Ghana 2013 National Education Assessment

Technical Report

May 2014 (Final Version)

Ministry of Education
Ghana Education Service
National Education Assessment Unit

Ghana 2013 National Education Assessment

Technical Report

May 2014 (Final Version)

This report was made possible by the support of the American people through the United States Agency for International Development (USAID). The contents of this report are the responsibility of RTI International and do not necessarily reflect the views of USAID or the United States Government.

Prepared under EdData II Technical and Managerial Assistance, Task Order No. 21, Contract Number AID-641-BC-13-00001, Strategic Objective 3. Submitted to Sarah Banashek, USAID/Ghana.

Prepared by Pierre Varly, Chris Cummiskey, Tracy Kline, and Liz Randolph, RTI International, 3040 Cornwallis Road, Post Office Box 12194, Research Triangle Park, North Carolina, 27709-2194, USA. RTI International is a trade name of Research Triangle Institute.

Table of Contents

Annexes

Figures

Tables

Acknowledgements

The authors wish to acknowledge the staff of the National Education Assessment Unit (NEAU; formerly Assessment Services Unit [ASU]) of Ghana for its leadership in the 2013 administration of the National Education Assessment (NEA), including the preliminary planning, enhancements to the NEA design, and the extended and improved test items making up the four instruments for 2013. We would also like to acknowledge the education staff from USAID/Ghana, the Ghana Education Service (GES), and current and former faculty from the University of Cape Coast and the University of Education, Winneba, for their individual and collective contributions to the enhanced design, training, and administration of the 2013 NEA.

Crucial support throughout and at key points along the way was provided by: Emmanuel Sam-Bossman, Ghana *Testing* Field Manager; Megan McCune, Ghana *Testing* Project Coordinator; Emily Kochetkova, Ghana *Testing* Home Office Technical Manager; Michael Fast, Assessment Specialist and Psychometrician; Lisa Thalji, Senior Survey Methodologist; Jamie Freedman; Survey Methodologist; Elizabeth Kinyara, Survey Methodologist; and Erin Newton, Senior Editor.

We are most grateful to the school administrators and pupils across the country who consented to participate in this study, and thereby contributed important knowledge to the citizens of Ghana; non-government organisations; and the Ministry of Education and the Ghana Education Service, all seeking to improve the quality of education and pupil learning across the country. In particular, we recognise all those in Ghana who welcomed the information, contributed their time and energy at the 12 District Cluster Forums (61 districts), and brought ideas to the National Policy Forum that resulted in the final recommendations which appear at the end of this report and offer Ghana a path forward.

Abbreviations

ASU	Assessment Services Unit (see also NEAU)
BECE	Basic Education Certification Examination
CI	confidence interval
CRDD	Curriculum Research Development Division
DAT	District Advocacy Team
EdData II	Education Data for Decision Making project
EGMA	Early Grade Mathematics Assessment
EGRA	Early Grade Reading Assessment
EMIS	education management information system
FCUBE	Free and Compulsory Universal Basic Education
G2G	Government-to-Government
GDP	gross domestic product
GER	gross enrolment rate
GES	Ghana Education Service
GETFund	Ghana Education Trust Fund
GH¢	Ghanaian cedi
IEA	International Association for the Evaluation of Educational Achievement
IGF	internally generated funds
JHS	junior high school
KR20	Kuder-Richardson-20 reliability coefficient
LCA	latent class analysis
MC35	minimum competency score, 35%
MOE	Ministry of Education
NALAP	National Literacy Acceleration Programme
NAR	net attendance ratio
NEA	National Education Assessment
NEAU	National Education Assessment Unit (new name for ASU as of September 6, 2013)
P2, P3, P4, P5, P6	primary grades 2–6
PF55	proficiency score, 55%
PIRLS	Progress in International Reading Literacy Study
PPP	purchasing power parity
QC	quality control
RTI	RTI International (trade name of Research Triangle Institute)
SD	standard deviation
TIMSS	Trends in International Mathematics and Science Study
UIS	UNESCO Institute for Statistics
UNESCO	United Nations Educational, Scientific, and Cultural Organization
UNICEF	United Nations Children's Fund
USAID	United States Agency for International Development

Executive Summary

This report presents the findings from the 2013 administration of the Ghana National Education Assessment (NEA), carried out by the Assessment Services Unit (ASU) within the Ghana Education Service (GES). In addition to the performance results, this volume offers a substantial amount of contextual background and technical detail regarding the methodology for the 2013 NEA test development, sampling, data collection, and data analysis. Available separately is a brief document titled *Ghana 2013 National Education Assessment: Summary of Results,* which focuses more narrowly on the findings. The summary version is intended for use in discussions of policies and recommendations around instruction and educational assessments in Ghana.

The NEA is a biennial nationally and regionally representative measure of student competency in mathematics and English in primary grades 3 and 6 (P3 and P6). The 2013 NEA was the fifth application of the NEA and it covered all 10 regions of Ghana, sampling 550 schools and testing 36,905 students over the course of three days in July 2013. The sample size was selected to provide precision data at the national and regional levels, but not at the district level.

The 2013 NEA tests were based on national curricula and were made up of 30 to 40 multiple-choice questions, administered to students in a group, in their classrooms. Test items covered skills and knowledge across the following domains:

English	**Mathematics**
Listening	Basic Operations
Reading Comprehension	Numbers and Numerals
Usage (Grammatical Structure)	Measurement / Shape and Space
	Collect and Handle Data

For the past two administrations (2011 and 2013), the United States Agency for International Development (USAID) has sponsored technical assistance to the ASU through the Education Data for Decision Making (EdData II) project. A part of this assistance served to enhance the test content, to improve comparability for detecting any historical trends, and to ensure that individual items within the test would measure the intended cognitive skills with reliability and an appropriate range of difficulty. With the enhancements, for the first time, the 2013 data analysts were able to make meaningful comparisons of student outcomes between some aspects of the current and a previous test.

In addition, because the NEA has consistently revealed that children in both grades were struggling to complete the English test and to perform grade-appropriate or even foundational mathematics, in 2013, a few items were added to the test specifically to help assess children's reading competency.

Test Results

All four subject tests (P3 mathematics, P3 English, P6 mathematics, and P6 English) used the same test score cut-points to indicate that a student had achieved the *minimum competency level* and the *proficiency level*. Students who scored 35% correct were defined as having reached minimum competency and students scoring 55% or better were defined as having reached proficiency.

Table ES1 provides information on student performance according to the two nationally defined cut-points. It includes the proportion of pupils who failed to achieve minimum competency. Less than 25% of the students met the proficiency cut-point in P3 mathematics and less than 30% of the pupils achieved proficiency in P3 English. A larger proportion of pupils (39%) reached proficiency in P6 English. A range of students from 30% to 40% failed to achieve even minimum competency levels in the subjects tested; that is, they failed to answer even 35% of the items correctly on a particular test.

Table ES1: Percentage of students meeting criteria for minimum competency and proficiency, by subject and grade

Competency level	Grade and subject tested			
	P3		P6	
	Maths	English	Maths	English
Below minimum competency	42.9%	41.9%	39.2%	31.3%
Minimum competency	35.0%	29.7%	50.0%	29.8%
Proficiency	22.1%	28.4%	10.9%	39.0%
Total	100.0%	100.0%	100.0%	100.0%

Performance Trends from 2011 to 2013

It can be seen in ***Table ES2*** that the average (percent correct) scores differed very little from 2011 to 2013. These small shifts in performance were not statistically significant.

Table ES2: Mean percent correct scores, 2011 and 2013 administrations

Subject	Year	Class	
		P3	P6
Maths			
	2011	38.6	39.5
	2013	41.1	38.2
English			
	2011	44.0	49.8
	2013	44.4	48.9

Performance According to Subject Domains

Analyses of student performance across the various subject domains within English and mathematics revealed some noteworthy patterns. In English, the reading domain presented the greatest challenge to students, in both P3 and P6. The fact that students were challenged by tasks involving reading was confirmed by an additional analysis that identified students who were likely to be non-readers. Based on this analysis (2013 data only), approximately 44% of P3 students and 35% of P6 students were considered likely to be non-readers.

In mathematics, both P3 and P6 students had difficulty with the higher-order cognitive tasks involving measurement of shapes and space. P3 students also did not do well on the tasks involving the collection and handling of data.

Performance According to Cognitive Skills

Improvements were made to the 2013 NEA to better balance the test items across various cognitive skill areas. The changes included items which tapped student ability on higher-order cognitive skills, including critical thinking in English and mathematical reasoning. In mathematics, neither P3 nor P6 students performed well on tasks that required applying mathematical knowledge and reasoning—that is, tasks that tapped cognitive skills beyond basic mathematics knowledge and understanding. On average, P3 and P6 students scored less than 40% correct on these types of maths questions. In English, P3 students had difficulty with tasks requiring basic understanding of English narrative text and with tasks requiring critical thinking; on average, they scored less than 40% correct. P6 students also performed poorly on the English tasks involving critical thinking.

Performance Across Sub-Populations

The NEA data also were examined by several sub-groups (see ***Table ES6***). For example, learning outcomes for males and females, in both P3 and P6, were similar. Small differences were observed on learning outcomes between males and females. Only the P6 results were significant, however, with a higher proportion of P6 males achieving proficiency than females, for both English and maths. On the other hand, the disparities in learning outcomes based on the location of the school (urban versus rural) and the type of school (public versus private) were substantial. Learning outcomes for students in urban areas were significantly higher than for students residing in rural areas, with the proportion of students reaching 'proficiency' levels in the urban schools twice that of the proportion of students reaching proficiency in the rural schools. Students residing in rural areas were much more likely to fall below the minimum competency cut-point of 35% correct than students residing in urban areas. The disparities were similar for students residing in deprived versus non-deprived districts. Not surprisingly, the performance of students residing in the three regions of northern Ghana (Northern, Upper East, Upper West)—where the majority of students sampled were residing in a deprived district—was substantially below that of students from all other regions of the country. With the exception of the Western Region, the proportion of students reaching proficiency in the non-northern regions of the country was double (or more) the proportion of students reaching proficiency in the three northern regions of the country.

Table ES6: Student performance by gender, location and type of school

Subject and competency level	Percentage of children performing at competency levels					
	Gender		Location		Type of school	
	Male	Female	Urban	Rural	Public	Private
P3 maths						
Below minimum competency	42.2	43.6	33.0	48.7	49.1	19.5
Minimum competency	36.0	33.9	34.0	34.8	35.7	32.3

Subject and competency level	Percentage of children performing at competency levels					
	Gender		Location		Type of school	
	Male	Female	Urban	Rural	Public	Private
Proficiency	21.8	22.5	33.0	16.5	15.3	48.2
P3 English						
Below minimum competency	42.2	41.4	29.6	49.0	49.0	14.9
Minimum competency	30.8	28.6	26.7	30.6	31.9	21.7
Proficiency	27.0	29.9	43.7	20.5	19.2	63.4
P6 maths						
Below minimum competency	37.3	41.0	27.7	46.1	44.5	18.9
Minimum competency	51.1	48.8	54.5	46.6	48.0	57.6
Proficiency	11.6	10.2	17.7	7.3	7.6	23.5
P6 English						
Below minimum competency	30.7	31.9	17.8	39.4	37.2	8.8
Minimum competency	30.9	28.5	23.1	33.6	33.1	17.0
Proficiency	38.4	39.6	59.2	27.1	29.7	74.2

With the exception of P6 English, the proportion of students attending private schools who reached the cut-point for 'proficiency' was three times the proportion reaching proficiency among students attending public schools. (The proportions were closer for P6 English, yet still more than double for private school students.) Children from rural areas depend much more on public education than in other parts of Ghana, especially in the northern regions, with only 4.3% of the enrolments in the three northern districts in private schools compared to the national average of 20% and 29% for the urban capitals of Ashanti and Greater Accra (see Section 1.2). Public schools in the northern regions and deprived areas were less likely to have qualified teachers, access to materials, or minimal physical infrastructure, and thus the higher dependency on public schools in these poor and hard-to-reach regions combined with inequities in public school inputs may account to a large degree for this pronounced gap between public and private school learning outcomes.

Recommendations

During the National Policy Forum in Ghana in February 2014, and a series of District Cluster Forums leading up to it, overall nearly 3,000 attendees thoroughly discussed the results and implications of the 2013 NEA and the 2013 EGRA/EGMA. The participants agreed on a set of specific policy and action recommendations in several areas. These areas were: changes in instructional methods, in both reading and mathematics; better methods and practices for training and coaching teachers, as well as placing them strategically by language, grade, and region for the greatest impact on pupil achievement; ensuring textbooks and supplementary materials for every pupil; and strongly encouraging parent and community involvement in children's learning.

Section 4 presents the group recommendations in full.

1. Background: Purpose and Objectives of the National Education Assessment

This report reviews the analyses and findings from the 2013 administration of the Ghana National Education Assessment (NEA), which was carried out by the Assessment Services Unit (ASU)[1] within the Ghana Education Service (GES). In addition, the report reviews a number of enhancements to the NEA that took place before the 2013 NEA was administered. The enhancements were based on recommendations from the 2011 NEA application. For both these efforts, the United States Agency for International Development (USAID) sponsored technical assistance to the ASU through task orders under the Education Data for Decision Making (EdData II) project, led by RTI International. Technical assistance in support of the 2013 NEA was provided for under the USAID/Ghana Partnership for Education: *Testing* activity.[2]

This was the fifth round of the biennial NEA. The NEA is a nationally and regionally representative measure of student competency in mathematics and English in grades 3 and 6 (known as Primary 3 [P3] and Primary 6 [P6]). The 2013 NEA (conducted in July 2013) covered all 10 regions of Ghana, sampling 550 schools and 36,905 students in the course of three days. As part of the analysis prepared for this report, the results were disaggregated on the following: gender, location (urban/rural), type of school (public vs. private), and whether schools were within a deprived district[3] or not.

The report is organised as follows. The remainder of Section 1 presents background and contextual information on education in Ghana of particular relevance to the NEA. Section 2 describes the methodology for the 2013 implementation of the NEA, including the rationale for and description of a number of improvements on the 2011 NEA that were made through a series of test development workshops that took place in March and April 2013. Section 3 presents the findings. Section 4 consists of overall recommendations informed by the District Cluster Forums that took place November 2013–January 2014, and especially the National Policy Forum in February 2014. Finally, the report concludes with several technical annexes that support material presented in the main text.

1.1 Education Expenditure

In the past decade, education has accounted for 18–27% of public expenditure, or approximately 5–6% of Ghana's gross domestic product (GDP). [4] The sector has seen steady growth from 5.6% of the GDP in 2003 to 6.3% in 2011, which is above the average for all African countries combined. Total government spending on education tripled from 2003 to 2011 (i.e., from 0.53 million Ghanaian cedi [GH¢] to 1.7 million), a trend that also occurred

[1] Although this unit was renamed 'National Education Assessment Unit' (NEAU) not long after the test administration in 2013, for chronological clarity, this report retains the name that was in effect when the preparations and data collection effort were under way.
[2] The *USAID Partnership for Education: Testing* is one of five interconnected components (*Learning, Testing, Evaluation, Funding, and Government to Government [G2G]*) of a partnership among USAID, the Ministry of Education (MOE), and the Ghana Education Service, called the *USAID Partnership for Education Program*.
[3] See Section 3.6.4, within the discussion of 2013 NEA results, for a definition of 'deprived'.
[4] Darvas, P., & Balwanz, D. (2013). *Basic education beyond Millennium Development Goals in Ghana.* Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/16307.

in other government spending over this period. In 2011, 25.8% of all public expenditure was for education, 34.6% of which was allocated to primary education (grades 1–6) and 49% for the full basic education (kindergarten through junior high school). Non-government sources of education finance in Ghana also increased from 2003 to 2011, including the Ghana Education Trust Fund (GETFund), internally generated funds (IGF), and donor and other external financing, accounting for approximately 28% of all funding to the sector in 2011.[5]

Recent data showed salaries and other personnel costs (e.g., travel, allowances) accounting for the majority of government expenditure on education, approximately 95.6%, and approximately 68.7% of all expenditure on education (i.e., including external sources). Almost the entire public allocation to primary school went to salaries (e.g., 99.4%). Furthermore, personnel costs have tended to run over budget (e.g., in 2011, 172%), which in turn squeezes the expenditures for services, which tend to run under budget. Taken together, in 2011 data, only 2.3% of public expenditure on education went to administration and services (i.e., 1.6% and 0.7%, respectively). This has implications for the sustainability of primary school national assessment programs, which involve considerable service and administrative costs. Currently, national assessment programs in primary education (namely the NEA and the Early Grade Reading and Mathematics Assessments [EGRA and EGMA][6]) depend almost entirely on donor funding. Per-pupil recurrent expenditure for primary school students (approximately US$413, or 16.5% of per capita GDP)[7] was above that of most sub-Saharan Africa countries when measured in 2008. Even so, only a fraction of this per-student recurrent expenditure was allocated to services.

In general, for primary schools, teacher deployment and resources flow from a central base. Primary schools have very little budget of their own and essentially no financial autonomy, despite ongoing reforms and decentralization. Districts rarely receive their full budget allocation for goods and services (e.g., approximately 30% of the GES budgeted allocation in 2012) and thus often depend on non-government funding for day-to-day management. This has implications for how information from assessment can be used effectively at subnational levels to address performance gaps, and it underscores the importance of providing a broad stakeholder reach in disseminating findings, as well as promoting local ownership and shared accountability in school performance and learning outcomes.

Even though salaries account for most allocation to primary schools, there are still budget inequities across districts. This speaks to the inequitable distribution of qualified and experienced teachers in remote and otherwise deprived areas.[8] On average, teachers in Accra have four more years of experience than those in the rest of the country. Teachers in deprived

[5] Ministry of Education, Republic of Ghana. (2012). *Education sector performance report, 2012* [ESPR]. Accra. http://www.ndpc.gov.gh/GPRS/Dist%20and%20Sec%20APR%202012/SECTOR%202012%20APRs/Ministry%20of%20Education.pdf

[6] The *Testing* activity also helped carry out an EGRA and EGMA in 2013 of nearly 8,000 pupils, in 12 Ghanaian languages and English, to gain a deeper understanding of students' skill levels and what impact they might be having on NEA results. The EGRA/EGMA findings are summarised in a separate report.

[7] Unit expenditure in purchasing power parity, or PPPs, based on 2008 data. From Motivans, A. (2010). *Education investment and commitment: Reassessing the international benchmarks.* Stockholm: United Nations Educational, Cultural and Scientific Organization, UNESCO Institute for Statistics (UIS).

[8] Ministry of Education, Republic of Ghana. (2013).*MOE education sector performance report 2008–2012.* Based on data from MOE education management information system (EMIS). Accra.

districts have two years' less experience than teachers in non-deprived areas, and receive lower pay.

1.2 Equitable Access, Retention, and Quality in Ghana's Primary Schools

Like other countries in sub-Saharan Africa and elsewhere, increased enrolment in primary school has introduced substantial supply-side barriers to learning. Examples include shortages of qualified teachers and support services, especially in remote areas; inadequate materials; and difficulties maintaining an effective and modern curriculum leading to acceptable learning outcomes. In addition, in the past 20 years, Ghana has seen population increases of approximately 70%, with most of the growth in the urban regions. The incidence of extreme poverty has been cut in half as a result of economic growth—partially attributable to new oil reserves—but disparities remain. Urban populations are more advantaged than rural ones in terms of access, retention, and learning outcomes. The differences between deprived and non-deprived regions are particularly stark.

Enrolments in primary education almost doubled in the past 15 years following the introduction of Free and Compulsory Universal Basic Education, or 'FCUBE' (e.g., 2.5 million in 1999/2000 school year to 4.45 million in 2011/2012), with a 40% estimated increase in net enrolment rate in the past decade (e.g., 58% in 2003/2004 to 82% in 2011/2012). Enrolment gains have been made across Ghana, even in some of the most impoverished and remote regions of the country, such as Upper East and Upper West regions.

In spite of overall gains in enrolment, irregular attendance and late entry into primary school remain serious problems for children from impoverished homes and rural settings. Data from 2010 indicated that in rural areas, approximately 60% of 6-year-old children and 45% of 7-year-old children were not in school, while in urban areas, 43% of age 6 and 23% of age 7 children were not in school.[9] Entering school on time is important for learning and primary completion. Attendance (as measured by net attendance ratio [NAR]) was, on the average, only 73% in 2011, suggesting children were not in school over 25% of the time.[10] Furthermore, attendance and therefore time-on-task in primary school were much higher for students from more wealthy families (highest quintile: 85%) and urban locations (80%). By contrast, NAR estimates were 61% for pupils in the lowest wealth quintile and 69% for those living in rural locations.

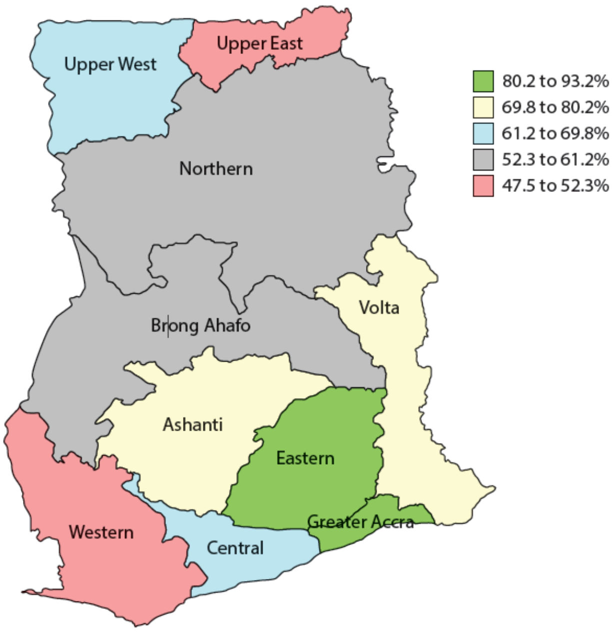
Children from rural areas, particularly in the north, depend heavily on public education. In the 2010/2011 school year, an average of 4.3% of enrolments in the three regions of the north combined were in private schools, compared to the national average of 20% and approximately 29% in the urban capitals of Ashanti and Greater Accra Region. The distribution of resources, particularly trained teachers, favours the urban and wealthier districts and thus plays an important factor in inequities observed in learning outcomes. In spite of increased expenditure in education in the past decade, inequities in education resources across urban/rural and poor/wealthy lines have exacerbated rather than attenuated.

