DIF items or testlets should be left in or removed from the analysis; however, as was done in the previous MLA (2008 Report), the rule was to exclude a testlet when it was found DIF across *all* 6 local languages: as reported earlier, in 2009, no testlet was found DIF for all 6 local languages.

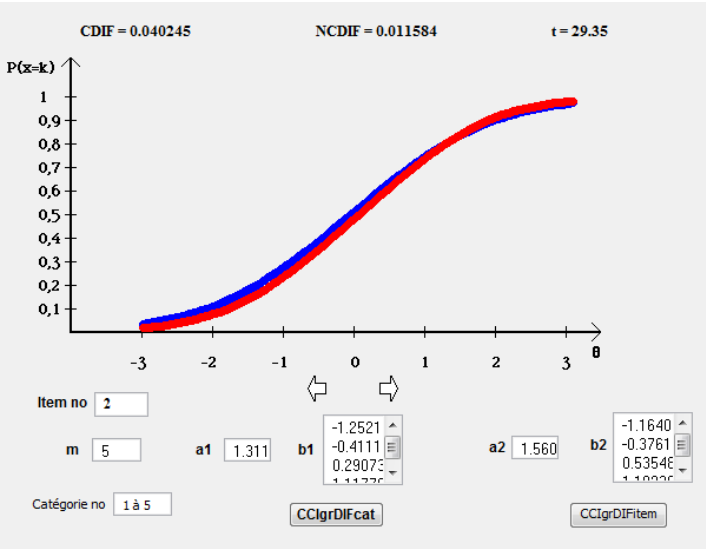
Figure 18: P3 graphs

1A. Example of "DIF" testlet



In order to make decisions about which items to retain for the MLA 2009 analysis, DIF was measured with groups of items, or testlets. Pictured at left is testlet e10t which consisted of 3 items, which has m=4 possible values: 0, 1, 2 and 3. A score greater than .054 for 4 values (or categories) in a testlet indicates DIF – i.e., that significant performance differences were observed in pupils of the same ability taking the test in Rukiga and English. The NCDIF value is a measure of the area between the Rukiga and the English curves. In this instance, the observed value of the DIF index, called here NCDIF at the top of the figure, is shown to be .639981. This value is in fact proportional to the area between the Rukiga and the English curves. This testlet is considered to be DIF! However, this testlet was not omitted from the analysis since the same level of DIF was not detected in all 6 local languages.

1B. Example of testlet that is not "DIF:



There are here m=5 categories associated with testlet e2t (pictured left) since this testlet is made of 4 items. However, since the value of NCDIF, .011584, is lower than .096, this testlet is not considered DIF. Note also that the area between the Rukiga curve and the English curve is much smaller than the corresponding area shown for testlet e10t above.

Threats to validity

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

1. **Multiple measures:** The 2009 MLA represents the third 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 third year of testing than it might have been in Years 1 and 2.
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 2009 MLA, some elevated missing data values were observed, especially return rates of interview instruments in certain districts and low percentages of pupils followed in the panel design (approximately 20% of 2007 pupils were tracked through 2009). 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: Analysis of results from cohort 1

As shown in Table 35, the subsample of 420 pupils in the panel is too small to be representative of the sample of 2,325 pupils of the initial 2007 cohort. In fact, scores of the 420 pupils in the panel were significantly higher than those of their counterparts. This constitutes a major drawback for our panel analysis and should be considered a serious limit for the interpretation of the subsequent results.

Table 35: Mean language and maths scores for the global cohort (2325 pupils) and the subsample (420 pupils) in P2

Results of the subsample

	N	Minimum	Maximum	Mean	Std. Deviation
LANGUAGE	420	2	40	26.02	10.322
MATHS	420	0	40	24.60	8.744
Valid N (listwise)	420				

Results of the global sample

	N	Minimum	Maximum	Mean	Std. Deviation
LANGUAGE	2325	0	40	19.98	11.456
MATHS	2325	0	40	20.09	9.692
Valid N (listwise)	2325				

As was done in the comparison of P3 2008 and P2 2007 above, all "true" achievement scores for 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 36 and Table 37 show the "true percent scores" in language and maths for the 420 pupils while they were in P2 (2007), P3 (2008) and P4 (2009). The most obvious finding is that the pupils in the control schools got much higher language and maths mean true scores than the pupils in the experimental schools in all 3 years – a result to be expected since most control schools were private schools. We

can also note a general increase in mean scores from P2 2007 to P3 2008 and then to P4 2009, with the exception of language from P2 in 2007 to P3 in 2008.