---

[9] UNICEF (2010), as cited in Darvas & Balwanz (2013); see footnote 4.
[10] A large proportion of children ages 6 to 11 years attend kindergarten or pre-school in Ghana but are not counted as attending (primary) school. Thus, the estimate of children enrolled but not attending (over 25%) may be inflated.

***Figure 1*,** taken from the *Education Sector Performance Report, 2012* (see footnote 5), based on 2011/2012 education management information system (EMIS) data, demonstrates this effect for trained primary teachers. Note that the lowest percentages of trained teachers were in the Western and Upper East Regions (48–52%), and the highest percentages in Greater Accra and Eastern Region (80–92%).

Figure 1: Percentage of trained primary teachers, by region, 2011/2012 school year



Source: ESPR 2012.

Source: Ministry of Education, Republic of Ghana. (2012). *Education sector performance report, 2012* [ESPR]. Accra. p. 18. http://www.ndpc.gov.gh/GPRS/Dist%20and%20Sec%20APR%202012/SECTOR%202012%20APRs/Ministry%20of%20Education.pdf

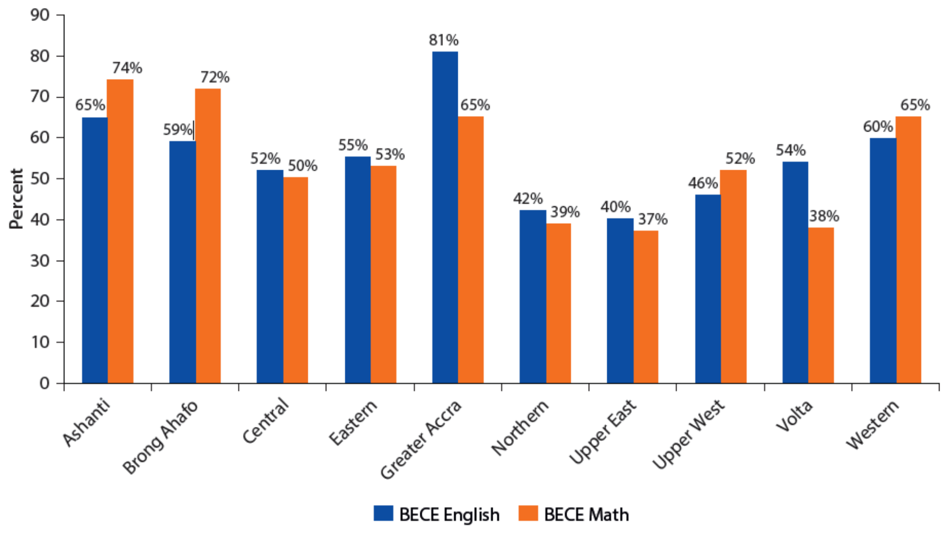
The quantity of untrained teachers was not the only relevant factor in regions that were poor, rural and/or in the north. In addition, classrooms were more crowded in schools in the poorest districts, with pupil–teacher ratios in the two highest wealth quintiles averaging 63 and in the two lowest wealth quintiles, 117 pupils per teacher (126.6 in the poorest wealth quintile, or twice as many as the average for the two highest [richest] quintiles). According to a 2011 World Bank study,[11] efficiency in teacher allocation decreased between the 2005/2006 and 2011/2012 school years. As compared to most other sub-Saharan African countries, a much larger proportion of Ghana's teachers was allocated at random (i.e., as opposed to teacher allocation being based on student enrolment), with 56% of the teacher allocation decisions not explained by enrolment.

Non-teacher inputs also fared worse in the poorest regions, especially in the deprived districts. Schools in the deprived districts had fewer primary textbooks, classrooms, potable water sources, and toilets. Findings from previous NEA administrations noted the disparities in learning outcomes, consistently showing poorer performance for pupils in the rural as

[11] As cited in Darvas, P., & Balwanz, D. (2013). *Basic education beyond Millennium Development Goals in Ghana*. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/16307.

opposed to urban regions, especially in the north. Finally, Basic Education Certificate Examination (BECE) pass rates also have underscored the poor performance of students in the rural and impoverished regions of the country. ***Figure 2*** presents the BECE English and mathematics pass rates by region, based on 2010/2011 EMIS data.

Figure 2: BECE English and mathematics pass rates, by region, 2010/2011

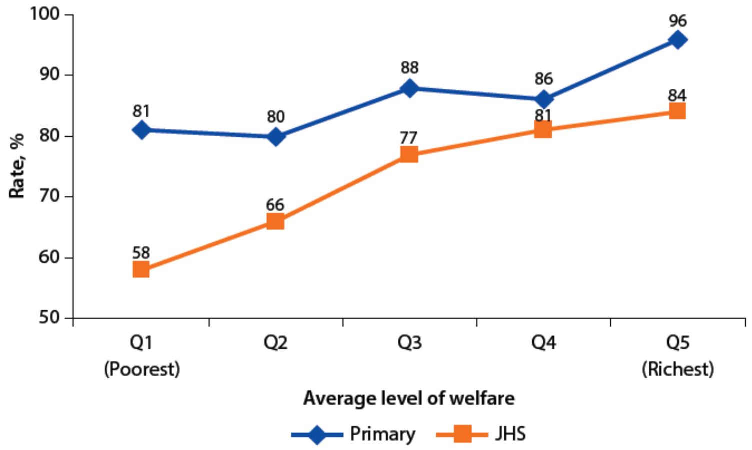


Source: EMIS 2011.

Thus, late age of enrolment, higher numbers of unqualified teachers, crowded classrooms, poorer infrastructure and fewer materials all contribute to poorer learning outcomes and in turn contribute to lower completion rates among the poorest regions of the country. Districts in the bottom wealth quintile (all located in the three northern regions) had completion rates in primary school of 81% compared to the highest wealth quintile at 96%.[12] The gap for junior high school (JHS) completion is even greater, with completion rates of only 58% in the lowest wealth quintile compared to 84% for the highest (see ***Figure 3***).

[12] World Bank. (2010). *Education in Ghana: Improving equity, efficiency and accountability of education service delivery.* Africa education country status report. Washington DC. http://documents.worldbank.org/curated/en/2010/02/17932091/education-ghana-improving-equity-efficiency-accountability-education-service-delivery

Figure 3: Completion rates according to wealth quintile*



*Source: World Bank 2010; estimates based on GLSS 2005/06 and EMIS 2008/09.

1.3 Relevance of Background to the 2013 NEA

As mentioned above, as measured at the national level, Ghana's expenditures on primary education are allocated almost entirely to personnel costs, with only 2.3% of funding on administration (1.6%) and services (0.7%). Even then, districts rarely receive their full budget allocation for goods and services (e.g., districts received only approximately 30% of the GES budgeted allocation for goods and services in 2012). Thus, the sustainability of Ghana's primary school national assessment programs is at risk, and until more funding can be allocated to services, will always depend on external funding. This situation also places at risk the degree to which districts themselves can directly address gaps in primary school performance and begin to reverse the status of primary school learning outcomes in the country.

The 2011 NEA research team was able to link student results to certain school characteristics based on data available from the EMIS. They identified the following as significant factors in learning: availability of textbooks; proportion of female teachers; proportion of teachers with training; visits from circuit supervisors; schools keeping administrative records; and student transfer rates. The background data presented in Section 1.2 underscores the continued decline in equitable resource allocation to schools, marginalising schools in the more rural and impoverished regions, including (but not limited to) the deprived districts. The 2013 NEA revealed similar gaps in performance based on location and wealth index.

1.4 Background of the NEA and EMIS Data Collection

As described earlier, the National Education Assessment involves tests administered every two years to a sample of P3 and P6 students, in English and mathematics. Given recommendations from 2011, some modifications were made in the tests, including reducing the number of test forms for each subject, reducing the number of items in each form, enhancing the comparability of the test with previous tests, and extending the ability levels assessed. These modifications are discussed in more detail in Section 2, Methodology. The tests are based on national curricula and are made up of 30 to 40 multiple-choice questions. The questions are presented in two forms, ordered differently on each. New questions are developed and added each year, but the tests retain a few common or 'anchor' items over time.

Less than a third of primary school children reached proficiency levels (i.e., percent correct scores of 55% or better) in English or in mathematics according to the NEA results of 2005, 2007, 2009, and 2011.[13] In 2011, an investigation of the factors associated with stronger or weaker performance was conducted by linking results to available school data given by the EMIS.[14] The NEA team agreed that these factors likely would not change from 2011 to 2013 and, therefore, factors associated with learning were not further analysed for the 2013 NEA. It is anticipated that the review of school and student factors associated with high and low performance will be seriously considered for the 2016 NEA. Like the 2011 NEA, the 2013 NEA analyses included some limited analyses at the subdomain level.[15]

2. Methodology

2.1 Content of the NEA Test

2.1.1 Justification for Revisions to the NEA Test Content

RTI support to the ASU for test and item development in preparation for the NEA 2013 was agreed upon as a result of recommendations made by RTI following the NEA 2011 testing. The objective was to extend the scope of skills tested, including foundational literacy and numeracy as well as higher cognitive abilities such as critical thinking. To allow a more rigorous evaluation of trends over time and possible comparisons with other tests such as the EGRA, EGMA, and international studies, the 2013 NEA also included some items in common with the 2011 test (the anchor items).

The tests followed international best practices in its design. Specifically, these recommendations[16] can be summarised as follows:

[13] As addressed in Section 3 below (NEA 2013 Results), across the NEA administrations, according to the GES recommendations, students who have scored 35% or better have been defined as having reached *minimum competency;* students scoring 55% or better have been defined as having reached *proficiency.*
[14] However, no questionnaires were administered to pupils, teachers, or directors; instead, the analysts relied on administrative data. Moreover, no measurements were available for some factors, such as pupils' socioeconomic status.
[15] The NEA content is based on topics and subtopics—known as domains and subdomains—from within the Ghana national curriculum. Sample domains from the P3 mathematics curriculum are 'Numbers,' 'Operations,' and 'Shape and Space.'
[16] See also Chapter 6 of the 2011 NEA report, pp. 51–54: Ministry of Education, Ghana Education Service, Assessment Services Unit. (2012). *Ghana National Education Assessment: 2011 findings report.* Prepared under USAID EdData II Technical and Managerial Assistance Task Order 12, Contract Number AID-641-BC-11-00001. Research Triangle Park, North Carolina: RTI International. https://www.eddataglobal.org/countries/index.cfm?fuseaction=pubDetail&ID=376

1. Greater attention to item design procedures, to eliminate errors,[17] including inappropriate distractor design and random distractors, especially in maths.
2. Integration of quality control (QC) protocols to eliminate typographical and graphics-based inconsistencies and errors, as well as the errors listed under point 1.
3. A focus on development of stringent criteria to ensure test items would measure the desired ability construct (i.e., testing reading rather than general knowledge).
4. Special care in constructing item sets for testing reading comprehension (and listening comprehension) to ensure passage and item alignment, and independence of items within the set.
5. Construction of a consistent format for form design, instructions, question layout, and examples.
6. Appropriate distribution of items per measured standard and across targeted ability levels among students.
7. QC procedures that would resolve dependence and clueing problems that often appear in tests involving assembling a collection of items, such as a series of reading comprehension questions all based on the same text.
8. Appropriate coverage of a range of cognitive abilities, especially in the P6 tests.
9. Appropriate structure of tests to allow subtest analyses and reporting by domain.
10. Inclusion of anchor items from NEA 2011 as well as from international tests in NEA 2013 tests.

A workshop held in Koforidua in February 2013 helped put these recommendations into practice by producing new items, which were then field-tested on a small sample of students. Several major decisions were made during the workshop and after formal piloting of the tests in April:

- Shortening the tests in terms of number of test items (from 40 to 30 for the P3 instrument and from 60 to 40 for the P6 instrument). The NEA 2011 results suggested diminishing validity of test results in the later items of the tests.
- Shortening the duration of the tests[18] from one hour to 40 minutes to avoid fatigue (and to fit with international test duration at these grade levels).
- Assembling two forms[19] for each instrument (instead of four), thereby permitting alternate distribution of the two forms in rows of four students sitting next to each other.
- Dropping the writing domain from the test, as the 2011 multiple-choice items[20] were considered not to be valid measures of students' own production.
- Including anchor items to allow linking with 2011 tests.

[17] Examples of errors to avoid include multiple correct response options; no correct response options; mismatch between stem and options; poor choice of reading passages; multiple constructs tested in a single item; lack of relationship between questions and reading passage; and lack of balance in the construction of options.
[18] The ratio of test length to duration remained comparable between the 2011 and 2013 tests.
[19] The two forms have the same content but differ in terms of item order and order of distractors within an item.
[20] Included sentence punctuation, capitalisation, and discourse organisation.

- Including international items accepted by the International Association for the Evaluation of Educational Achievement (IEA).[21]

With regard to the introduction of open-ended (constructed response) items, the benefits of these types of questions are clearly recognised, for both mathematics and English-language tests. They can measure student production and higher cognitive abilities such as critical thinking, and they prevent random answers. Arguments against incorporating such items, however, were based on finance and level of effort. It was agreed that further study was needed of future budget changes that could make feasible open-ended items, which are the only way to test writing.

2.1.2 Domains Covered by the Test in 2011 and 2013

Changes in domain content from the 2011[22] to the 2013 test are summarised in ***Table 1.***

Table 1: 2011–2013 test content: Number of items by specific domain skills

Subject and domain	No. of P3 items		No. of P6 items	
	2011	**2013**	**2011**	**2013**
Maths				
Basic Operations	13	10	22	15
Collect and Handle Data	6	5	3	6
Measurement and Shape-Space	14	9	7	12
Numbers and Numerals	7	6	7	7
Total	**40**	**30**	**39**	**40**
English				
Grammar	10	10	21	16
Listening	10	8	15	8
Reading	13	12	14	16
Writing	7	0	10	
Total	**40**	**30**	**60**	**40**

The distribution by domain content in the test was not dramatically changed from 2011, especially for P3—although as noted above, the Writing domain was dropped. Each 2013 domain was covered by sufficient numbers of items to produce accurate pupil scores by domain. For P6, compared to 2011, more items were included for the 2013 core domains (i.e., Reading and Basic Operations) to be able to measure foundational skills with more accuracy.

[21] i.e., from the Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS).

[22] 2011 definition of the domains and cognitive abilities are in the NEA operations manual, pp. 6–9: Ministry of Education, Ghana Education Service, Assessment Services Unit. (2011). *Ghana National Education Assessment: Operations manual.* Prepared under USAID EdData II Technical and Managerial Assistance, Task Order 12, Contract Number AID-641-BC-11-00001. Research Triangle Park, North Carolina: RTI International. https://www.eddataglobal.org/countries/index.cfm?fuseaction=pubDetail&ID=381

2.1.3 Cognitive Skills Covered by the Test

The cognitive abilities covered by the test were distributed across four levels (according to Bloom's Taxonomy):

- Knowledge
- Understanding
- Application
- Reasoning/Critical Thinking

Summary definitions appear in ***Table 2*** for English and ***Table 3*** for maths.

Table 2: Definitions of cognitive processing type by domain: English language

Content Domain	Recall of Knowledge	Understanding of Concepts	Critical Thinking
Listening	– Answer appears word for word in text – All options contain language appropriate for targeted grade – To the extent possible, distractors are directly referenced in text	Answer is derived from multiple pieces of information in text	Answer is derived from various sources in the text or outside the text
Grammar	– Answer correctly uses 1-step rule	Answer correctly uses multi-step rule	Answer correctly uses multi-step rule
Reading	– Answer appears word for word in text – All options contain language appropriate for targeted grade – Distractors are directly referenced in text	Answer is derived from multiple pieces of information in text	Answer is derived from various sources in the text or outside the text

For the mathematics tests, three slightly different labels for cognitive processing types or levels were used. These did not differ in any substantive way from the levels agreed upon for English language but were more relevant to the domains measured in mathematics, as shown in Table 3.

Table 3: Definitions of cognitive processing type by domain: Mathematics

Content Domain	Knowledge and Understanding*	Application*	Reasoning**
Numbers and Numerals Basic Operations Collect and Handle Data Measurement Shape and Space	remember, recall, identify, define, describe, list, name, match, state principles, facts and concepts	summarise, translate, rewrite, paraphrase, give examples, generalise, estimate or predict consequences based upon a trend	capacity to apply logical, systematic thinking, including intuitive and inductive reasoning based on patterns and regularities that can be used to arrive at solutions to non-routine problems*

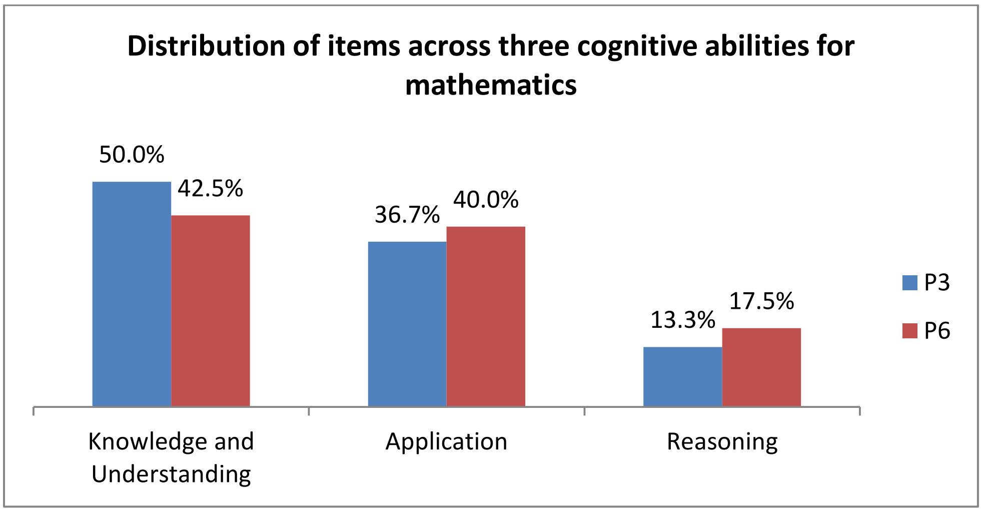
* Definition source: Ghana national curriculum and test framework.
** Definition source: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS) 2011 Framework.

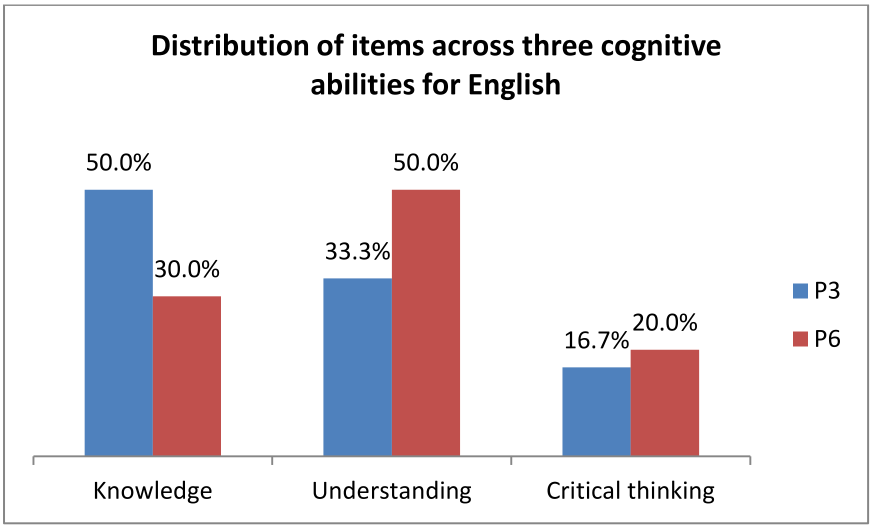
In 2011, Reasoning and Critical Thinking abilities were not tested according to the assessment framework. Instead, the tests were divided into Knowledge and Understanding items (70%) and Application items (30%) for P3 and 65% and 35% respectively for P6 (see the 2011 NEA operations manual, pp. 6–9 [citation in footnote 22]). In the 2011 NEA these distributions of cognitive abilities for P3 and P6 tests were not properly aligned with the curricula objectives.

In 2013, the test development team introduced significant changes in the structure of the tests with regard to cognitive abilities to address this issue, as follows.

For P3, 50% of the 2013 maths items and 83% of the English items were classified as assessing Knowledge and Understanding cognitive abilities. ***Figure 4*** shows the progression of the cognitive abilities assessed by the test from P3 to P6. Less than 20% of the test, for all grades and subjects, targeted higher cognitive abilities (such as critical thinking and mathematical reasoning). As said before, the key limitation of multiple-choice questions is that they do not allow testing of the highest cognitive skills, such as creation.

Figure 4: Distribution of items across three cognitive abilities, for mathematics and English, P3 and P6





The Ghana primary school curriculum focuses predominantly on narrative text and occasionally on descriptive; other types of texts, such as informative and persuasive, are not featured extensively in the curriculum. By extension, teachers tend to limit reading instruction to predominantly narrative with some descriptive texts. The NEA, a curriculum-based test, also reflects this focus, with English items featuring passages similar to what would be seen in the children's textbooks. Moreover, reading items in the test are presented as if they were texts in a reading book or textbook. The instruments do not test the capacity to retrieve information from nonliteral texts such as forms, tables, web pages, or tables of contents.

The NEA mathematics items are designed to ensure that literacy does not present a barrier to valid assessment of a child's basic numeracy skills. Thus, reading is not a requirement for completing the mathematics items that assess lower-order cognitive abilities. However, to test reasoning and application, word problems are necessary.

In mathematics, insofar as possible, the 2013 items avoided including text, to prevent incidentally testing skills other than the targeted maths learning outcomes.

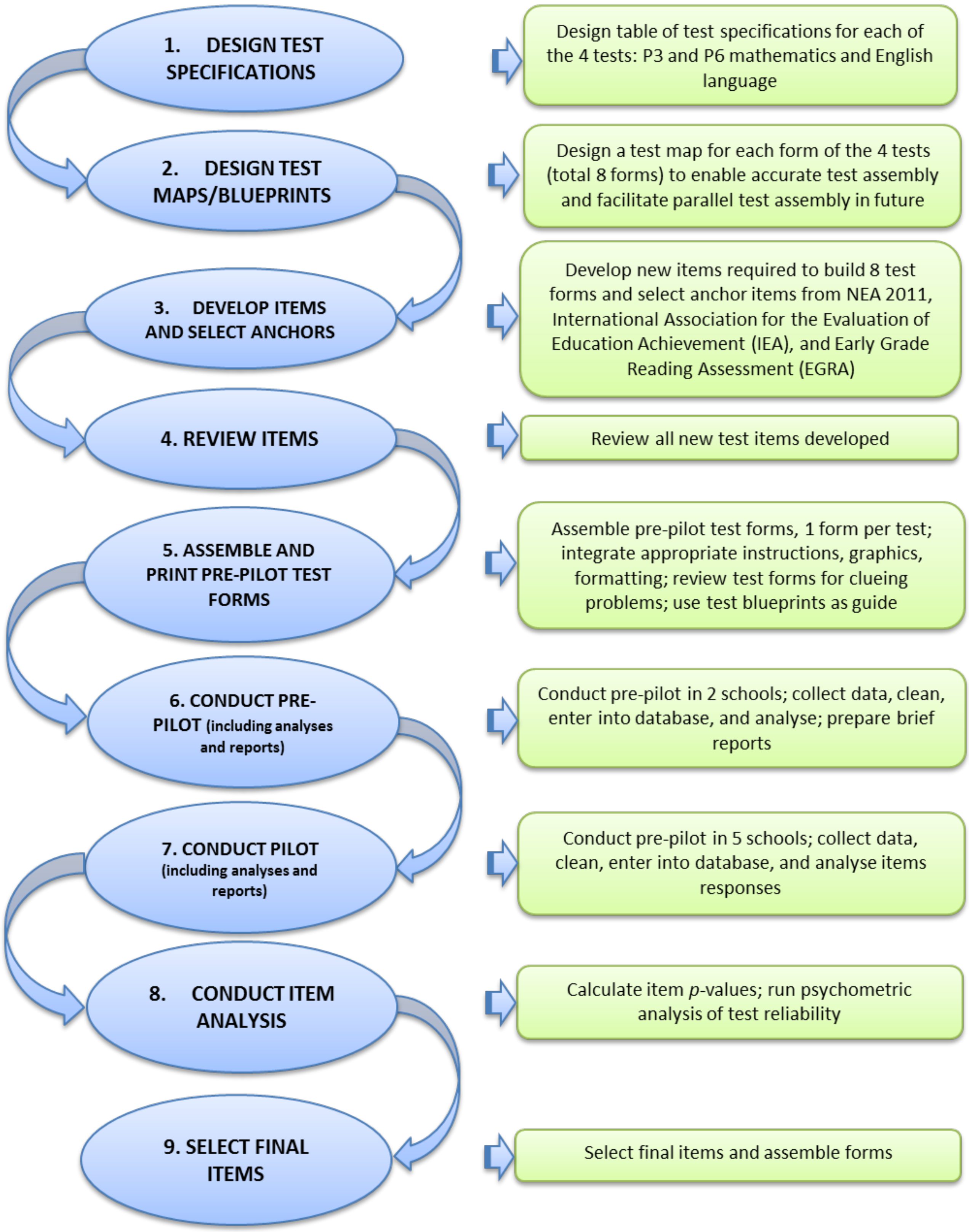
2.1.4 Procedures to Develop New Items

Based on current test theory, substantial improvements in the technical quality of test forms were achieved following the process in ***Figure 5.***

Emphasis was placed on several aspects of item development:

- Avoid testing the general knowledge domain in the English reading items
- Develop content for distractors and format similarly (same size, same length, same lexical level…)
- Include typical errors and plausible answers as maths distractors
- Simplify data in the word problems to test reasoning rather than basic operations
- Improve the format and design of the test
- Reduce and simplify text in mathematics word problems
- Harmonise the type of instructions
- Avoid use of text in basic mathematics operations as much as possible

Figure 5: Process of test development



2.1.5 Anchor Items

The design of the NEA 2013 tests introduced features that were not possible in the 2011 and earlier versions of the test instruments. Before the 2013 revision, it was not possible to compare scores either across forms within the same administration (four forms were administered in 2011 and these forms were not equated) or from one administration to the next.

To remedy this, linking strategies were applied in 2013 in order to facilitate the comparison of the NEA 2011 results with those of the NEA 2013. Maintaining these anchor items in successive years will enable trend analyses in the future, from 2011. The instruments were linked by including a subset of NEA 2011 items (to the extent possible, representing all of the subject domains) in the NEA 2013 instruments.

A second source of linking, specifically on the English-language test, was the inclusion of EGRA-like items. Test results from the NEA 2011 suggested that a number of students were unable to read, which also had potentially serious implications for performance on the mathematics tests, considering that some mathematics items required reading of simple text. Since the NEA was not originally designed to yield additional information about the specific language problems that students were having (in fact, the test assumed reading comprehension), it was not possible to determine whether students performed poorly because they could not read, or for some other content-based reason.

Therefore, the test developers added to the 2013 P3 English instruments a series of EGRA-like items, which had originally been tailored for Nigeria and specifically for this purpose. These items required children to read a passage silently and then answer a series of comprehension questions, thus tapping foundation-level skills that had not been separately assessed in the 2011 NEA.

A third source of linking was to include a small number of publicly released items from international tests for both mathematics and English. Some Trends in International Mathematics and Science Study (TIMSS) mathematics items (especially those testing higher-order cognitive processing skills such as mathematical reasoning) and Progress in International Reading Literacy Study (PIRLS) English-language items (for reading comprehension) were embedded in the NEA 2013 test design in order to provide a means of comparing student performance against international content benchmarks.[23] Only items that aligned with the P3 and P6 curriculum were selected.

The number of anchor items by types can be found in ***Table 4.***

[23] However, these items cannot be used to produce TIMSS or PIRLS equivalent scores.

Table 4: Number and proportion of anchor items, by test by type

Grade and subject	Anchor NEA 2011		IEA		EGRA		Common to P3 and P6	
	No. of items	% items	No. of items	% items	No. of items	% items	No. of items	% items
P3 maths	6	20.0%					**1**	3.3%
P3 English	6	20.0%			4	13.3%		
P6 maths	6	15.0%	5	12.5%			**1**	2.5%
P6 English	10	25.0%						

A number of conditions[24] must be satisfied for test equating or linking analyses to be conducted:

1. The number of anchor items must represent at least 20% of the total number of items in the test.
2. Both tests to be equated must measure the same construct (uni-dimensionality).
3. Anchor items must be independent.
4. The set of anchor items must be representative of the test as a whole in both content and psychometric characteristics.
5. Sample size must be a minimum of 1,800 cases.
6. The tests must be of the same length.

With the exception of point 6 (considering the length and time modifications made from 2011 to 2013), these guidelines were adhered to.

2.1.6 Specification Tables

Mathematics and English-language experts in Ghana developed the specifications for the NEA. To determine the strands and learning outcomes for each domain, experts used the curriculum descriptions for English and mathematics for P3 and P6. Generally speaking, the curriculum-defined learning outcomes for P3 and P6 English and mathematics are measured in the NEA through a relatively short test made up of multiple-choice test items.

Section 2.7 further discusses the psychometric properties of the items and the tests.

2.2 Population and Sample Design

The 2013 NEA sample used 2011/2012 EMIS data, which was essentially a census of all primary schools in Ghana. After exclusion of schools that contained a P3 or P6 pupil enrolment less than 10 pupils[25] (*n* = 3,156 schools), 15,609 schools remained in the sample frame. Schools were stratified by region and sorted by district, locality (urban or rural), school type (public or private), and enrolment size. For each region, 55 schools were

[24] See Angoff, W. H. (1984). *Scales, norms and equivalent scores.* Princeton, New Jersey: Educational Testing Service.
[25] Exclusions based on enrolment less than 10 P3 or P6 pupils were implemented in the previous three NEAs (2007, 2009, 2011).

randomly sampled with equal probability, for a total of 550 schools. All P3 and P6 pupils attending a selected school on the day the NEA was administered (July 9, 2013) were automatically selected to complete the assessment. In all, 19,458 P3 pupils and 17,447 P6 pupils completed the NEA administration (total 36,905). ***Table 5*** summarizes of the number of schools included in the sample and the number of P3 and P6 pupils who took the NEA 2013 for each region.

Table 5: Number of sampled schools and number of completed P3 and P6 NEA 2013 completed tests

Region	No. of sampled schools*	No. of pupils completing NEA	
		P3 pupils	P6 pupils
Ashanti	54	1,787	1,555
Brong Ahafo	55	2,049	1,792
Central	55	1,462	1,466
Eastern	55	1,731	1,559
Greater Accra	55	2,224	2,145
Northern	55	1,661	1,521
Upper East	55	2,690	2,246
Upper West	55	2,489	2,128
Volta	55	1,782	1,609
Western	55	1,583	1,426
Total	**550**	**19,458**	**17,447**

* One school in Ashanti and one school in Northern Region could not be assessed because the schools were not in session during July 9–11.

Although at least one school from each of the 170 districts[26] was randomly sampled, the sample size was insufficient to make valid statistical inferences at the district level. In other words, the sample size was selected to analyse data at the national and regional levels, not at the district level.

Sample weights were generated at the school level as the total number of schools divided by the sampled number of schools in each Region. For further information about the sample weights, please see ***Annex A.***

2.3 Training and Preparation of Test Materials

2.3.1 Preparation of Test Materials

RTI worked with CSX Customer Services, a South African firm, to format and create double-sided answer sheets. These new forms had additional features that added value to the NEA test administration and collected data, including:

[26] In June 2012, the federal government increased the number of districts from 170 to 216. The EMIS unit had not yet had time to update the current districts within the EMIS data; therefore, we were able to comment only on the 170 districts.

- Improved efficiency, with pupils using only one answer sheet for both maths and English tests. This facilitated the linking of each pupil's performance across the two subjects.
- Additional demographic information, such as age and class in the previous school year. The test forms were designed so that pupils entered their demographic information only at the beginning of the first test, reducing administration time.

RTI also worked with CSX to format and create a scan-enabled test monitoring form. In 2013, NEA test monitors used this form to record their observations during test administration. This provided for more rigorous analysis of monitoring information, which in turn produced information to help understand best practices and challenges during the 2013 NEA and will guide and enhance administration in the future.

2.3.2 Training of Master Trainers

Together with ASU, RTI staff conducted a training-of-trainers session for the 2013 NEA. Held over three days, May 20–22, this workshop trained nine individuals, selected either from one of the Regional Education Offices or from the Curriculum Research Development Division (CRDD). These individuals conducted 10 regional trainings at five training centres from late May through mid-June.

Based on lessons learned from the 2011 training, the goals of the 2013 training were extended and enhanced to ensure that participants:

- Developed the ability to present a standardised training program to instruct test administrators and test monitors, and obtained all instruction needed to oversee the test monitors and administrators throughout the NEA implementation.
- Developed a thorough understanding of the details of the 2013 implementation plan, protocol, and procedures.
- Became conversant in the reading of instructions and learned to present them and train others in using them in a standardised fashion.
- Developed the ability to administer and train for the test in a uniform manner.
- Developed the skill of creating, modifying and using PowerPoint for training.

2.3.3 Training of Test Administrators

An average of four Master Trainers led the training in each of six training centres, and over 800 test administrators and test monitors participated in the three-day training. The training objectives were to ensure that all test administration was standard; to give participants an opportunity to familiarise themselves with the implementation plan, protocol, and procedures; and to practice presenting the instructions in a standard fashion.

Improvements in the test administrator training from 2011 were the increased liveliness and interactive 'learning-while-doing' (versus lecture) approaches applied. There were important guided role-playing activities, peer learning experiences, and self-evaluation facilitated by the trainers. The 2013 NEA witnessed an unprecedented number of new test administrators and monitors. As a result, facilitators used veterans' questions as a way to provide answers that would distinguish NEA 2011 test administration from NEA 2013. It enabled all participants to seek clarification on both the procedural and the material changes affecting the 2013 NEA test administration.

The importance of an active-learning process was clear and will be applied and improved upon in future trainings—including, but not limited to, the use of real practical experiences. Even though there was time for this in the 2013 training, test administrators continued to struggle with reading and explaining general instructions clearly, and they could have used additional practice administering the listening comprehension part of the test.

2.3.4 Packaging of Test Material

During the NEA regional trainings from May to mid-June, ASU collected district enrolment figures from the participants. The week of June 17, RTI entered these data into a spreadsheet used to create the packing allocation (materials) forms. Nine schools had to be replaced because of inaccessible roads during the rainy season, school closings, or a shift from primary to secondary. Materials included instruments, envelope labels for packing and color-coded plastic tags used to identify each school bag.

NEA test forms and storage bags

One day before packing, ASU and RTI conducted a mini-training session with the packing staff. In total, 14 packers, three CRDD staff, two ASU staff, and two RTI staff packed over 44,000 sets of testing materials and controlled the quality of each school material, with two packers for each school. The packing phase took nine days.

The packaged material was distributed by truck in six days, an improvement over the 2011 distribution, which took 10 days. In addition to the testing materials, districts received an envelope containing:

- a list identifying the sampled schools;
- a letter to each school, informing them of the test dates;
- monitoring forms;
- a note to the test monitors listing the contents of the district envelope and instructions on how and when to distribute the contents; and
- copies of the test administrator scripts, translated into 11 local languages (i.e., to be given to the test administrators).

After the packing and distribution, a debriefing meeting was held with ASU and ASU-based RTI staff to discuss the master and regional training program and packing issues. This provided an opportunity to identify lessons and areas for moving forward, for capacity building and an improved process in 2016.

2.4 Data Collection

Data collection for the 2013 NEA took place July 9–11, with administration completed in approximately 60% of the schools on the first day, July 9. By July 11, all schools had completed the tests.

The test booklets were collected from the Regional Centres July 21–28. All the test booklets and answer sheets were returned in a timely fashion. The ASU administrative assistant made a record in a field data inventory of all the forms and materials from the field as they were returned.

The following factors increased the efficiency of the NEA 2013 data collection based on lessons from 2011:

- Fewer test forms
- Two-sided bubble sheets (i.e., manually completed forms for optical scanning)
- Numbers of items on bubble sheets aligned with number of items on the test
- Improved Master Trainer training
- Introduction of standardised translated instructions to enable them to be given in the local language as well as English
- Creation of bubble sheets for the monitor forms

2.5 Data Cleaning and Scanning

Bubbled answer sheets were cleaned prior to scanning. Typical problems addressed included partially filled forms, empty sheets to be removed, poor bubble completion, extra marks on the forms, and torn and mangled forms. The ASU team advised that the lack of knowledge among students about how to fill in a bubble sheet was extensive and problematic, and therefore suggested introducing a more extended session on the training day to train children on how to appropriately fill in the bubble sheets to record their multiple choice answers.

Scanning of the data took place August 5–9. This was the first year that the monitor forms were developed for scanning (as opposed to being recorded manually). Although substantially more effective, the monitor bubble sheets did not have the school codes pre-coded on the sheet, nor were these codes filled in manually by the monitors prior to the monitoring visit; thus, until this was corrected, it was impossible to link the monitor with the school itself. Once this was corrected by hand-entering the school code, the administration review was quite productive. In the future, either the school codes should be entered before the test administration, or the monitors should be trained to fill in the bubble sheet with a school code before they visit the school.

2.6 Data Processing and Quality Control

2.6.1 NEA Test Data

Data were checked for accuracy and completion. Sample weights[27] were adjusted at the school level and applied to the achievement test data set. Answer keys were cross-checked with the pupils' responses and the test scores were calculated.

Item-level data were documented. Anchor items were verified one by one to allow test linking.

2.6.2 Contextual Data

Contextual data were collected for schools and pupils using EMIS information at the school level and a brief pupil questionnaire (i.e., to record age, gender, grade attended in the last school year and the current year—to measure repetition). School data included location (district, region, urban/rural), type (public/private) and whether the school was placed within a deprived district or not.

The number of pupils by region is given in Table 5 above. ***Table 6*** provides information on the number of P3 and P6 pupils—as well as missing data from the 2011 and 2013 NEA administrations—by (1) location, (2) sex of pupil, and (3) average age.

Table 6: Pupil demographics in 2011 and 2013

	2011 NEA				2013 NEA			
Category	**P3**	**% missing**	**P6**	**% missing**	**P3**	**% missing**	**P6**	**% missing**
Urban	32.3%	13.9%	35.1%	14.4%	31.5%	14.6%	34.3%	15.3%
Female	49.7%	0.2%	48.5%	0.1%	48.3%	0.3%	48.3%	0.1%
Mean age (years)	Not collected	Not collected	Not collected	Not collected	10.7	0.4%	13.5	0.1%

Pupil demographics were very similar for the 2011 NEA and the 2013 NEA.

2.7 Psychometric Analyses

Item analyses included distractor analyses, item difficulty analyses, the progression of difficulty level along the continuum of the test (e.g., progressing from lesser to more difficulty levels), and reliability analyses. The difficulty levels of each item (i.e., percentage of respondents marking the item correct, or '*p*-values') were plotted to evaluate the item difficulty along the continuum of the test and also to evaluate the variability among *p*-values. Ideally, the difficulty of items should increase from the first item (least difficult) to the last item on the test (most difficult). In English, this requirement was almost perfectly met for each 2013 instrument. In maths, this is harder to achieve given the variety of strands measured along the test. The actual plots of item difficulty across the 30 test items on each instrument are provided in ***Annex B.***

[27] For sampling information, see Section 2.2. For more information on weighting, see Annex A.

It can be seen in ***Table 7*** that the test reliability across instruments and forms met or exceeded the conventional acceptable value for tests of this nature (Cronbach's alpha ≥.70), suggesting that, overall, internal consistency was acceptable for all tests and forms. However, reliability findings presented in Table 7 show that the P6 mathematics tests only just met this standard; thus, further analysis and measures to improve the reliability of the P6 mathematics tests should be undertaken before the next NEA administration.

The reliability of the tests increased from 2011 to 2013 for the English portions of the NEA and remained the same for P3 maths. In 2013, there was considerably greater variation in the range of the skills tested than in 2011, given the inclusion of Reasoning/Critical Thinking. These changes had no impact on test reliability except for P6 maths.

Table 7: Comparison of test reliability results, 2011 to 2013*

Test Year	P3				P6			
	Maths		English		Maths		English	
	Form 1	Form 2	Form 1	Form 2	Form 1	Form 2	Form 1	Form 2
2011	0.82		0.84		0.84		0.89	
2013	0.79	0.80	0.86	0.86	0.70	0.71	0.90	0.90

* As measured by SPSS Kuder-Richardson-20 (KR20) reliability analysis.

As noted, more details from the test reliability analysis are presented in Annex B, including Wright plots of pupils' ability and item difficulty.

Further analyses using item response theory and standard Rasch model analysis provided additional psychometric information, with improvements in the psychometric characteristics of the tests discussed in the following.

In 2011, in maths at both P3 and P6, the person and item targeting was a standard deviation off, indicating a test likely too difficult for the target population; the distribution of pupils' ability was relatively normal, and the spread of items was compacted, with substantial item redundancy.

For the English forms in 2011, the pupil and item targeting was close or even perfectly aligned, indicating that the test forms were pretty well aligned to the target population; the distribution of persons was relatively normal, albeit with quite a bit of measurement redundancy.

The 2013 instruments demonstrated reduced measurement redundancy (i.e., fewer numbers of test items, but more diverse content and ability level) for both subjects and grades. The mathematics instruments achieved better alignment of the maths test with pupils' ability level. However, although aligned to curricula, the mathematics instrument was still likely too difficult for the target population, especially at P6.

Despite the efforts to reduce the amount of text and written instructions associated with the operations problems, some word problems were necessary to test mathematical reasoning in P6 mathematics. Thus, given some students' limited capacity to read and understand, it is likely that the reading requirement negatively effected some students' maths performance on the NEA.

2.8 Test Linking

As noted previously, in 2013 the NEA itself and the administration procedures were modified to address limitations in the 2011 administration. As a result, scores from 2011 and 2013 were expressed on different score scales.

Because *equating* methods are intended for test forms created using the same specifications, the procedures we used to compare the 2011 and 2013 data are more appropriately referred to as *linking* rather than equating. ***Annex C*** explains the technical details of how we used equating methods to link the 2011 and 2013 test forms to a common scale, so as to facilitate score comparisons across administrations. Equating procedures help to determine how students participating in the 2011 NEA would perform on the 2013 assessment; in other words, it is desirable to determine whether the examinee population in 2011 would score similarly to the 2013 examinee population on the 2013 NEA.

In the analyses, we used a *non-equivalent groups* equating design to link the 2011 and 2013 test forms. To control for differences across groups, students also were assessed on a subset of test items that was common across the forms. Four types of equating were conducted: mean, linear, circle-arc, and equipercentile equating. The sample sizes supported the application of the frequency estimation equipercentile equating method, considered to be the most appropriate method in this case, yielding similar results as the mean and linear equating methods (see highlighted rows in the Annex C tables). According to this analysis, students participating in the 2011 NEA and in the 2013 NEA *would* perform similarly from one assessment to the other (i.e., 2011 examinees would perform similarly to the 2013 examinees on the 2013 NEA instruments). The estimated mean (percent correct) scores were similar within grades and subjects, comparing 2011 and 2013. When frequency estimation equipercentile equating was used to convert 2011 scores to the 2013 scale, the estimated 2013 score distribution was similar to that of the 2013 sample population, indicating that student performance in English and maths was similar across years. It should be noted that this is merely a comparison of descriptive statistics; no statistical tests were conducted to examine the score differences further.[28]

2.9 Methods for Estimating the Proportion of Pupils Likely to Be Nonreaders

For the 2011 NEA test forms, only six items were included to assess reading comprehension at the Knowledge and Understanding cognitive ability levels. To estimate the proportion of pupils likely to be non-readers, only the scores for the six P3 items were used (non-readers proportion was not estimated for P6 pupils in 2011). When a P3 pupil had 0 or 1 item correct out of 6 (i.e., a score lower than what would be obtained by guessing, or 25%), that pupil was considered likely to be a non-reader. Indeed, these were six very simple items for which explicit information could easily be retrieved from the text.

In 2013, additional items were added for assessing reading comprehension (in total, 12 items at P3 and 16 items at P6, Knowledge–Understanding cognitive ability level) and when

[28] Albano, T. (2013). *Ghana NEA equating report.* Prepared for the Ghana *Testing* activity under the USAID EdData II project, Task Order No. AID-641-BC-13-00001. Research Triangle Park, NC, USA: RTI International.

evaluated, the language tests had excellent psychometric properties. The P3 test forms included the aforementioned EGRA-like reading comprehension questions.