Table 36: Mean language true scores for the pupils of the control and the experimental schools in P2 2007, P3 2008 and P4 2009

Descriptive Statistics

		Mean	Std. Deviation	N
True language P2 2007	Control	90.7838	11.50246	148
	Experimental	70.6213	26.85412	272
	Total	77.7262	24.61365	420
True language P3 2008	Control	90.5203	15.65377	148
	Experimental	70.4743	29.62308	272
	Total	77.5381	27.30305	420
True language P4 2009	Control	96.8986	6.58150	148
	Experimental	84.5993	20.78632	272
	Total	88.9333	18.14540	420

Table 37: Mean maths true scores for the pupils of the control and the experimental schools in P2 2007, P3 2008 and P4 2009

Descriptive Statistics

	Contr_exp	Mean	Std. Deviation	N
True maths P2 2007	Control	74.4189	14.17145	148
	Experimental	64.1581	19.86692	272
	Total	67.7738	18.70354	420
True maths P3 2008	Control	84.3378	11.51134	148
	Experimental	72.7500	17.75053	272
	Total	76.8333	16.76286	420
True maths P4 2009	Control	91.0338	10.59227	148
	Experimental	81.5956	16.10179	272
	Total	84.9214	15.08072	420

Figure 17 and Figure 18 below summarize the results shown above in a graphical way. Note that, for the language true scores (Figure 17), the two lines, the first one

for the pupils of the control schools and the second one for the pupils of the experimental schools, are not parallel as they generally do for the maths true scores (Figure 18).

Figure 17 Plot of the mean equated/linked language true scores for pupils in the control and the experimental schools in P2 2007, P3 2008 and P4 2009