Therefore it was possible to create four indicators (referred to below as Methods 1 through 4) to distinguish non-readers:

1. Pupils having ≤ 25% correct answers on the Knowledge and Understanding reading items
2. Pupils having zero scores on the four items, assessing lower cognitive skills (Knowledge cognitive ability level)
3. Pupils having zero scores on the four EGRA-like items
4. Pupils in a particular class (i.e., category) that evidenced a lower probability of responding correctly to the item (using latent class analysis, or LCA)

It is worth noting that for the typical EGRA, the mode of administration is different from that of the NEA EGRA-like items. That is, in the standard EGRA, the pupils read a passage aloud and then answer open-ended comprehension questions asked by the assessor, without being able to see and refer to a printed stimulus sheet. In the NEA, the pupils are allowed to refer to the passages when answering the comprehension questions, which are in multiple-choice format. Therefore, data from the NEA EGRA-like items cannot be compared directly to standard EGRA test results.

Method 1 for 2013 was consistent with the 2011 method, although incorporating both anchor items and a larger number of items testing the same cognitive abilities and at the same level of difficulty.

Methods 2 and 3 were very restrictive, somewhat like binary decisions, and based on only four items.

For Method 4, we used latent class analysis (LCA) on nine reading items spread over various cognitive skills and based on different stimuli/texts. LCA can identify classes of pupils and their associated probable responses to each item. Criteria such as homogeneity, separation, and goodness of fit are used to determine the optimum number of classes.[29] For this study, the class with the lowest probability of correct answers would be designated as the non-readers class.

All four methods were applied and results were compared by crossing the various classifications of non-readers. Each method was also crossed with the minimum competency level in English. The use of latent class analysis allowed us to make statements about the validity of the different methods.

For comparison, note that such methods have been used in secondary analysis of data from the Ghana 2011 NEA and from Cameroon.[30] In 2011 in Ghana, the proportion of non-readers estimated using the LCA approach under a three-class model was 49.5%. The Cameroon data

[29] See (1) Lanza, S. T., & Collins, L. M. (2010). *Latent class and latent transition analysis with application in the social, behavioral, and health sciences.* New York: Wiley; and (2) Haertel, H. E. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301–321, http://dx.doi.org/10.1111/j.1745-3984.1989.tb00336.x.

[30] Varly P. & Abarda. A. (2013). *Utilisation de l'analyse en classes latentes pour l'évaluation des acquis scolaires : application aux données du Ghana et du Cameroun.* Rabat, Morocco: Varlyproject.

included both responses to paper-and-pencil items[31] and results for individual oral fluency assessments. The LCA method was useful in this case to estimate the proportion of non-readers based on the paper-and-pencil test as compared to the oral fluency test. If, as should have been the case, the LCA method more accurately estimated the proportion of non-readers (i.e., 49.3% in grade 5 in Cameroon) then 25% of the pupils were misclassified as either readers or non-readers.

Last but not least, during the NEA 2013 pilot testing, administrators observed a proportion of students responding randomly to the test—that is, pupils were just filling in the bubble sheets randomly, working more quickly than possible if they were reading the questions, marking the same lettered answer for each item, or marking in a visual or other pattern. It is impossible to distinguish pupils who cannot read from pupils who can read but for some reason choose to respond randomly to the test. However, during the pilot phase, when the research team combined the NEA and EGRA individual oral fluency tests on a small sample of pupils, most pupils who could not read aloud also responded randomly to the written test. Nevertheless, if the estimated proportion of non-readers included both pupils who truly were non-readers and pupils who responded randomly for any other reason, the real proportion of non-readers would be somewhat overstated.

Finally, Method 1 was selected for the 2013 analyses because it allowed comparison of results between 2011 and 2013, the results were consistent with those of more sophisticated methods (LCA), it was based on a sufficient number of items, and it is easy to understand.

3. NEA 2013 Results

The following presents the NEA 2013 findings. In this section results are presented as either average (percent correct) scores or as a percentage of students falling within the defined cut-points for *minimum competency* (i.e., 35% correct) or *proficiency* (i.e., 55% correct or greater) along with the percentage of students who failed to achieve minimum competency (i.e., <35% correct).

Sections 3.1 and 3.2 present the overall performance for subject area (English and mathematics) and class (P3 and P6). Section 3.3 presents a brief comparison of learning outcomes between the 2011 and the 2013 administration. Section 3.4 contains overall findings according to the sub-domains for the English and mathematics tests and for three different cognitive ability levels required to perform tasks. Section 3.5 looks at the relationship between pupil learning outcomes on the English and the mathematics subtest. Section 3.6 presents learning outcomes among the various sub-populations, including gender; urban/rural; deprived districts/not-deprived districts; public/private schools; and region. Section 3.7 reports on the findings from a set of multiple regression analyses, and Section 3.8 presents results of the four estimation methods for non-readers.

3.1 Overall Scores

Before we present the information on overall scores, note that all four subject tests (P3 mathematics, P3 English, P6 mathematics, and P6 English) used the definitions of minimum

[31] Including open-ended questions to limit random guessing.

competency and proficiency that have been applied in all previous NEA applications. Thirty-five percent was set as the minimum competency because it would indicate that students had achieved a score higher than if they had guessed all the questions or answered randomly.[32]

Note, however, that international standards generally classify students as proficient if they have achieved a minimum score of 70%. For comparability reasons, this report presents NEA's traditional classifications of proficiency (≥55%), but readers should be aware that this definition defines proficiency as obtaining just over half of the items correctly and does not effectively identify students who have a full grasp of the curriculum—that is, pupils who are truly proficient in the subject area. The proportion of pupils reaching 70% discussed within ***Annex D.***

The scores in ***Table 8*** represent the average percentage of correct answers (i.e., mean 'percent correct' scores) for all students, by grade and subject.

Table 8: Mean percent correct scores, by class and subject

Subject	Mean scores by grade	
	P3 (95% confidence interval)	P6 (95% confidence interval)
Maths	41.1	38.2
	(40.0 – 42.2)	(37.4 – 39.0)
English	44.4	48.9
	(42.8 – 45.9)	(47.3 – 50.6)

On average, students in all grades and subject areas obtained less than 50% of the items correct. To understand the range of student performance, it is helpful to look at the distribution of students who fell below minimum competency, who achieved minimum competency, and who met the criteria for proficiency.

3.2 Pupils Reaching Minimum Competency and Proficiency

Table 9 shows the proportion of pupils achieving minimum competency and proficiency as defined by the 'at least 35%' and '55% and above' response thresholds. We have also incorporated a third category to present the percentage of pupils whose performance fell below the minimum competency level (had less than 35% of the items correct).

In P6 maths, only 10.9% of the pupils reached proficiency, versus 22.1% in P3. In English, the results were better, with 28.4% of the pupils reaching proficiency at P3 and 39% at P6. For both P3 and P6, approximately 40% of the students failed to achieve even minimum competency in mathematics and approximately 40% of the P3 students failed to achieve minimum competency in English.

[32] As there are always three distractors and one correct answer in every multiple-choice item, pupils answering randomly can expect to get 25% correct answers.

Table 9: Percentages of students achieving minimum competency and proficiency, by subject and class

Cut-points	P3		P6	
	Maths	English	Maths	English
Below minimum competency	42.9%	41.9%	39.2%	31.3%
Minimum competency	35.0%	29.7%	50.0%	29.8%
Proficiency	22.1%	28.4%	10.9%	39.0%
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**

3.3 Trend Analysis

Improvements in the NEA 2013 test forms facilitated score comparisons for the 2011 and 2013 administrations. The 2011 and 2013 score scales were statistically linked to one another using what is referred to as the frequency estimation equipercentile method. The linking process is intended to account for differences in the score scales due to differences in the length and difficulty of the test forms themselves. Once these differences have been accounted for, any remaining differences in the linked score distributions are expected to result from performance difference for the groups being compared.

Table 10 illustrates which portions of the information necessary for year-to-year comparisons were present for each year of the NEA data collection.

Table 10: Status of data for each year of NEA administration

Data requirements for multi-year analysis	Year of data collection				
	2005	2007	2009	2011	2013
Item-level data	NO	YES	NO	YES	YES
Item content	NO	NO	YES	YES	YES
Test linking	NO	NO	NO	NO	YES

Findings given in ***Table 11*** demonstrate that pupils in 2011 and 2013 performed similarly on their respective assessments. The 2013 mean (percent correct) score was not dramatically above or below the 2011 score equivalents.

Table 11: Overall mean percent correct scores, by class and subject, 2011 and 2013

Subject	Year	Class	
		P3	P6
Maths			
	2011	38.6	39.5
	2013	41.1	38.2
English			
	2011	44.0	49.8
	2013	44.4	48.9

3.4 Results by Domains and Cognitive Abilities

3.4.1 Results by Domains

The questions in mathematics and English covered multiple domains, providing an opportunity for deepening our understanding about students' relative strengths across domains and for identifying potential gaps in performance according to the different domains tested. ***Table 12*** presents the English mean (percent correct) scores by domain. See Section 2.1.2 for an explanation of the number of items per domain.

On the English tests, performance on Reading was low relative to Listening and Grammar. Although in general the scores were higher for English Listening Comprehension, performance on English Listening Comprehension was still lower than what one would expect, especially for grade 6.

Table 12: NEA 2013 mean English test scores, by grade and domain

Domain	Class	
	P3 Mean % correct (95% confidence interval)	P6 Mean % correct (95% confidence interval)
Listening	57.5 (55.9 – 59.1)	54.7 (53.1 – 56.3)
Grammar	41.1 39.5 – 42.7	51.7 (50.1 – 53.4)
Reading	38.3 (36.8 – 42.7)	43.2 (41.4 – 45.1)

In mathematics, shown in ***Table 13***, students had the most difficulty with tasks in the Measurement/Shape and Space domain, for both P3 and P6. Questions assessing mathematics skills in the Collect and Handle Data domain also presented special challenges for P3 students, and for P6 pupils to a lesser extent. Pupils did poorly in completing these types of tasks in the 2011 NEA as well, as shown in ***Table 14.***[33]

Table 13: NEA 2013 average maths test scores, by grade and domain

Domain	Class	
	P3 Mean % correct (95% confidence interval)	P6 Mean % correct (95% confidence interval)
Operations	47 (45.6 – 48.3)	44.7 (43.8 – 45.6)
Numbers	45.6 (44.4 – 46.8)	37.9 (36.9 – 38.9)
Measurement / Shape and Space	35.5 (34.7 – 36.3)	26.9 (26.4 – 27.5)
Collect and Handle Data	34.0 (32.6 – 35.5)	44.7 (43.4 – 46.0)

Table 14: 2011 average test scores, by grade and domain

Domain	% correct answers	
	P3	P6
English		
Listening	63.6%	72.8%
Grammar	35.1%	44.1%
Reading	40.6%	44.1%
Writing	33.9%	32.4%
Maths		
Numbers	39.3%	40.7%
Operations	42.9%	42.4%
Measurement	38.8%	Not applicable
Shape and Space	41.2%	37.3%
Collect and Handle Data	29.0%	44.0%

3.4.2 Results by Cognitive Abilities

The 2013 test assessed a greater variety of cognitive skills than previously, especially in the area of Critical Thinking and Reasoning.

[33] There were not enough anchor items by domain to draw conclusions about changes over time by content domain from 2011 to 2013.

Table 15 presents the average percent correct scores for P3 and P6 students according to the cognitive ability levels of the test items in mathematics. It can be seen in Table 15 that the higher the cognitive ability level, the lower the overall performance of students, in both P3 and P6. Although mathematics items involving the mathematical reasoning were challenging for both P3 and P6 pupils, P6 pupils' performance in particular was quite low. The average percent correct score was 23.3% for P6 students, which is lower than the score that a pupil could have obtained by answering randomly or 'by chance' without any reference to the questions.

Table 15: Scores by cognitive skills and grade: Maths

Cognitive level, mathematics	Class	
	P3 Mean % correct (95% confidence interval)	P6 Mean % correct (95% confidence interval)
Knowledge and Understanding	45.8 (44.6 – 46.9)	44.9 (44.0 – 45.8)
Application	38 (36.9 – 39.0)	37.5 (36.6 – 38.4)
Reasoning	33.5 (32.3 – 34.6)	23.3 (22.6 – 23.8)

Table 16 presents the average (percent correct) scores for P3 and P6 students across three cognitive ability levels assessed in English. Similarly to mathematics, the higher the cognitive ability level, the lower the overall performance of students, in both P3 and P6. Results were more encouraging for English, however, especially for P6 students. Over half of the P6 pupils answered at least 50% of the test items correctly for items involving both Knowledge and Understanding cognitive abilities, but items involving Critical Thinking were difficult for P6 pupils. P3 pupils' performance also was low on questions involving Understanding cognitive abilities and Critical Thinking (i.e., 38.0% and 36%, respectively).

Table 16: Average scores by cognitive skills and grade: English

Cognitive level, English	Class	
	P3	P6
	Mean % correct (95% confidence interval)	Mean % correct (95% confidence interval)
Knowledge	51.8 (50.0 – 53.4)	54.1 (52.1 – 55.9)
Understanding	38 (36.5 – 39.4)	50.2 (48.5 – 51.8)
Critical Thinking	36 (34.6 – 37.3)	39.4 (37.9 – 40.8)

3.5 Relationship Between Maths and English Results

As noted earlier, in 2011, there was no possibility to link maths and English results. To remedy this situation, in 2013, a double-sided coding sheet was introduced that allowed the test analysts to correlate the two sets of scores. The correlations between pupils' performance in maths and English in 2013 were 0.68 at P3 and 0.66 at P6 and were statistically significant. In other words, there was a significant relationship between pupils' performance on the English and maths tests; pupils who performed well in mathematics were likely to perform well in English. See ***Annex E*** for further analyses of this relationship.

Table 17 and ***Table 18*** cross-tabulate pupil performance in maths and English based on the NEA cut-points for minimum competency and proficiency for P3 and P6. These tables present the percentages of pupils who fell into the same category of performance across English and mathematics, and they speak to the positive relationship given by the correlations above. Clearly it is unlikely that students proficient in one subject (e.g., mathematics or vice versa) will only achieve minimum competency or fail to meet the minimum competency in the other subject (e.g., English or vice versa). Very few pupils were proficient in one subject without having minimum competency in the other subject. Moreover, only 17.2% of P3 pupils were proficient in both English and maths. Even fewer (12.9%) P3 pupils reached the minimum competency level on maths without having the minimum competency level in English.

Table 17: Scale level in maths compared to scale level in English, P3 (in %)

	Competency levels in English		
Competency levels in maths	Below minimum competency	Minimum competency	Proficiency
Below minimum competency	27.7	12.7	2.6
Minimum competency	12.9	13.5	8.6
Proficiency	1.3	3.5	17.2

Even though the test designers limited as much as possible the amount of text that had to be read in the maths test, both the P3 and the P6 mathematics tests included some word problems which required pupils to read. Thus, to perform well in maths, pupils had to demonstrate some level of proficiency in English.

In ***Table 18*** it can be seen that for P6, only 9.9% of the pupils reached the NEA definition of proficiency (i.e., achieved ≥55% correct) in both maths and English. Equally disconcerting is that 21.7% of the P6 pupils failed to achieve even minimum competency in either English or mathematics.

Table 18: Scale level in maths compared to scale level in English, P6 (in %)

	Competency levels in English		
Competency levels in maths	Fail to meet minimum competency	Minimum competency	Proficiency
Below minimum competency	21.7	13.0	4.4
Minimum competency	9.4	15.9	24.7
Proficiency	0.2	0.8	9.9

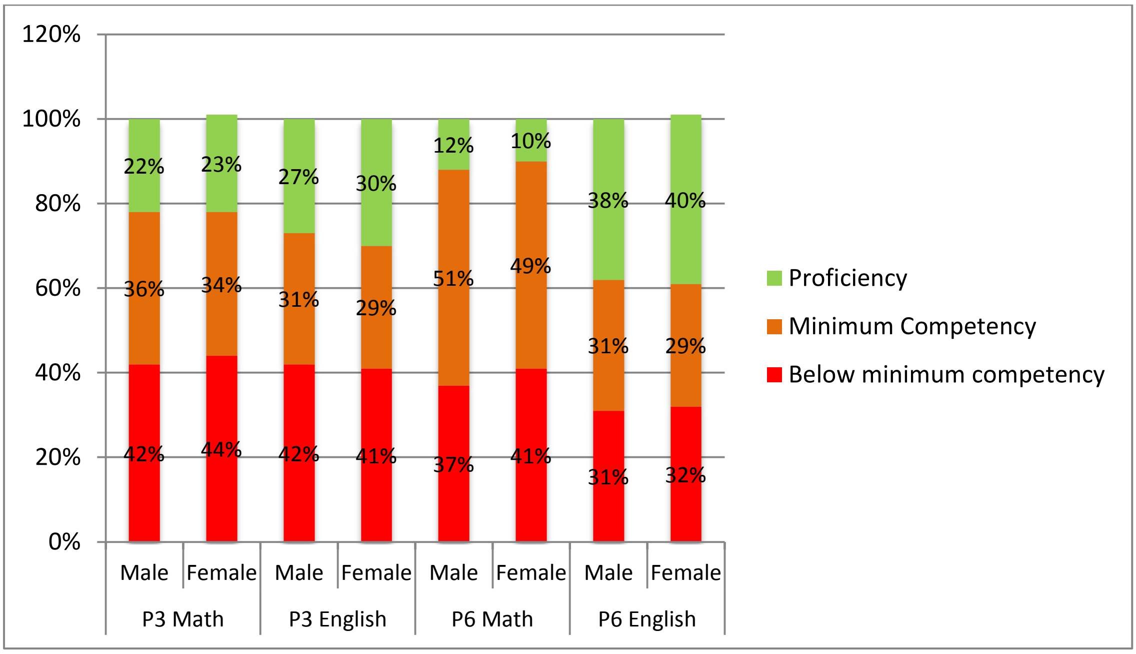
3.6 Analysis of Pupils' Performance by Core Demographic Variables

Test results were also analysed by certain sub-populations, including: students' characteristics, male/female; students attending schools in urban/rural locations, deprived districts/non-deprived districts, and different regions of the country; and students attending public/private schools. The following subsections present results for these student groupings and provide group comparisons on learning outcomes.

3.6.1 Gender

Gender gaps were small when compared to other demographic variables (see ***Figure 6***) and were significant only for P6. For P6 students, males outperformed females in mathematics, with a larger proportion of males than females making the cut-point for both minimum competency and proficiency. Overall, males also outperformed females in P6 English, with a larger percentage of males meeting the cut-point for minimum competency. For P3 in both mathematics and English, there were small differences in the distribution of students across the proficiency and minimum-competency achievement categories, but these differences were not statistically significant. Similar results—small performance differences between males and females which were statistically significant only for P6—were found in 2011.

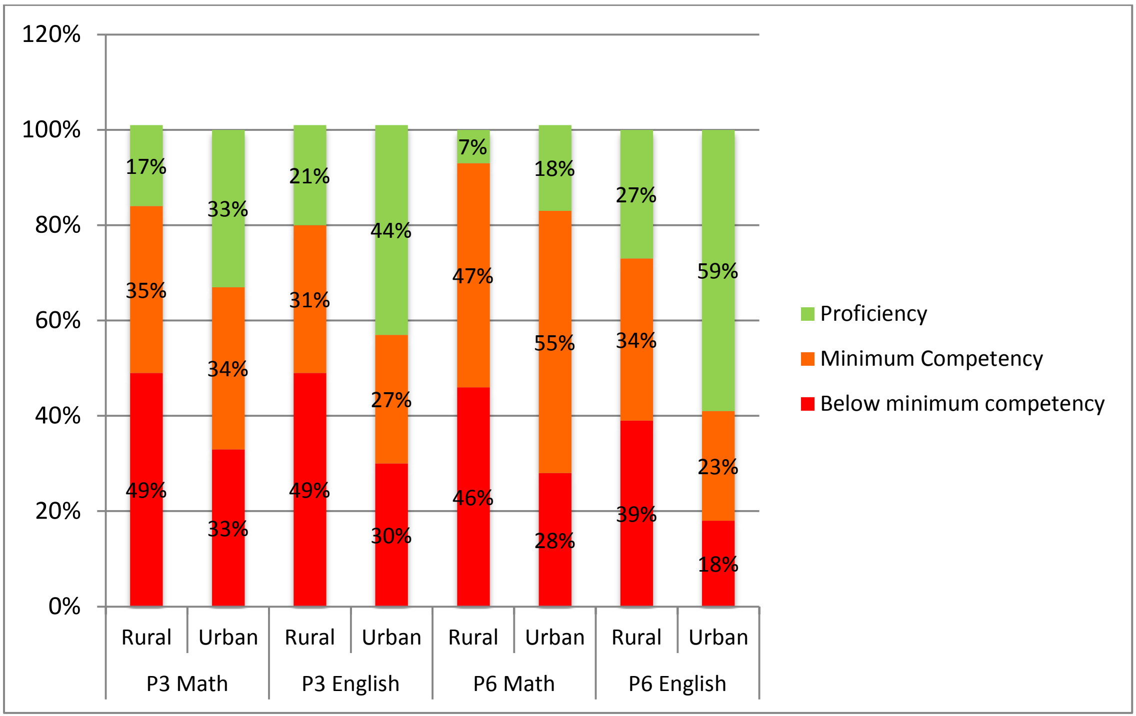
Figure 6: Percentages of students achieving minimum competency and proficiency levels, by gender



3.6.2 Urban vs. Rural

Performance gaps between children residing in rural versus urban areas were considerable and statistically significant for both P3 and P6 and for English and mathematics. In all tests, children residing in urban areas outperformed children residing in rural areas. ***Figure 7*** presents the percentage of pupils achieving the NEA score cut-points across rural and urban locations. The gap between rural and urban for P6 English was particularly large. More than half (59.2%) of P6 English pupils reached proficiency in urban areas versus 27.1% in rural areas. Only 7.3% of the P6 rural pupils reached proficiency in maths. A larger percentage of students in rural locations failed to achieve even minimum competency than did students in urban locations. With the exception of P6 English, approximately half of the children in rural locations fell below the minimum competency cut-point; that is, for children attending rural schools, less than 35% of the test items were answered correctly (almost chance-level responses).

Figure 7: Percentages of students achieving minimum competency and proficiency levels, by school location

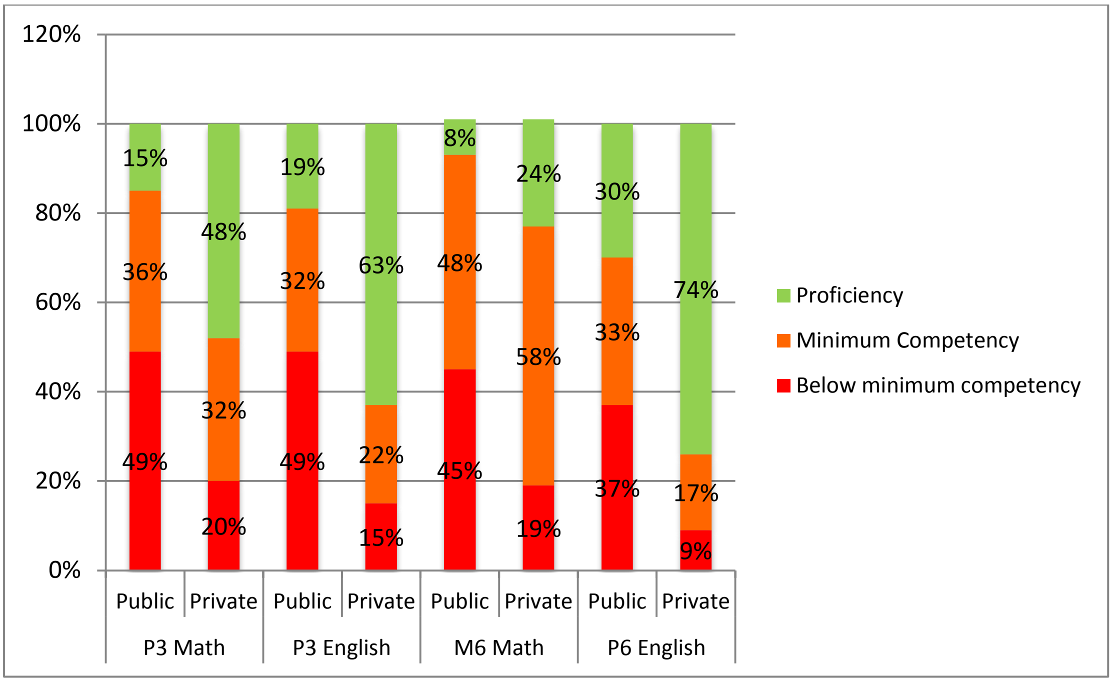


3.6.3 School Type

The NEA is administered in both private and public schools in Ghana. Private schools had much better results overall in 2013 than public schools ***(Figure 8).*** Among all the demographic variables, it was the factor having the largest effect in the regression models (see Section 3.7 below). With the exception of P6 mathematics, the majority of pupils in private schools reached proficiency level. As previously mentioned, few P6 pupils did well in mathematics, even in private schools, where only 23.5% of the students met or surpassed the proficiency cut-point of 55% correct. Fewer than 20% of private-school students failed to achieve minimum competency, in both grades and both subjects. By contrast, with the exception of P6 English, close to 50% of the pupils attending public schools failed to meet the cut-point for minimum competency (35% correct). The results in maths were remarkably low in the public schools, with only 7.6% of the pupils reaching proficiency at P6 and 15.3% at P3.

In English, 63.4% of P3 pupils attending private schools achieved proficiency, compared to 19.2% of the children attending public schools. As a rule, pupils in private schools tended to be more likely to speak English at home (as documented in the 2011 pupil characteristics), and this could explain part of the gap.

Figure 8: Percentages of students achieving minimum competency and proficiency levels, by school type

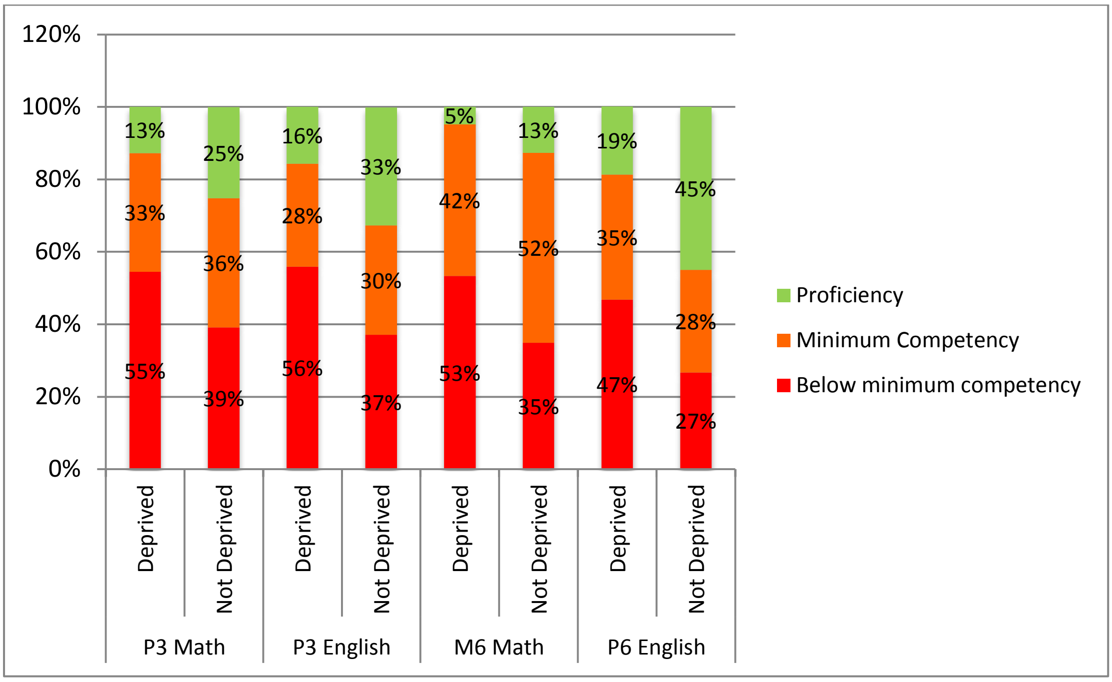


3.6.4 Deprived Districts

Since 1999, the Ghanaian government has classified roughly one third of the districts (originally it was 40 and gradually it increased to 68 as the district mapping was redrawn) as 'deprived.' The classification was drawn from a combination of education outcome and resource indicators including gross enrolment rate (GER) in primary, gender parity, seats and core textbooks per pupil, share of schools needing major repairs, BECE pass rates in both English and maths, per pupil expenditure in primary, pupil–teacher ratio in primary, and the share of qualified primary teachers. The majority of districts in Ghana classified as 'deprived' are in one of the three northern regions of Ghana.

The performance of students attending public schools in deprived districts and non-deprived districts is presented in ***Figure 9.*** For all grades and subjects, the proportion of pupils from deprived districts that met the cut-point for proficiency was less than half the proportion for pupils from non-deprived districts. A much larger proportion of students from deprived districts fell below the cut-point for minimum competency compared to pupils from non-deprived districts.

Figure 9: Percentages of students achieving minimum competency and proficiency levels, by deprived and not-deprived district status



3.6.5 Region

Table 19 presents relevant contextual information by region. As discussed earlier in Section 1.2, public schools in the northern regions and deprived areas are less likely to have qualified teachers or access to materials, and more likely to have minimal physical infrastructure. The higher dependency on public schools and the lower levels of literacy (see Table 19, ***Figure 10,*** and Section 1.2) in these poor and hard-to-reach regions, combined with inequities in public school inputs, may account in part for the lower scores in the most impoverished regions of the country, particularly the three regions of northern Ghana (see Table 19, BECE pass rates; and ***Table 20***); and may also have an impact on completion rates, which are much lower among districts in the two lowest wealth quintiles (see Figure 3 above).

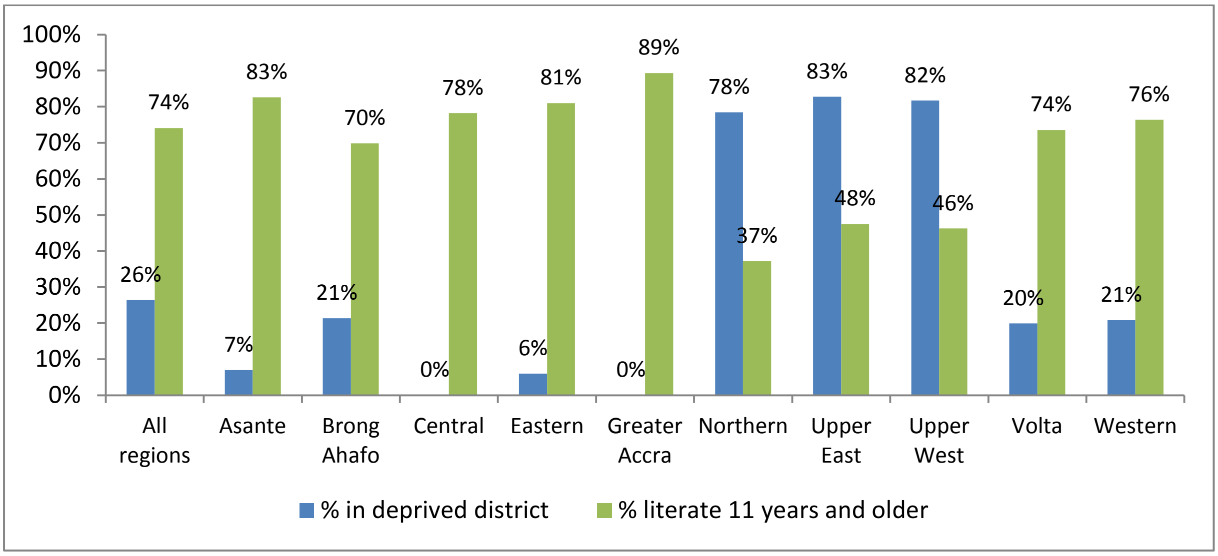
The data presented in Table 19 and Figure 10 underscore noticeable regional differences, especially along three lines: (1) the regions of northern Ghana (shaded pink), (2) Greater Accra (shaded blue), and (3) the remaining six regions (shaded green). Compared to the rest of the country, the three regions of Northern Ghana—where the bulk of deprived districts are located—are characteristically rural, are dependent on public school education, and have much lower percentages of literacy. In contrast, in Greater Accra—which has no deprived districts, is urban, and is much less dependent on public education—literacy rates are high: Approximately 90% of the 11+ population were literate according to 2010 census data, compared to 43.6% on average for the three regions in northern Ghana.

Table 19: Region characteristics

Region	Percent sample in deprived district	Percent sample in urban areas	Percent sample in public schools	Percent literate 11+ years	Net enrol-ment rate (NER)	Percent districts with sub-standard PCE*	BECE pass rates: Maths	BECE pass rates: English
Ashante	7.0%	43.1%	75.8%	82.6%	74%	29%	74%	65%
Brong Ahafo	21.3%	43.0%	82.3%	69.8%	72%	35%	72%	59%
Central	0.0%	23.6%	79.5%	78.2%	73%	No Data	50%	52%
Eastern	6.0%	15.2%	77.0%	81%	75%	15%	53%	55%
Greater Accra	0.0%	75.6%	62.5%	89.3%	82%	7%	65%	81%
Northern	78.4%	27.3%	94.9%	37.2%	59%	65%	39%	42%
Upper East	82.7%	8.5%	91.8%	47.5%	72%	66%	37%	40%
Upper West	81.7%	14.2%	93.8%	46.2%	65%	45%	52%	46%
Volta	19.9%	17.8%	79.5%	73.5%	73%	0%	38%	54%
Western	20.8%	19.6%	75.8%	76.4%	77%	18%	65%	60%
All regions	26.4%	32.8%	81.2%	74.1%	73%	No Data	No Data	No Data
Source	NEA 2013	NEA 2013	NEA 2013	2010 Census	Ghana Stat. Service 2011	World Bank 2010	EMIS 2011	EMIS 2011

* Per child recurrent expenditure.

Figure 10: Regional characteristics: Percentage of pupils in deprived districts and percentage of literate 11+-year-olds



These three regions of the north (i.e., Northern, Upper East and Upper West—see pink shading in Table 20) also share important language features and have considerable more linguistic diversity than most other regions (although this is a feature of Volta as well).

Table 20 shows pupils' performance on the 2013 NEA by region. Note that the confidence intervals about the proportion estimates presented in Table 20 can be seen in ***Annex F.*** The findings presented in Table 20 underscore the impact that attending public schools in the most impoverished and hard-to-reach areas of the country has on a child's learning, especially in contrast to children residing in the urban centres. For all grades and subjects, the percentage of children from one of the three northern regions who met the cut-point for proficiency was half the percentage reaching proficiency in the other regions of the country. The disparity in performance between the northern regions and the urban centres such as Greater Accra was even larger. For all grades and subjects, the percentage of students residing in Greater Accra who met the cut-point for proficiency was three times the percentage of students achieving proficiency in the three northern regions. For example, while 77.7% of children residing in Greater Accra met the criteria for proficiency in P6 English, less than 25% of the children residing in northern Ghana did so. P6 mathematics fared even worse: Averaging across the three northern regions, only 4.9% reached or surpassed the cut-point for proficiency (<3% in the Northern Region). In P3, only about 10% of the children living in one of the northern regions achieved proficiency in mathematics and no more than 14% reached proficiency in English.

Table 20: Percentages of students achieving minimum competency and proficiency levels, NEA 2013, by region

Competency level, by grade and subject	Regions									
	Ashanti	Brong Ahafo	Central	Eastern	Greater Accra	Northern	Upper East	Upper West	Volta	Western
P3 maths										
Below minimum competency	39.4	39.4	44.6	40.8	24.8	62.2	52.7	57.1	43.5	45.0
Minimum competency	37.4	37.3	35.4	34.9	31.3	28.0	37.1	32.5	35.0	38.8
Proficiency	23.2	23.2	20.0	24.3	43.9	9.8	10.2	10.4	21.5	16.2
P3 English										
Below minimum competency	39.6	36.8	38.9	38.0	20.4	62.1	60.8	58.9	40.6	47.9
Minimum competency	31.2	33.5	33.5	31.0	22.5	24.0	28.4	28.8	32.5	31.0
Proficiency	29.2	29.6	27.6	31.0	57.1	13.8	10.8	12.3	26.9	21.0
P6 maths										
Below minimum competency	34.8	35.8	48.4	37.3	17.9	58.9	50.0	46.1	37.3	44.5
Minimum competency	54.8	53.6	44.5	50.2	57.1	38.3	44.4	47.6	51.8	46.9
Proficiency	10.4	10.6	7.1	12.5	24.9	2.8	5.6	6.4	10.9	8.6
P6 English										
Below minimum competency	29.1	31.0	37.7	30.0	7.2	48.2	45.1	40.1	26.5	37.8
Minimum competency	34.0	28.7	33.4	29.7	15.1	35.8	32.2	35.6	27.5	31.1

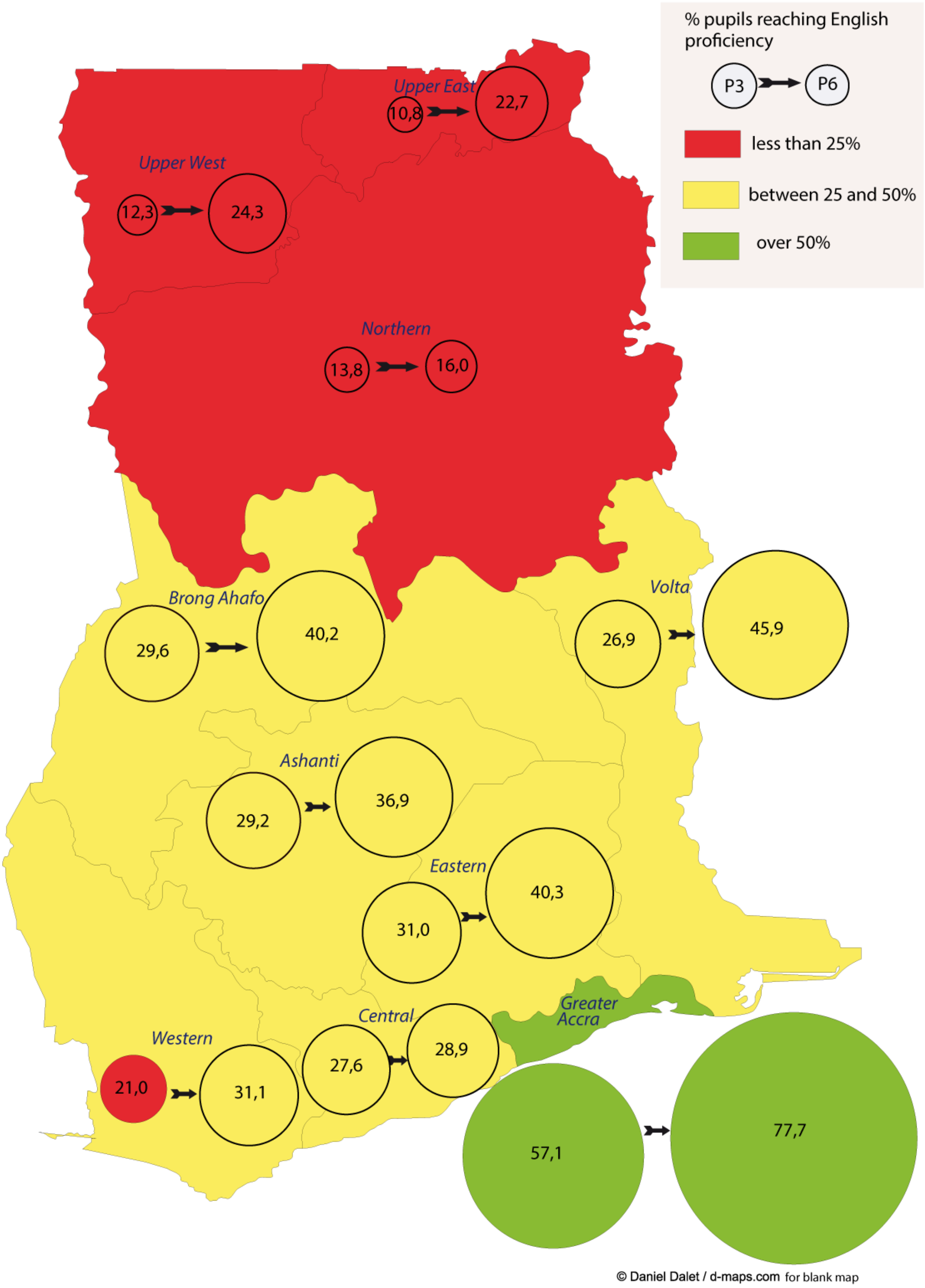
Competency level, by grade and subject	Regions									
	Ashanti	Brong Ahafo	Central	Eastern	Greater Accra	Northern	Upper East	Upper West	Volta	Western
Proficiency	36.9	40.2	28.9	40.3	77.7	16.0	22.7	24.3	45.9	31.1

Source: NEA 2013.

Figure 11 and ***Figure 12*** are maps showing the percentages of pupils achieving the NEA cut-point for proficiency at P3 and P6 in English and in maths. P6 students fared worse with the P6 mathematics curriculum than P3 students did with the P3 mathematics curriculum. The proportion of P6 students achieving proficiency in mathematics was lower than the proportion of P3 students achieving proficiency.

Children without strong foundations tend to decline in their ability to fully benefit from schooling in the later grades, particularly for more complex problem-solving tasks that are more prominent in the later grades. As mentioned earlier, the P6 mathematics tests—and the P6 mathematics curriculum itself (as the tests are curriculum based)—are difficult for students.

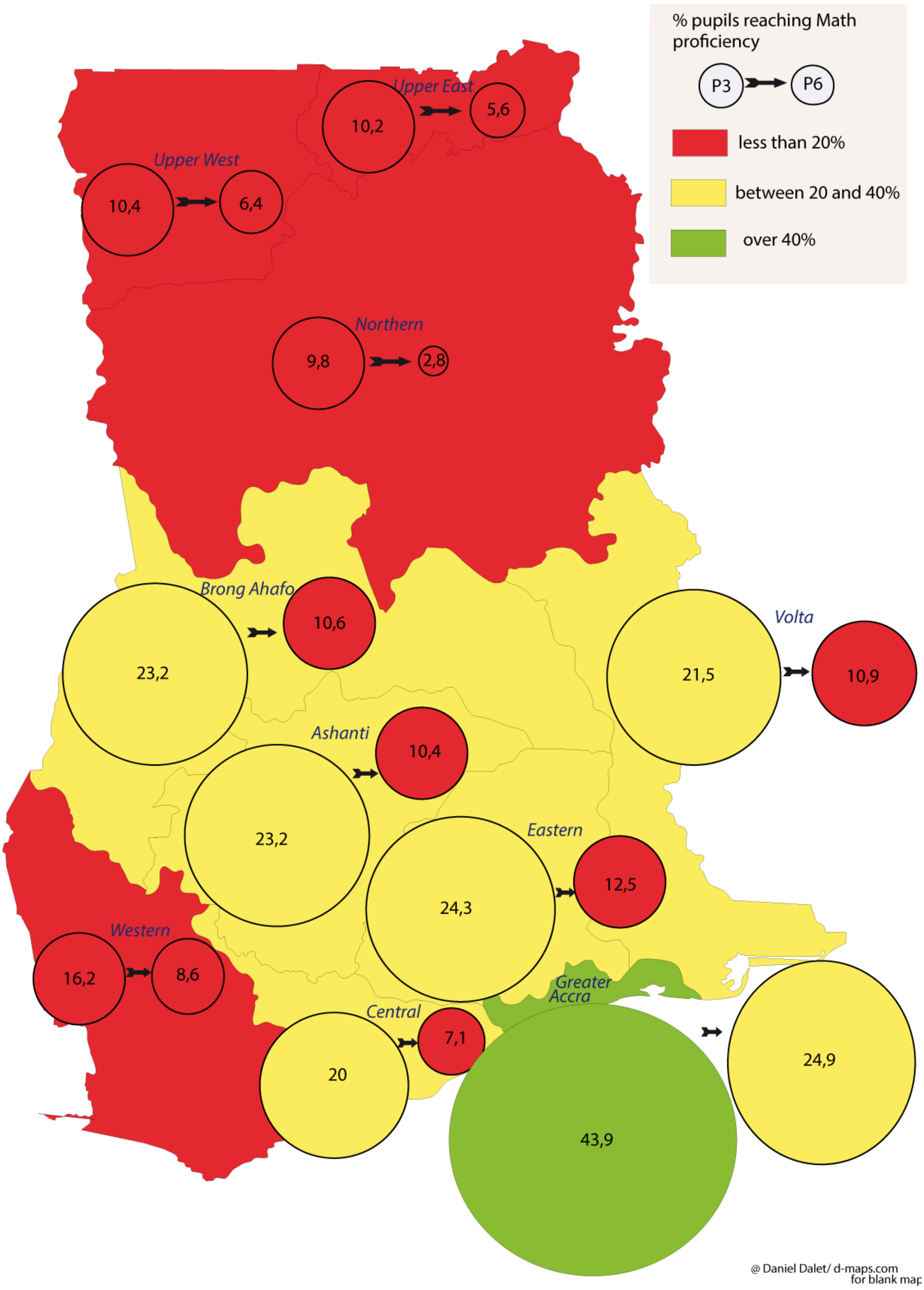
Figure 11: Proportion of pupils reaching 'proficiency' (scores ≥55%), 2013 NEA English, by region, P3 and P6



Data source: NEA 2013

Outline map source: http://d-maps.com/ carte.php?num_car=4676&lang=en

Figure 12: Proportion of pupils reaching 'proficiency' (scores ≥55%), 2013 NEA mathematics, by region, P3 and P6



Data source: NEA 2013

Outline map source: http://d-maps.com/ carte.php?num_car=4676&lang=en

3.7 Multiple Regression Analysis

A multiple regression analysis was conducted to study the relationship between a variety of individual and school characteristics and maths and English learning outcomes for P3 and P6 (see ***Table 21***). According to the analysis of regression coefficients, a number of factors had a significant correlation with learning outcomes. For all subjects and grades (all four tests), the following were identified as being a predictor of academic performance: (1) geographical location in one of the three regions in the north; (2) type of school (public vs. private); and (3) location (urban vs. rural). Students who were enrolled in a private school and students residing in an urban area performed better than children in public schools and children residing in rural areas. Students living in one of the three regions in the north consistently did poorer than students in other parts of the country.