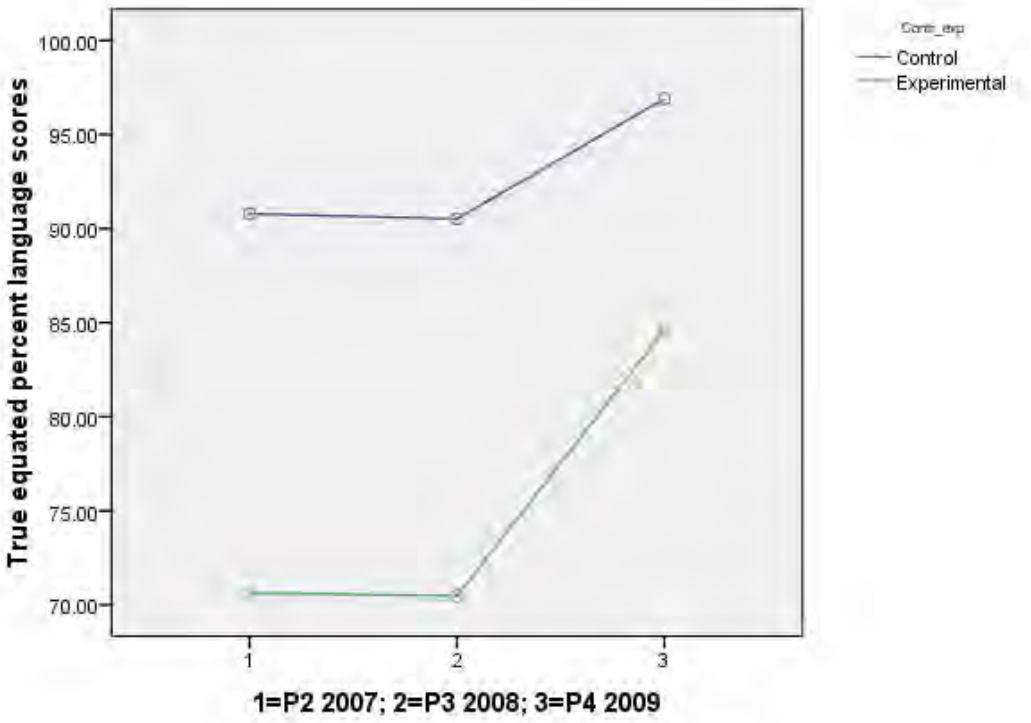
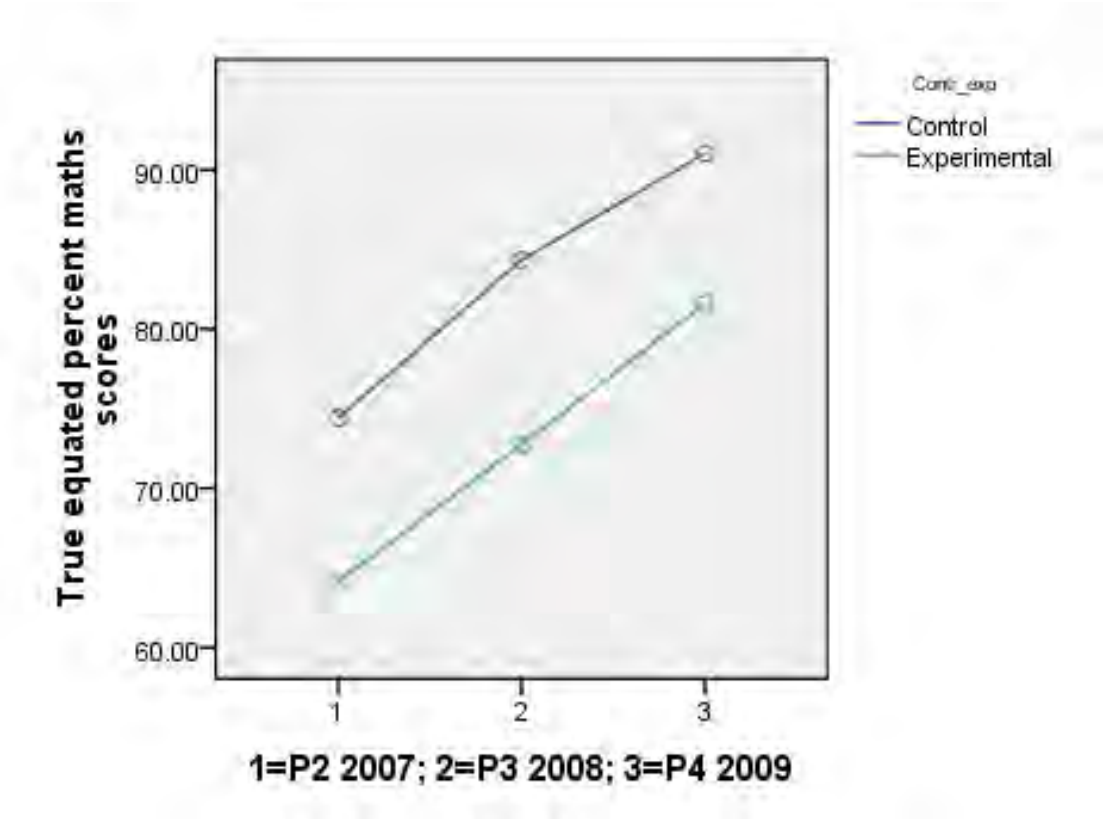


Figure 18 Plot of the mean equated/linked maths true scores for pupils in the control and the experimental schools in P2 2007, P3 2008 and P4 2009


Cont_exp
Control
Experimental
True equated percent maths scores
90.00
80.00
70.00
60.00
1
2
3
1=P2 2007; 2=P3 2008; 3=P4 2009

References

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

Bertrand, R. & 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.

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.

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

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

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

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

Wainer, H., Bradlow, E.T., & Wang, X. (2007). Testlet Response Theory and its Applications. New York: Cambridge University Press.