Age was not associated with performance on P3 mathematics, but there were negative and significant correlations between age and P3 English, P6 mathematics, and P6 English. The negative correlations suggest that older children tended to perform less well than their younger peers in the same grade. This may be, in part, a result of repetition and/or late entry into primary school. Gender was not identified as a predictor of performance in P3 English or P3 mathematics, but there was a significant correlation between sex and performance in P6, as noted above. In general, females in P6 obtained lower scores than males in both mathematics and English. Although 'deprived district' was not seen to be a factor related to performance, it is likely that the variable 'residing in the regions of the north' accounted for most of the variance associated with deprived versus non-deprived district. That is, 80% of the pupils residing in one of the three northern regions were also in a deprived district and thus there was a situation of multi-collinearity between the factors 'reside in the regions of the north' and 'reside in a deprived district.' Overall, a greater share of the variance could be accounted for by this set of predictors for English than for mathematics, for both grades.

The relative advantage of being in an urban school versus a rural school is demonstrated from the positive and significant beta coefficients for 'reside in urban location,' for all grades and all subjects tested. The disadvantage of being in a public versus a private school in any region of the country and of being in one of the three regions of the north is demonstrated from the significant negative coefficients, for all grades and all subjects. Thus, children in urban areas tended to perform better than children in rural areas. Children in public schools were more likely to obtain low scores than their counterparts in private schools, and children residing in one of the three northern regions were more likely to perform poorly than children outside of the three northern regions.

Table 21: Regression coefficients: Effect of demographic variables on scores, with confidence intervals (in parentheses)

Multiple-regression statistics	P3		P6	
	Maths	English	Maths	English
Being female	**-0.369**	**0.234**	**-1.610*** **	**-1.158** **
	(-0.990, 0.251)	(-0.502, 0.969)	(-2.063, -1.157)	(-1.946, -0.369)
Age	**-0.223**	**-1.361*** **	**-0.982*** **	**-2.234*** **
	(-0.664, 0.218)	(-1.924, -0.798)	(-1.316, -0.647)	(-2.870, -1.598)

Multiple-regression statistics	P3		P6	
	Maths	English	Maths	English
Reside in urban location	**6.225*** **	**9.879*** **	**4.564*** **	**12.492*** **
	(3.510, 8.939)	(6.385, 13.373)	(2.642, 6.485)	(8.731, 16.253)
Public	**-13.931*** **	**-19.841*** **	**-6.472*** **	**-16.527*** **
	(-17.103, -10.760)	(-23.713, -15.970)	(-8.725, -4.218)	(-20.525, -12.529)
Reside in deprived district	**0.53**	**1.36**	**-1.013**	**-2.261**
	(-1.931, 2.992)	(-1.333, 4.053)	(-3.220, 1.194)	(-5.880, 1.358)
Reside in one of the 3 regions of the north	**-5.459*** **	**-7.203*** **	**-2.527* **	**-4.548* **
	(-7.758, -3.160)	(-10.001, -4.405)	(-4.833, -0.220)	(-8.092, -1.004)
Constant	53.781***	72.553***	56.620***	90.002***
	(48.915, 58.647)	(66.662, 78.444)	(51.850, 61.390)	(81.285, 98.718)
R^2	**0.158**	**0.257**	**0.143**	**0.286**
N	16,500	16,500	14,743	14,743

*$p < 0.1$; **$p < 0.01$; ***$p < 0.001$.

3.8 Proportion of Non-readers

In this section we present the results of the non-reader estimates for both 2011 and 2013, for comparison.

First, as stated in Section 2.9, in 2011, estimating the proportion of pupils likely to be non-readers involved examining scores on six P3 reading comprehension items (the P6 items were not considered reliable enough to produce a good estimate). Using Method 1, pupils who had 0 or 1 item correct out of 6 (i.e., a score lower than they could have obtained by responding randomly to the questions or 25%) were considered likely to be non-readers.

To illustrate what being 'non-reader' meant in 2011, below is an example of a P3 reading comprehension item (i.e., testing only basic Knowledge and Understanding cognitive abilities). Note that the answer is explicit in the text, such that a pupil with basic reading skills should have been able to answer the comprehension questions easily. Yet only 34% of the P3 public-school pupils answered this item correctly in 2013.

Ebo visited Uncle Ato at Cape Coast. His uncle is a fisherman. He saw his uncle fishing. The fishermen happily sang together while fishing. Ebo smiled to himself and said, "One day I will also learn to catch fish".

Where does Ebo's uncle live?

A **Chereponi**

B **Kade**

C **Cape Coast**

D Komenda

Table 22 shows the estimated proportion of non-readers at P3 by school type, using Method 1, for 2011.

Table 22: NEA 2011: Percentages of pupils likely to be non-readers

Grade and type of school	Proportion (95% confidence intervals)
P3 (6 items)	
Public schools	46.5 (44.7 – 48.3)
Private schools	20.6 (16.3 – 24.8)
All	42.4 (40.6 – 44.2)

Overall in 2011, 42.4% of the P3 pupils were likely to be non-readers and 20.6% of the students attending private schools were considered to be non-readers.

Again as noted in Section 2.9,the 2013 version of the NEA added more items testing reading comprehension (12 items for P3 and 16 items for P6, at the Knowledge and Understanding levels) and analysis of the language tests themselves showed high internal reliability. The data on non-readers from the 2013 NEA ***(Table 23)*** were very consistent with the 2011 estimates at P3.

Table 23*:* NEA 2013: Percentages of pupils likely to be non-readers

Grade and type of school	Proportion (95% confidence intervals)
P3 (12 items)	
Public	50.0 (48.1 – 52.1)
Private	22.8 (19.8 – 25.7)
All	44.4 (42.4 – 46.3)
P6 (16 items)	
Public	41.4% (38.9 – 43.7)
Private	11.8% (9.0 – 14.1)
All	35.2% (32.9 – 37.4)

The proportion of P6 non-readers identified from the 2013 analysis is noteworthy as 41.4% of the P6 students in public schools were considered to be non-readers. These proportions are consistent with the total (public and private schools combined) proportions of students failing to reach minimum competency in English (score <35%).

The results of the additional three methods used to cross-check the estimated percentages of non-readers in 2013 appear in ***Annex G.***

Finally, recall that parallel to the NEA 2013 administration, under the same USAID *Testing* project, teams were also separately administering a Ghana-adapted EGRA to a sample of over 800 schools (nearly 15,000 students) in 11 national languages plus English. Some of the key findings from the EGRA intersected with the EGRA-like items in the 2013 NEA. For example, although the definition of a non-reader in a standard EGRA administration is different from the one used for the EGRA-like items in the NEA (that is, for the EGRA instrument, pupils are classified as non-readers if they cannot read aloud a single word of a text passage), for P3 the NEA estimates yielded exactly the same figure as the EGRA-determined proportion of non-readers in English at the end of P2 in public schools: 50%. At P6, still 41.4% of the pupils were likely to be non-readers in public schools.

4. Conclusions and Recommendations

The NEA 2013 research team, including NEAU representatives, presented draft conclusions and recommendations at a dissemination workshop—the National Policy Forum—in Ghana in February 2014 for the Ministry of Education and other stakeholders. Feedback from these reviewers, as well as input from the earlier District Cluster Forums managed by the NEAU, was incorporated into the draft recommendations and used to create this section of the final version of the NEA Technical Report. As noted in the Executive Summary, the recommendations centred on instructional methods, teacher training and support, availability of teaching and learning materials, and parent and community involvement.

4.1 Reading Instructional Methods: How Pupils Learn vs. What Pupils Learn

Study results:

The EGRA study showed that when children have a grasp of some of the basic 'building blocks' in learning to read, such as understanding of letter sounds and the ability to decode or 'sound out' new words, they are more likely to be able to read fluently (e.g., quickly) and to understand what they read. Few pupils demonstrated these basic skills on the EGRA, in any of the languages assessed; and therefore it should not be surprising that few children in the EGRA study (on average, less than 2%) were able to read a passage fluently and with comprehension.

Recommendations:

Instructional methods for teaching reading need to shift from the conventional 'chalk and talk' methods to classroom instructional practices that focus on the critical components of successful literacy acquisition, so that students learn to read in the early grades. When teachers develop their students' oral language skills (e.g., phonological awareness and vocabulary) and teach the relationship between letters and sounds in a systematic and explicit fashion, their students have the foundation for success in recognising words and reading with comprehension. Ghanaian teachers need to be trained in these teaching methods through both in-service and pre-service programmes.

Reinvigorating the National Literacy Acceleration Programme (NALAP)[34] would be an important first step toward reaching this goal. Within the NALAP curriculum is a clearly stated timetable for literacy instruction. A reasonable literacy timetable should be established as policy. The timetable should incorporate shared and independent reading as appropriate, to ensure that students have time to practice their new skills.

4.2 <u>Mathematics</u> Instructional Methods: How Pupils Learn vs. What Pupils Learn

Study results:

On the EGMA study, students did reasonably well on the most procedural of items, such as the basic addition and subtraction facts. However, on the more conceptual items, there was a sharp drop-off in performance, with nearly 70% of the pupils unable to answer a single subtraction level 2 item correctly—the easiest of these being: 19 – 6 = ☐. This stark difference in performance between the more procedural and more conceptual subtasks suggests a lot about how children in Ghana are likely to experience school mathematics. That is, it is likely that they experience mathematics as a subject in which you have to know (remember) the answer rather than having a strategy for developing it; or as the memorisation of facts, rules and procedures rather than as a meaningful, sense-making, problem-solving activity.

Recommendations:

Either through their own resources or with the assistance of technical experts, the Ministry of Education and the GES need to identify effective, evidence-based practices regarding the teaching of early grade mathematics. Such an approach would support the acquisition of foundational mathematics *and* reading skills, with an emphasis on students' conceptual understanding.

Once the Ministry and the GES have established an evidence-based approach to teaching early grade mathematics for Ghana, attention should shift to implementing the approach.

The implementation should be achieved through both in-service and pre-service teacher training programmes. Teachers need to receive specific training on how to teach mathematics in the early grades. In addition, suitable learning materials need to be developed.

[34] As summarised in the 2013 EGRA/EGMA analysis report, "NALAP provides for instruction in the predominant Ghanaian language of the local community through grade 3, with English introduced gradually in the early grades, and pupils making the full transition in grade 4. By grade 4 the programme assumes pupils will have first become fluent speakers and readers of the Ghanaian language of instruction, followed by English" (p. 2).

4.3 Teacher Development and Management

Study results:

International research on learning shows that students who attend classrooms where teachers are qualified, engaged with their pupils, and well supervised are more likely to do well in school. In many countries, schools in remote regions and early grade classrooms tend to have fewer qualified teachers than in urban areas and in upper primary classrooms. In this case, the study findings demonstrated that students in urban settings consistently outperformed students in rural settings, especially in the three regions of northern Ghana.

Teachers and teaching practice, teacher management, supervision and support, and teacher placement and incentives were the subjects of much discussion at the National Policy Forum. The importance of addressing these issues was the most prominent of the recommendations emerging from the policy dialogue.

Recommendations:

Districts need to establish systems for regular school, teacher and student performance monitoring, tied to clear performance targets. Data from such district monitoring systems should be used to inform school- and district-wide interventions for improving student learning outcomes.

District and school management should work together to ensure that sufficient supervision or coaching is available to assist teachers as they learn and apply new and effective teaching methods. To this end, training on effective coaching for circuit supervisors and head teachers is needed. Budget allocations to support regular school coaching visits by district circuit supervisors must be provided for.

Policy related to placement and distribution of teachers in primary schools should be reviewed and improved upon to better support early grade literacy and numeracy attainment. Qualified teachers are needed in the lower primary grades and in rural areas. Furthermore, whenever and wherever possible, teachers placed in the lower primary classrooms should be fluent and literate in the language of learning and instruction of their placement school. To support this, it is recommended that Ghanaian language pedagogy (e.g., teaching reading and mathematics in the local language and bridging to English in the mid-primary grades per Ghanaian language policy) be considered as a required, examinable course in the Colleges of Education.

4.4 Time to Practice and the Availability of Materials

Study results:

The EGRA and EGMA studies showed that the majority of pupils had an exercise book, but less than half of the children had an English or mathematics textbook. Fewer pupils (only approximately 35%) had a supplemental reader, and only 20% of the pupils reported that they were able to take materials home for practice. Study findings showed that the few children who were able to read with comprehension had access to materials and also practiced reading at school and at home.

Recommendations:

Unless pupils gain the basic reading and mathematics skills in the early grades and are given ample opportunity to practice, they will fall farther and farther behind in school in the later years. Reading and mathematics textbooks and supplemental materials that children can take home to practice are important for children's learning.

The GES, District and Regional Education Officers, the District Assembly, and head teachers should work together to ensure that all students have textbooks in school and are allowed to take texts home.

The GES, District and Regional Officers, the District Assembly and District Education Oversight Committee, School Committees, parent-teacher associations, community members, religious organisations, language bureaus, nongovernmental organisations, and the private sector should come together to contribute to building a strong base of supplementary readers for students to use at school and at home for independent reading and practice—in local languages and in English. Establishing 'classroom book boxes' and reinvigorating the community library are examples of what could be done.

4.5 Parental Involvement

Study results:

The study findings showed that pupils who had higher learning outcomes on the national assessments were more likely to: attend school regularly; have books to take home and use for practice; have homework assignments that are graded by teachers; and have someone at home who helps them with their homework (e.g., someone to read to or do mathematics problems with).

Recommendations:

As part of the GES/NEAU nationwide dissemination programme from November 2013 to January 2014, District Advocacy Teams (DATs) were established to champion advocacy for children's learning, in districts and communities. Support from the Regional and District Education Offices and the District Assemblies to keep these DAT teams active is recommended.

Parents and communities should work closely with teachers and schools to develop and implement programmes to support children's learning at home and in the community. A few ideas from participants included: (1) Organise community-level parent advocacy groups focusing specifically on supporting early grade learning in reading and mathematics; (2) encourage regular school attendance by all children; (3) organise storytelling, shared reading, and after-school programmes; and (4) strongly encourage parents to visit schools often, meet with teachers and discuss how they can help their child at home.

Annex A: Population Size and Sample Weights

This annex supplements the information about sample weighting presented in Section 2.2. As mentioned there, schools were stratified by region and sampled with equal probability. All P3 and P6 students present on the day of assessment were assessed. Thus, the sample methodology was a stratified one-stage cluster sample of schools. The population of interest was P3 and P6 pupils who were present on the day of the assessment and were attending a school with P3 enrolment and P6 enrolment larger than 10 pupils.

Sample Weights

The sample weights were created by dividing the total number of primary schools (which had at least ten P6 and ten P3 students) by the number of schools that were sampled and that completed the NEA tests in each region.

$$Wt_{Sample} = \frac{Total\ schools\ (by\ region)}{Sampled\ schools\ completed\ NEA\ (by\ region)}$$

For example, there were 2,923 primary schools in the Asante Region and 54 sampled schools in Asante that completed the NEA. Thus, the sample weight for Asante was 54.13 (2,923/54 = 54.13). Because the Upper West Region had only 467 schools, its sample weight was much smaller at 8.49 (467/55= 8.49). In other words, each school that completed the NEA in Asante represented 54.13 schools, whereas all 55 schools that completed the NEA in Upper West were equivalent to only 8.49 schools after weighting.

Table A1 provides the unweighted and weighted counts and percentages for schools, showing how the weights affect the data. Again using Asante as an example, it had 2,923 schools in the region. The weighted value of the combined 54 Asante schools that completed the NEA was 2,923, or 18.76% of all the sampled schools. Similarly, Upper West's 55 schools that completed the NEA had a weighted value of 467, or 3% of all the sampled schools.

Table A1: Schools: Unweighted and weighted counts and percentages by region

Region	Unweighted number and percentage of schools	Weighted number and percentage of schools
Asante	54 (9.85%)	2,923 (18.76%)
Brong Ahafo	55 (10.04%)	1,709 (10.97%)
Central	55 (10.04%)	1,639 (10.52%)
Eastern	55 (10.04%)	1,821 (11.69%)
Greater Accra	55 (10.04%)	1,542 (9.9%)

Region	Unweighted number and percentage of schools	Weighted number and percentage of schools
Northern	54 (9.85%)	1,703 (10.94%)
Upper East	55 (10.04%)	609 (3.91%)
Upper West	55 (10.04%)	467 (3%)
Volta	55 (10.04%)	1,355 (8.7%)
Western	55 (10.04%)	1,809 (11.61%)
Total	**548**	**15,577**

Because all P3 and P6 pupils present on the day of assessment took the test, each completed P3 and P6 test received the same weight as in the school weight discussion above. ***Tables A2 and A3*** provide the unweighted and weighted counts and percentages for P3 pupils and P6 pupils. The 1,787 completed P3 test forms in Asante represented 96,730 P3 pupils (18.5% of the total P3 population); the 2,489 completed P3 test forms in Upper West represented only 29,786 pupils (5.7% of the total P3 population).

Table A2: P3 pupils: Unweighted and weighted counts and percentages, by region

Region	Unweighted number and percentage of P3 pupils	Weighted number and percentage of P3 pupils
Asante	1,787 (9.18%)	96,730 (18.5%)
Brong Ahafo	2,049 (10.53%)	63,668 (12.18%)
Central	1,462 (7.51%)	43,568 (8.33%)
Eastern	1,731 (8.9%)	57,312 (10.96%)
Greater Accra	2,224 (11.43%)	62,353 (11.92%)
Northern	1,661 (8.54%)	52,397 (10.02%)
Upper East	2,690 (13.82%)	29,786 (5.7%)
Upper West	2,489 (12.79%)	21,134 (4.04%)
Volta	1,782 (9.16%)	43,902 (8.4%)

Region	Unweighted number and percentage of P3 pupils	Weighted number and percentage of P3 pupils
Western	1,583 (8.14%)	52,066 (9.96%)
Total	**19,458**	**522,916**

Table A3: P6 pupils: Unweighted and weighted counts and percentages, by region

Region	Unweighted number and percentage of P6 pupils	Weighted number and percentage of P6 pupils
Asante	1,555 (8.91%)	84,172 (17.8%)
Brong Ahafo	1,792 (10.27%)	55,682 (11.78%)
Central	1,466 (8.4%)	43,687 (9.24%)
Eastern	1,559 (8.94%)	51,617 (10.92%)
Greater Accra	2,145 (12.29%)	60,138 (12.72%)
Northern	1,521 (8.72%)	47,981 (10.15%)
Upper East	2,246 (12.87%)	24,869 (5.26%)
Upper West	2,128 (12.2%)	18,069 (3.82%)
Volta	1,609 (9.22%)	39,640 (8.38%)
Western	1,426 (8.17%)	46,902 (9.92%)
Total	**17,447**	**472,757**

Annex B: Item-Level Evaluation

Primer: How to Read Outputs of the Rasch Model

The example Wright plot below ***(Figure B1)*** is an illustration of the interval-level properties enforced on the data by the Rasch model that allow items to be examined in relation to the persons taking those items. The ‘Measure’ axis provides the logit scale on which both persons and items are placed. Students are depicted on the left side (‘#’ marker) of the chart and the items are shown on the right (‘X’ marker). This plot can be seen as an ability continuum (for persons), and difficulty continuum (for items), with more able persons as well as more difficult items positioned near the top of the chart. In an ideal instrumentation situation, a normal distribution of individuals would appear on the left with a flat spread of one to two items for each person ability level. The item and person means (‘M’ highlighted in turquoise) are expected to be ‘targeted,’ or in close alignment. The ‘S’ represents the designation of one standard deviation from the mean, suggesting less alignment between person ability and difficulty level.

In addition, item spread is very important in that there is at least one representative item for each place in the ability distribution. Evidence of ‘stacking’ (yellow highlight), or a redundancy of measurement, indicates the presence of more than one item at a given level of the ability distribution (or level of effort). Ideally, each targeted skill should be represented by only one or two items, with more potentially affecting performance by unnecessarily fatiguing examinees with redundant items which do not provide any new information. An examination of item stacking is particularly useful for reducing the number of items or creating item banks for instruments that will be used over time.

Finally, discussions of ‘unexpected responses’ refer to items that are not eliciting the responses expected, given the relative difficulty of the item. Such ‘unexpected responses’ are often outliers, but can also identify situations where, for example, high-ability examinees respond incorrectly to an easy item or lower-ability respondents are successful in getting a difficult item correct. These situations can flag validity problems with the items such as ambiguous language in the test question itself.

The remainder of this annex presents the Wright plots from the Rasch model outputs for each subject, grade, and form of the 2013 NEA. Each figure is followed by a summary of the important findings.

Figure B1: Sample Wright plot of ‘persons’ vs. test items

```
MEASURE                 |                MEASURE
 <more> --------------------- PERS  -+- ITEMS  --------------------- <rare>
  3               ### S+S X                3
                .###### | X
               ######## | X
  2          ############# + XX                2
            ############## | XXX
            .############## | XX
  1   .########################### + XX                1
         .################### M| X
            .############## | XXX
  0       .################### +M XXXX               0
          .################# | X
      .######################### | XXX
  -1      .#################### + XXXX              -1
            .############## | XX
          ################### | XX
  -2          ############# + X                -2
              .########## | XXXXX
               .####### S| XXXXXXX
  -3             .####### + XXXXX              -3
               .######## | XXX
                .###### | XXX
  -4            .######## +S X                -4
 <less> --------------------- PERS  -+- ITEMS  ------------------<frequent>
 EACH '#' IN THE PER COLUMN IS 5 PERS; EACH '.' IS 1 TO 4
```

Figure B2: English P3 Form 1 – Cronbach’s alpha (KR20) person raw score ‘test’ reliability = .86

```
MEASURE  PERSON - MAP - ITEM
        <more>|<rare>
  4       . +
           |
           |
         .# |
           |
           |
           |
  3        +
        .## |
           |
           |
        .## |
           |
           |
  2      .## T+
           |
        .### |
        .### |
           | c13e
        .### |T
        .### |
  1      .## + c26e
          S|
        .### |S c14e  c29e
        .### | c25e
        .### | c15e  c17e  c18e  c21e  c28e
       .#### | c16e  c23e
       .#### | c10e  c20e  c27e  c30e
  0    .##### +M c19e  c22e  c24e  c8e
      .###### |
    .######## M| c11e  c12e
   .########## | c7e
   .########### | c3e
  .############ |S c4e  c5e  c6e
  .############ |
  -1         + c1e
  .############ |
   .########## |T
          S| c9e
    .######## |
      .##### | c2e
           |
  -2     .### +
           |
           |
         .# |
          T|
           |
         . |
  -3        +
           |
           |
           |
         . |
           |
           |
  -4       . +
        <less>|<frequent>
 EACH ‘#’ IS 63: EACH ‘.’ IS 1 TO 62
```

- The reliability for this assessment ***(Figure B2),*** 0.86, is well above the conventional cut-off for acceptable reliability (0.70). Therefore this assessment has adequate internal consistency.
- The distribution of students is mostly normal, but with a skew toward the lower end of the ability range. However, there is a near perfect alignment of the student ability and item difficulty means (turquoise highlight).
- There is evidence of measurement redundancy (yellow highlight), where items are replicating the same level of effort for participants; however, they could be from different subscale sections of the assessment.
- There are two items not performing as expected, or showing misfit to the model (c13e, c26e). In this case, it appears that several moderate-ability students responded incorrectly, but several low-ability students were able to provide a correct response.
- One assumption of Rasch measurement is that there is only a single construct in an assessment (that is, a maths test is made of only mathematics items, not listening comprehension). Another assumption of the Rasch model, tied to the single construct measurement, is that all items would be locally independent, or uncorrelated, after accounting for the single dimension they share. Based upon further examination of additional output (not shown), this assessment meets both assumptions of uni-dimensionality and local independence.

Figure B3: English P3 Form 2 – Cronbach's alpha (KR20) person raw score 'test' reliability = .86

```
MEASURE  PERSON - MAP - ITEM
        <more>|<rare>
  4       . +
            |
            |
         .# |
            |
            |
            |
  3         +
            |
         .# |
            |
            |
        .## |
            |
  2      .## +
          T|
        .## |
            |
       .### |
       .### |T
        .## |
  1         +
       .### | c12e  c18e  c30e
       .### S| c21e  c25e  c9e
       .### |S
       .### | c15e
      .#### | c16e  c17e  c19e  c24e  c27e  c29e  c8e
      .#### | c26e
  0     .##### +M c23e  c28e  c7e
     .###### | c14e  c22e
     .####### | c10e  c11e  c20e  c6e
    .######### M|
   .########### |S
   .############ | c3e  c5e
  .############# |
 -1         + c2e  c4e
  .############# |
   .############ |T c13e  c1e
    .######### S|
            |
     .###### |
            |
 -2      .### +
            |
         .# |
            |
          T|
            |
          . |
 -3         +
            |
            |
            |
          . |
            |
            |
 -4       . +
        <less>|<frequent>
 EACH '#' IS 63: EACH '.' IS 1 TO 62
```

- The reliability for this Form 2 assessment ***(Figure B3),*** 0.86, is also well above the conventional cut-off for acceptable reliability (0.70). Therefore this assessment has adequate internal consistency.
- The distribution of students is mostly normal, but with a slight skew toward the lower end of the ability range. However, there is close alignment of the student ability and item difficulty means (turquoise highlight).
- There is evidence of measurement redundancy (yellow highlight); however, the items could be from different subscale sections of the assessment.
- There are two items not performing as expected, or showing misfit to the model (c12e, c30e). In this case, it appears that several moderate-ability students responded incorrectly, but many low-ability students were able to provide a correct response.
- Based upon further examination of additional output (not shown), this assessment meets both assumptions of uni-dimensionality and local independence.

Figure B4: English P6 Form 1 – Cronbach's alpha (KR20) person raw score 'test' reliability = .90

```
MEASURE   PERSON - MAP - ITEM
        <more>|<rare>
  4       . +
         . |
           |
           |
           |
           |
           |
         . |
  3        +
           |
         .# |
           |
           |
         .# |
           |
        .## T|
  2        +
        .## |
        .## |
           | c21e
        .## |
        .## |T
       .### | c2e
        .## |
  1        S+ c31e
       .### | c22e  c40e
       .### |S c18e  c32e
     .###### | c27e  c38e  c8e
       .### | c35e
       .### | c39e
       .### | c16e  c9e
       .### | c15e  c26e  c28e  c34e
  0    .### +M c24e  c25e  c29e  c33e  c36e
       .### M| c1e
      .#### | c10e  c19e  c30e
       #### | c17e  c37e
      ##### | c23e
     .##### | c11e  c7e
  .############ |S c4e
           | c12e  c13e  c3e  c6e
 -1    .###### + c20e
     .###### |
      .##### S| c14e  c5e
      .##### |T
       .### |
           |
        .## |
         .# |
 -2        +
         . |
           |
         . T|
           |
           |
         . |
           |
 -3        +
        <less>|<frequent>
 EACH '#' IS 72: EACH '.' IS 1 TO 71
```

- This assessment ***(Figure B4)*** has adequate internal consistency (reliability = 0.90).
- The distribution of students is fairly flat, but there is close alignment of the student ability and item difficulty means (turquoise highlight).
- There is evidence of measurement redundancy (yellow highlight); however, the items could be from different subscale sections of the assessment.
- There are four items showing misfit (c2e, c18e, c21e, c22e). It appears that several moderate-ability students were responding incorrectly, but many low-ability students were able to provide a correct response.
- This assessment meets both assumptions of uni-dimensionality and local independence.

Figure B5: English P6 Form 2 – Cronbach's alpha (KR20) person raw score 'test' reliability = .90

```
MEASURE  PERSON - MAP - ITEM
        <more>|<rare>
  4        . +
          . |
            |
            |
            |
            |
          . |
  3          +
            |
          .# |
            |
          .# |
           T|
         .## |
  2          +
         .## |
         .## | c22e
         .## |
        .### |T
            |
        .### S|
  1     .###### + c2e  c32e  c35e
        .### | c27e  c36e  c8e
         .## |S c17e  c21e
        .### | c39e
         ### | c31e
      .###### | c26e  c30e
        .### | c16e  c1e  c24e  c28e  c33e  c37e  c38e  c9e
  0      #### M+M c15e  c40e
        .### | c10e  c11e  c25e
    .######### | c18e  c23e  c34e
       .##### | c29e
       .##### | c3e  c7e
      .####### |S
      .###### | c12e  c13e  c19e
 -1     .###### + c20e  c4e  c5e
  .############ S| c14e
            | c6e
       .##### |T
        #### |
         .## |
          .# |
 -2          +
          . |
           T|
          . |
            |
          . |
            |
 -3          +
          . |
            |
            |
            |
            |
            |
 -4          +
        <less>|<frequent>
 EACH '#' IS 69: EACH '.' IS 1 TO 68
```

- This assessment ***(Figure B5)*** has adequate internal consistency (reliability = 0.90).
- The distribution of students is mostly flat, but slightly curves toward the bottom end of the ability distribution. However there is perfect alignment of the student ability and item difficulty means (turquoise highlight).
- There is evidence of measurement redundancy (yellow highlight); however, the items could be from different subscale sections of the assessment.
- There are four items showing misfit (c2e, c17e, c21e, c22e). It appears that several moderate-ability students were responding incorrectly, but many low-ability students were able to provide a correct response.
- Based upon further examination of additional output (not shown), this assessment meets both assumptions of uni-dimensionality and local independence.

Figure B6: Mathematics P3 Form 1 – Cronbach’s alpha (KR20) person raw score ‘test’ reliability = .79

```
MEASURE  PERSON - MAP - ITEM
        <more>|<rare>
  4        . +
            |
            |
          . |
            |
            |
            |
  3         +
          . |
            |
            |
          . |
            |
            |
  2       .# +
            |
          .# |
          .# |
           T|
        .## |T c23m
        .### |
  1     .### + c19m  c24m
            | c20m
        .### |S
       .#### |
       .#### S| c22m  c25m  c27m  c30m  c7m
       .##### | c13m  c15m  c28m  c29m  c6m
      .###### | c26m  c8m
  0    .###### +M c10m  c11m
     .######## | c3m
     .######### | c9m
    .########## M| c17m  c5m
    .########### | c16m  c21m  c2m
   .############ |S c12m  c14m
   .############ |
  -1           + c4m
   .############ |
    .########### |T c18m
           S|
     .######## | c1m
       .##### |
            |
  -2      .### +
            |
           T|
          .# |
            |
            |
          . |
  -3         +
            |
            |
            |
          . |
            |
            |
  -4         . +
        <less>|<frequent>
 EACH ‘#’ IS 64: EACH ‘.’ IS 1 TO 63
```

- The reliability for this assessment ***(Figure B6),*** 0.79, is above the conventional cut-off for acceptable reliability (0.70). Therefore this assessment has adequate internal consistency.
- The distribution of students is normal with a slight skew toward the lower end of the ability range. The student ability and item difficulty means (turquoise highlight) are off-center by about a half of a standard deviation.
- There is some evidence of measurement redundancy (yellow highlight), where items are replicating the same level of effort for participants; however, they could be from different subscale sections of the assessment.
- There are two items not performing as expected, or showing misfit to the model (c19m, c23m). It appears that a couple of high-ability students responded incorrectly, but many moderate- to low-ability students were able to provide a correct response.
- Based upon further examination of additional output (not shown), this assessment meets both assumptions of uni-dimensionality and local independence.

Figure B7: Mathematics P3 Form 2 – Cronbach's alpha (KR20) person raw score 'test' reliability = .80

```
MEASURE  PERSON - MAP - ITEM
        <more>|<rare>
  4       . +
            |
            |
          . |
            |
            |
            |
  3         +
          . |
            |
            |
            |
          . |
            |
  2      .# +
            |
         .# |
            |
        .## T|
        .## |T
        .## | c24m
  1         + c18m  c23m
       .### | c19m
       .### |
      .#### |S c25m
      .#### S| c15m  c27m  c30m
      .#### | c10m  c16m  c22m  c26m  c28m  c29m  c6m
     .##### |
  0  .####### +M c12m  c7m  c8m
    .######## |
    .######### | c17m  c3m  c9m
   ########### M| c4m
   .########### |S c1m
   .########### | c11m  c13m  c14m  c21m
  .############ |
 -1         + c5m
  .############ | c20m  c2m
   ########### |T
    .######## S|
            |
     .##### |
            |
 -2      .### +
            |
         .# T|
            |
            |
            |
          . |
 -3         +
            |
            |
            |
          . |
            |
            |
 -4         . +
        <less>|<frequent>
 EACH '#' IS 63: EACH '.' IS 1 TO 62
```

- The reliability for this assessment ***(Figure B7),*** 0.80, indicates that it has adequate internal consistency.
- The distribution of students is normal with a slight skew toward the lower end of the ability range. Alignment of the student ability and item difficulty means (turquoise highlight) is off-center by about a standard deviation.
- There is evidence of measurement redundancy (yellow highlight), where items are replicating the same level of effort for participants; however, they could be from different subscale sections of the assessment.
- There are two items not performing as expected, or showing misfit to the model (c18m, c24m). It appears that many moderate- to low-ability students were able to provide a correct response to these difficult items.
- Based upon further examination of additional output (not shown), this assessment meets both assumptions of uni-dimensionality and local independence.

Figure B8: Mathematics P6 Form 1 – Cronbach’s alpha (KR20) person raw score ‘test’ reliability = .70

```
MEASURE  PERSON - MAP - ITEM
           <more>|<rare>
  4              +
                 |
                 |
                 |
              .  |
                 |
  3              +
              .  |
                 |
              .  |
              .  |
                 |
  2           .  +T
              .  |
              .  | c23m
              .  | c21m  c24m  c29m
              .  |
              .  |
  1           .  +S c14m  c28m  c8m
             .# T| c10m  c20m  c22m  c39m
             .#  | c25m
             .#  | c12m  c30m  c32m  c33m  c36m
           .###  | c15m  c34m  c5m
            .## S|
  0        .###  +M c18m  c26m
      .########  | c11m  c13m  c7m
         .#####  | c31m
         .#####  | c17m  c27m  c35m  c40m
  .############ M| c4m
        .######  | c19m
 -1      .#####  +S c37m  c6m
         .#####  | c38m
          .#### S| c2m  c3m
           .###  |
            .##  | c16m
             .#  |
 -2           . T+T
                 |
              .  | c9m
                 |
              .  |
                 | c1m
 -3           .  +
                 |
                 |
              .  |
                 |
                 |
 -4              +
              .  |
                 |
                 |
                 |
                 |
 -5              +
           <less>|<frequent>
 EACH ‘#’ IS 115: EACH ‘.’ IS 1 TO 114
```

- The reliability for this assessment ***(Figure B8),*** 0.70, just makes the cut-off for adequate internal consistency.
- The distribution of students is compact and almost normal. The student ability and item difficulty means (turquoise highlight) are off-center by more than a half of a standard deviation.
- There is evidence of measurement redundancy (yellow highlight); however, the items could be from different subscale sections of the assessment.
- There is one item showing misfit to the model (c23m). It appears that many moderate- to low-ability students were able to provide a correct response to this difficult item.
- Based upon further examination of additional output (not shown), this assessment meets both assumptions of uni-dimensionality and local independence.

Figure B9: Mathematics P6 Form 2 – Cronbach's alpha (KR20) person raw score 'test' reliability = .71

```
MEASURE  PERSON - MAP - ITEM
        <more>|<rare>
  4         +
            |
            |
            |
          . |
            |
  3         +
          . |
            |
          . |
          . |
            |T
  2        . +
          . |
          . | c22m
          . | c14m  c23m  c29m
          . |
          . | c21m  c25m
  1       .# +S c24m  c28m  c8m
          . T| c12m  c19m  c36m
          . | c39m
         .# | c10m  c30m  c32m  c33m
        .### | c15m  c26m
        .## | c34m  c4m
  0     .### S+M c13m
     .######## | c16m  c7m
       .##### | c11m  c18m  c27m
       .##### | c31m  c35m
   .############ M| c20m  c37m  c40m
      .###### | c5m
 -1     .##### +S
       .##### | c38m
       .#### S| c2m
        .### | c1m  c6m
        .## |
         .# | c17m
 -2          T+
          . |T
          . | c9m
            | c3m
          . |
            |
 -3        . +
            |
            |
          . |
            |
            |
 -4          +
          . |
            |
            |
            |
            |
 -5        . +
        <less>|<frequent>
 EACH '#' IS 107: EACH '.' IS 1 TO 106
```

- The reliability for this assessment ***(Figure B9),*** 0.71, just makes the cut-off for adequate internal consistency.
- The distribution of students is compact and almost normal. The student ability and item difficulty means (turquoise highlight) are off-center by almost a standard deviation.
- There is evidence of measurement redundancy (yellow highlight); however, the items could be from different subscale sections of the assessment.
- There is one item showing misfit (c25m). It appears that a many moderate- to low-ability students were able to provide a correct response to this moderately difficult item.
- Based upon further examination of additional output (not shown), this assessment meets both assumptions of uni-dimensionality and local independence.

Annex C: Test Linking

As pointed out in the main text of this report, in 2013 the NEA instrument and the administration procedures were modified to address limitations in the 2011 administration. As a result, scores from 2011 and 2013 were expressed on different score scales. This annex explains how we used equating methods to link the 2011 and 2013 test forms to a common scale, so as to facilitate score comparisons across administrations.

In the analyses below, we used a *non-equivalent groups* equating design to link the 2011 and 2013 test forms. In this type of design, examinees are assigned a test form by group. These groups are not assumed to have the same ability or background characteristics, however. Instead, to control for differences across groups, students are also assessed on what is referred to as an *anchor test* or *subscale:* a subset of test items that is common across the forms. Scores on this anchor test are used to statistically control for ability differences across groups so as to detect any differences in score distributions that can be attributed to form difficulty. ***Table C1*** contains descriptive statistics for the 2011 and 2013 complete tests as compared to the anchor test.

Table C1: Total and anchor test descriptive statistics

			Total test					Anchor test			
Year	**Gr**	**Sub**	**Ave%**	**M**	**SD**	**Skew**	**Kurt**	**M**	**SD**	**Skew**	**Kurt**
2011	3	Eng	44	17.69	7.36	0.86	3.18	1.55	1.19	0.48	2.35
		Maths	36	14.55	6.4	1.04	3.97	1.74	1.31	0.63	2.8
	6	Eng	50	29.93	10.54	0.41	2.3	1.52	1.04	0.04	1.83
		Maths	38	23.07	8.59	0.82	3.52	1.83	1.28	0.57	3.05
2013	3	Eng	43	12.86	6.45	0.77	2.73	1.65	1.28	0.41	2.1
		Maths	40	12.04	5.4	0.62	2.91	1.97	1.42	0.44	2.35
	6	Eng	48	19.19	8.64	0.44	2.09	1.63	1.03	-0.09	1.83
		Maths	38	15.16	5.02	0.6	.63	1.86	1.27	0.56	3.08

Notes: Gr = grade; Sub = subject; Ave% = average percent correct; M = mean; SD = standard deviation; Skew = skewness; Kurt = kurtosis.

Equating statistics procedures—which were applied to link the 2011 and 2012 tests—are designed to convert a scale based on one score distribution, referred to as *X*, to a scale from another score distribution, referred to as *Y*. The purpose of equating is to find the score on *Y* that students taking *X* would have received had they taken *Y*. In this study the question was, how would students in the 2011 administration have performed on the 2013 assessment? Various methods exist for performing the necessary score conversions.[35] Four types of equating were conducted in this study: mean, linear, circle-arc, and equipercentile equating.

[35] For details, see Kalen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer Science + Business Media. http://dx.doi.org/10.1007/978-1-4757-4310-4

Table C2 contains equating statistics for P3, and ***Table C3*** for P6. The first two rows in each table include descriptive statistics for 2011 and 2013 raw score distributions. The remaining rows include descriptive statistics for the equated distributions according to each equating type and method. The tables demonstrate the focus of the different methods. The **mean** equating methods adjust only for the mean. The **linear** methods adjust both the mean and the standard deviation, but not other distribution characteristics such as measures of skewness and kurtosis. The remaining non-linear methods adjust for the mean, standard deviation, skewness and kurtosis. The frequency estimation euipercentile method was used to calculate to link the 2011 and 2013 score distributions, as discussed in Section 3.3 above. This row is highlighted in Table C2 and C3 below.

Table C2: P3 equating statistics, by type and method

	English				Maths			
Source	**M**	**SD**	**Skew**	**Kurt**	**M**	**SD**	**Skew**	**Kurt**
2011	17.69	7.36	0.86	3.18	14.55	6.40	1.04	3.97
2013	12.86	6.45	0.77	2.73	12.04	5.40	0.62	2.91
Mean: Tucker	12.40	7.36	0.86	3.18	11.29	6.40	1.04	3.97
Mean: Levine	12.00	7.36	0.86	3.18	10.55	6.40	1.04	3.97
Linear: Tucker	12.46	6.22	0.86	3.18	11.40	5.17	1.04	3.97
Linear: Levine	12.16	5.71	0.86	3.18	10.89	4.61	1.04	3.97
Circle: Tucker	8.88	4.92	1.21	3.94	8.31	4.26	1.31	4.78
Circle: Chain	9.18	4.96	1.18	3.88	8.73	4.34	1.27	4.63
Equip: FE	12.86	6.45	0.77	2.74	12.03	5.40	0.63	2.91
Equip: Chain	12.28	5.95	0.88	3.19	11.19	4.93	0.80	3.51

Notes: M = mean; SD = standard deviation; Skew = skewness; Kurt = kurtosis; Circle = circle-arc; Equip = equipercentile; FE = frequency estimation.

Table C3: P6 equating statistics, by type and method

	English				Maths			
Source	**M**	**SD**	**Skew**	**Kurt**	**M**	**SD**	**Skew**	**Kurt**
2011	29.93	10.54	0.41	2.30	23.07	8.59	0.82	3.52
2013	19.19	8.64	0.44	2.09	15.16	5.02	0.60	3.63
Mean: Tucker	18.43	10.54	0.41	2.30	15.04	8.59	0.82	3.52
Mean: Levine	17.73	10.54	0.41	2.30	14.85	8.59	0.82	3.52
Linear: Tucker	18.54	8.65	0.41	2.30	15.08	5.03	0.82	3.52
Linear: Levine	17.94	8.71	0.41	2.30	14.96	5.10	0.82	3.52
Circle: Tucker	10.19	5.65	0.76	2.69	9.13	4.47	1.21	4.62
Circle: Chain	10.76	5.71	0.74	2.66	9.21	4.49	1.21	4.60
Equip: FE	19.18	8.64	0.44	2.10	15.16	5.03	0.63	3.69
Equip: Chain	18.35	8.54	0.57	2.26	15.04	5.06	0.63	3.64

Notes: M = mean; SD = standard deviation; Skew = skewness; Kurt = kurtosis; Circle = circle-arc; Equip = equipercentile; FE = frequency estimation.

Annex D: Analysis of Score Distribution

The findings in this report were presented in terms of average test score and percentage of pupils reaching a certain level. It is standard to report data on test score distribution in order to get a better picture of how pupils perform on the test. This annex presents the analyses of score distribution, comparing performance across grades and subject.

The analyses also include the proportion of pupils reaching a certain level when considering other cut scores than the ones used in the report. While the NEA official cut scores are 35% (minimum competency) and 55% (proficiency), other thresholds can be considered, for example: (1) 25%, the score that a pupil would get by answering randomly to the test; and (2) 70% for proficiency (based on international literature and curricula objectives).

First, it is useful to consider the spread of data across subjects and grades, which is represented in ***Table D1*** by the standard deviation of scores.

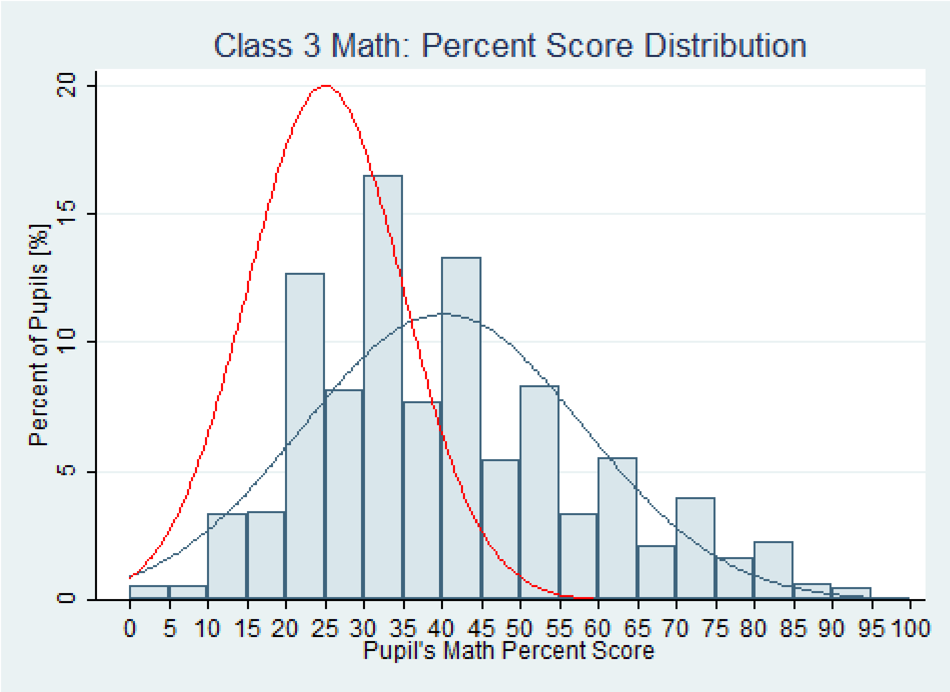
Table D1: Total scores, by grade and subject (% correct answers)

Subject	P3				P6			
	Score	**CI**		**SD**	**Score**	**CI**		**SD**
Maths	**41.1**	40.0	42.2	18.3	**38.2**	37.4	39.0	12.4
English	**44.4**	42.8	45.9	21.9	**48.9**	47.3	50.6	21.7

Notes: CI = confidence interval; SD = standard deviation.

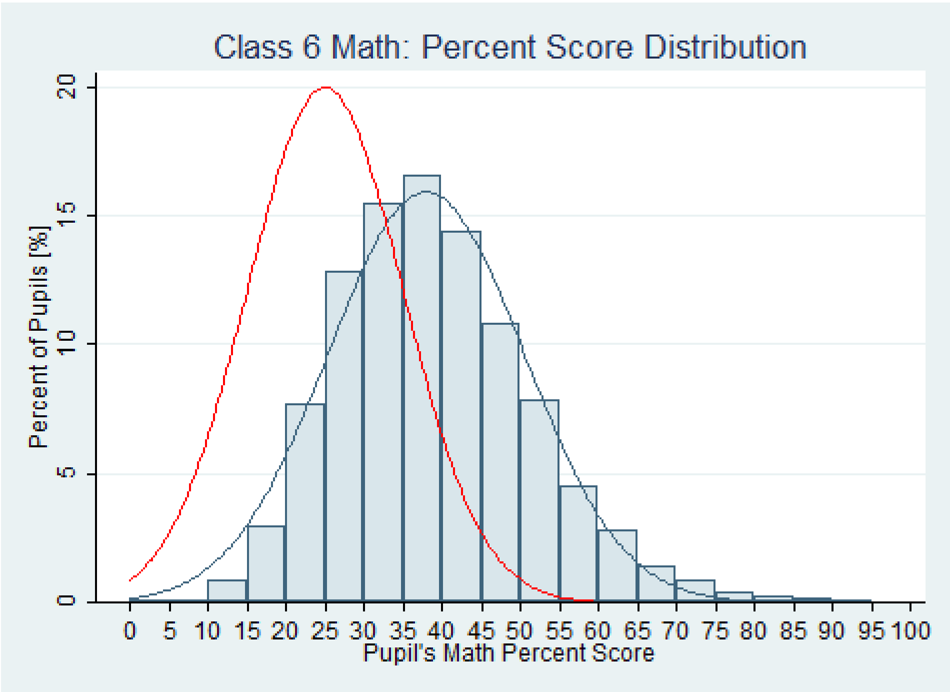
The standard deviation is slightly larger in English than in mathematics, where there was little variation among pupils' performance and a more normal distribution (i.e., in mathematics). The distribution was somewhat flat for English and slightly skewed by a small percentage of very high performers scoring between 80% and 100% correct. To have a better sense of how pupils performed, it is useful to plot the histogram (see ***Figure D1***), the normal distribution (blue curve), and the distribution of a simulated group of pupils if they were to respond randomly to the test (red curve).

Figure D1: Distribution of maths scores, P3



In maths in P6, there was less variance than in P3, and only a few pupils had more than 70% correct answers ***(Figure D2).***

Figure D2: Distribution of maths scores, P6



In English, the situation was quite different (see ***Figures D3 and D4***). The curves deviated from a normal distribution with two groups of pupils: low-average performers and high performers. A significant proportion of students did answer more than 70% of answers correctly in English.

Figure D3: Distribution of English scores, P3

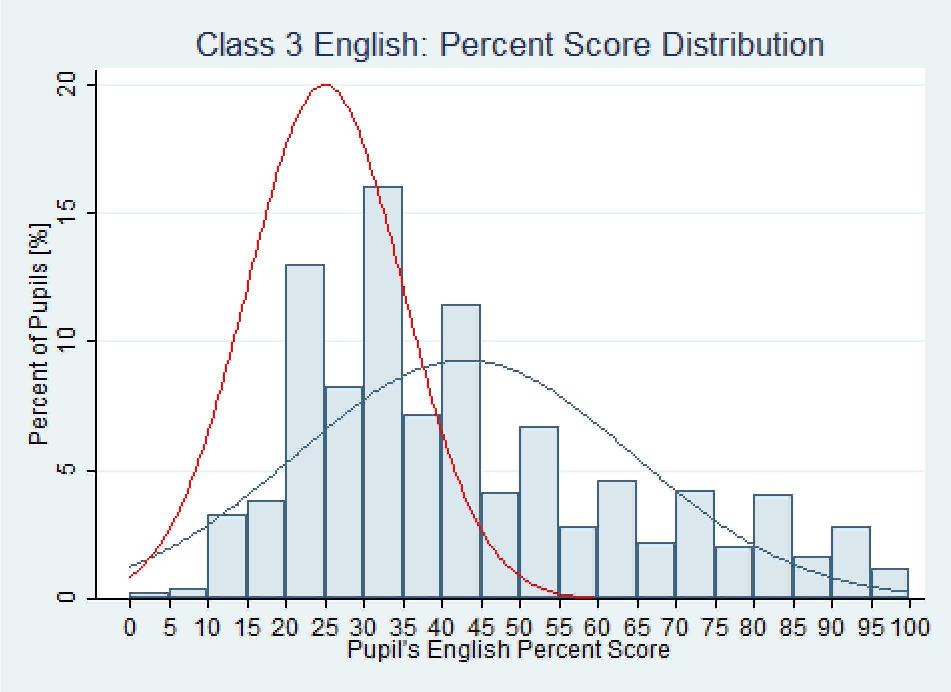
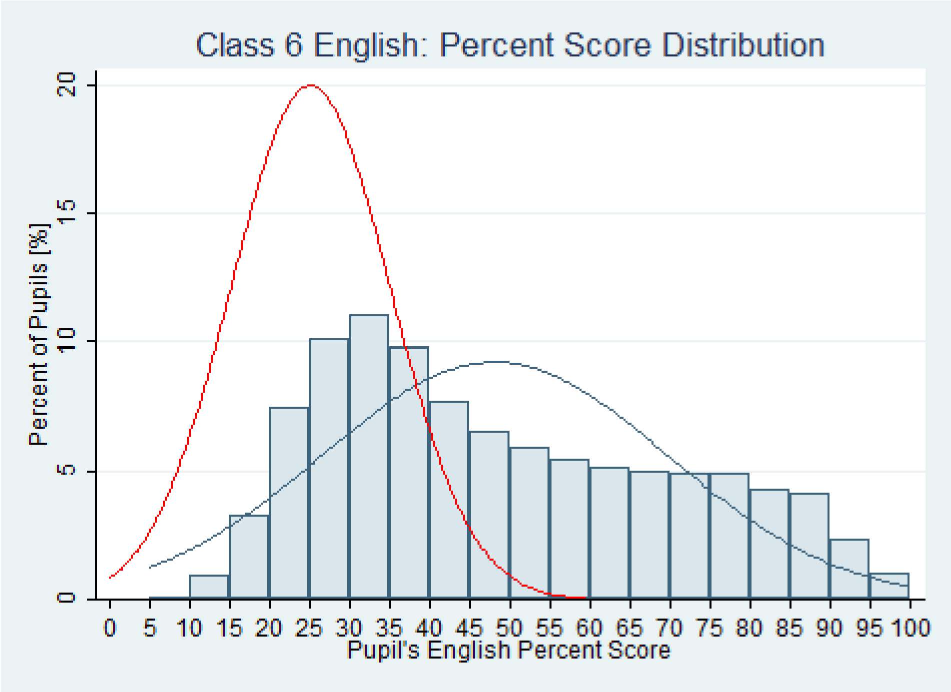


Figure D4: Distribution of English scores, P6



It can be seen in ***Table D2*** that for P6 mathematics, only 1.5% of the pupils answered more than 70% of the questions correctly and the highest score was 92.5%. Students performed equally poorly on P6 maths in 2011; only 4.2% of the pupils were able to get more than 70% of the questions correct. As mentioned above, the variance was greater for English, in both grades, with a larger proportion of students obtaining scores >70% and a small percentage of pupils successfully answering 90–100% of the questions correctly in English, for both P3 and P6.

Table D2: Proportion of pupils with more than 70% correct answers

Subject and grade	% of pupils with over 70% correct answers	Confidence interval	
Maths			
P3	10.1%	8.5%	11.6%
P6	1.5%	1.0%	2.0%
English			
P3	17.7%	15.3%	20.1%
P6	22.8%	19.8%	25.9%

Annex E: Relationship Between Individual Pupils' Performance in English and Maths

As discussed in Section 3.5 of this report, in 2013, the NEA research team modified the instruments to ensure pupils' English and mathematics tests were linked, thus allowing analysis of the relationship between English and maths results.

One of the reasons this is important is to understand the impact that reading may have on a pupil's mathematics performance. In order to test higher cognitive abilities in maths, word problems must be used. To understand the mathematics word problems, pupils must be able to read with some level of comprehension, especially at P6. As noted in Section 3.5, analysis findings suggested a strong relationship between English and mathematics performance, as suspected. Pupils who failed to achieve minimum competency in English rarely scored in the proficiency range in mathematics.

The correlation analyses documented in ***Table E1*** were carried out as a way to examine in more detail the relationship between English and mathematics and to determine the minimum performance level in English required to respond correctly to the mathematics test items. ***Table E1*** presents the bivariate correlations between subject domains in English and mathematics.

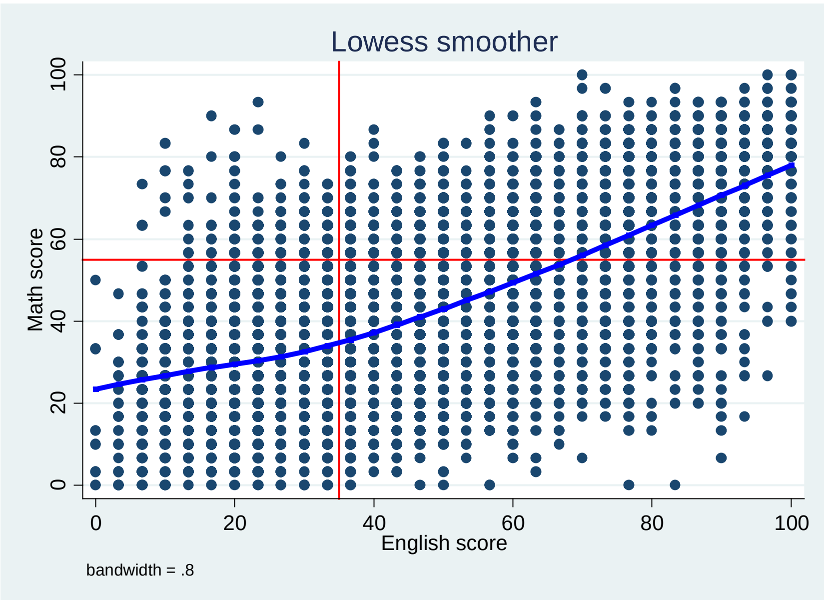
All correlations were significant. A correlation that fell above 0.50 was considered to be meaningful in the analysis. For P3, only within-subject relations met the criterion of $r > 0.50$. Reading and Grammar within the English subject area and Operations and Numbers within the mathematics subject area met this criterion. In P6, Reading and Grammar had an even stronger relationship. In addition, P6 Reading was correlated with mathematics Operations and both Reading and Grammar were correlated with Collect and Handle Data problems, which included word problems.

Table E1: Bivariate correlations between subject domains: English, maths

		English domains			Mathematics domains			
		Listening	Reading	Grammar	Operations	Numbers	Meas / Shape and Space	Collect / Handle Data
P3								
English domains	Listening	1						
	Reading	**0.4996**	1					
	Grammar	**0.531**	**0.6464**	1				
Mathematics domains	Operations	0.4941	0.489	0.492	1			
	Numbers	0.4645	0.4562	0.4532	0.5536	1		
	Measurement / Shape and Space	0.3525	0.3576	0.3537	0.4055	0.4042	1	
	Collect and Handle Data	0.3224	0.4709	0.4554	0.3929	0.3606	0.3597	1
P6								
English domains	Listening	1						
	Reading	**0.5567**	1					
	Grammar	**0.6102**	**0.7311**	1				
Mathematics domains	Operations	0.4239	0.483	**0.501**	1			
	Numbers	0.3331	0.4033	0.4033	0.42	1		
	Measurement/ Shape and Space	0.2392	0.2971	0.2903	0.2711	0.2534	1	
	Collect and Handle Data	0.4415	**0.5242**	**0.5261**	0.4231	0.3511	0.2682	1

Figure E1 shows the relationship between P3 English and maths scores.

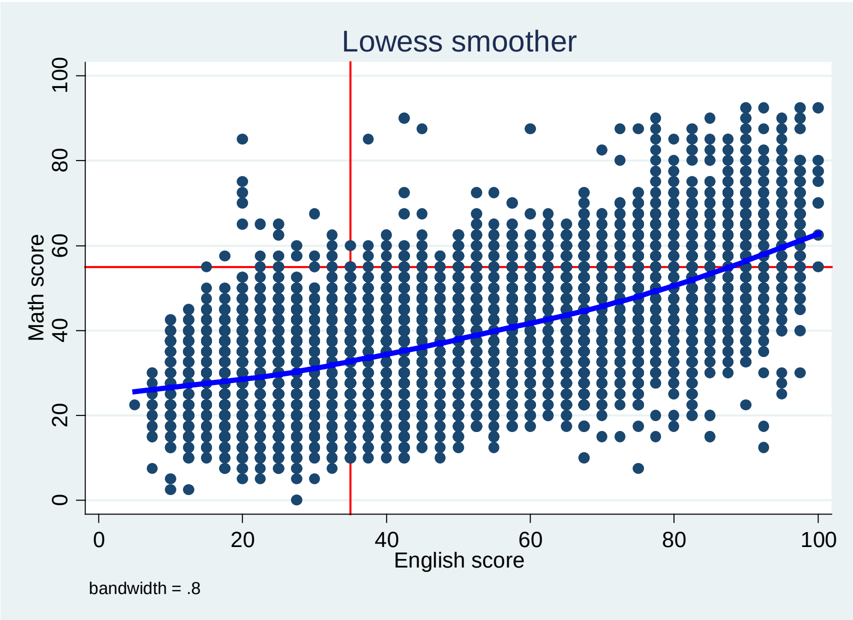
Figure E1: Non-parametric estimation of the relationship between P3 English and maths test scores



Note the inflection point that occurs at 35 in English, which means that P3 pupils could succeed on the maths portion of the test even with limited competency in English. Also, a certain proportion of P3 pupils who failed to achieve minimum competency in English did meet the criteria for proficiency in P3 mathematics (upper left quadrant). This situation was much less prominent in P6, where there were a larger number of word problems for which the need for both language and mathematics to answer the question correctly was required.

For the P6 test (see ***Figure E2***), which had more word problems than P3, very few pupils who failed to achieve minimum competency in English achieved proficiency in maths. Thus, the reading requirement in these problems may have presented a strong barrier to successfully completing these (word) mathematics problems.

Figure E2: Non-parametric estimation of the relationship between P6 English and maths test scores



Annex F: Scores and Percentages of Pupils Reaching Minimum Competency and Proficiency, with Confidence Intervals

The following provides additional information about the spread of scores given by the 95% confidence bands—the 95% confidence intervals about the means ***(Tables F1, F2, F4–F7)*** and the 95% confidence intervals about the proportions within the performance categories ***(Tables F3, F8–F13).***

Table F1: Total scores, by class and subject (% correct answers)

Subject	P3				P6			
	Mean	CI		SD	Mean	CI		SD
Maths	**41.1**	40.0	42.2	18.3	**38.2**	37.4	39.0	12.4
English	**44.4**	42.8	45.9	21.9	**48.9**	47.3	50.6	21.7

CI = confidence interval; SD = standard deviation.

Table F2: Mean (% correct) scores, by grade and subject, 2011 and 2013

Subject and year	P3			P6		
	Mean	CI		Mean	CI	
Maths						
2011	**38.6**	37.3	39.8	**39.5**	38.1	40.8
2013	**41.1**	40.0	42.2	**38.2**	37.4	39.0
English	**Score**	**CI**		**Score**	**CI**	
2011	**44.0**	42.5	45.5	**49.8**	48.3	51.4
2013	**44.4**	42.8	45.9	**48.9**	47.3	50.6

CI = confidence interval.

Table F3: Distribution of students across performance categories, by subject and class

Competency level	P3 maths			P3 English			P6 maths			P6 English		
	%	CI		%	CI		%	CI		%	CI	
Below minimum competency	**42.9**	40.7	45.2	**41.9**	39.5	44.3	**39.2**	36.8	41.6	**31.3**	28.9	33.8
Minimum competency	**35.0**	33.7	36.3	**29.7**	28.5	31.0	**50.0**	48.4	51.5	**29.8**	28.2	31.4
Proficiency	**22.1**	19.9	24.4	**28.4**	25.6	31.3	**10.9**	9.3	12.7	**39.0**	35.7	42.3
Total	**100**			**100**			**100**			**100**		

CI = confidence interval.

Table F4: NEA 2013 mean (% correct) scores in English, by grade and domain

Domain	P3			P6		
	Mean	CI		Mean	CI	
Listening	**57.5**	55.9	59.1	**54.7**	53.1	56.3
Grammar	**41.1**	39.5	42.7	**51.7**	50.1	53.4
Reading	**38.3**	36.8	39.9	**43.2**	41.4	45.1

CI = confidence interval.

Table F5: NEA 2013 mean (% correct) scores in mathematics, by grade and domain

Domain	P3			P6		
	Mean	CI		Mean	CI	
Operations	**47.0**	45.6	48.3	**44.7**	43.8	45.6
Numbers	**45.6**	44.4	46.8	**37.9**	36.9	38.9
Measurement /Shape & Space	**35.5**	34.8	36.3	**26.9**	26.4	27.5
Collect and Handle Data	**34.0**	32.6	35.5	**44.7**	43.4	46.0

CI = confidence interval.

Table F6: NEA 2013 mean (% correct) scores in mathematics, by cognitive skills and grade

Cognitive skill and grade	Mean	95% confidence interval	
Knowledge and Understanding			
P3	**45.8**	44.6	46.9
P6	**44.9**	44.0	45.8
Application			
P3	**38.0**	36.9	39.0
P6	**37.5**	36.6	38.4
Reasoning			
P3	**33.5**	32.3	34.6
P6	**23.3**	22.6	23.8

Table F7: NEA 2013 Mean (% correct) Scores in English by Cognitive Skills and Grade

Cognitive skill and grade	Mean	95% confidence interval	
Knowledge			
P3	**51.8**	50.0	53.4
P6	**54.1**	52.1	55.9
Understanding			
P3	**38.0**	36.5	39.4
P6	**50.2**	48.5	51.8
Critical Thinking			
P3	**36.0**	34.6	37.3
P6	**39.4**	37.9	40.8

Table F8: Cross-tabulation of P3 English and P3 mathematics according to distribution across NEA performance categories

	Distribution across performance categories: P3 English								
	Below minimum competency			Minimum competency			Proficiency		
Distribution across performance categories: P3 mathematics	%	CI		%	CI		%	CI	
Below minimum competency	**27.7**	25.8	29.6	**12.7**	11.9	13.5	**2.6**	2.2	3.0
Minimum competency	**12.9**	12.0	13.9	**13.5**	12.8	14.3	**8.6**	7.8	9.5
Proficiency	**1.3**	1.1	1.6	**3.5**	3.2	3.9	**17.2**	15.1	19.5

CI = confidence interval about the percentage estimate.

Table F9: Cross-tabulation of P6 English and P6 mathematics according to distribution across NEA performance categories

	Distribution Across Performance Categories: P6 English								
	Below Minimum Competency			Minimum Competency			Proficiency		
Distribution across performance categories: P6 mathematics	%	CI		%	CI		%	CI	
Below minimum competency	**21.7**	19.9	23.7	**13.0**	12.1	14.0	**4.4**	3.9	4.9
Minimum competency	**9.4**	8.5	10.2	**15.9**	14.9	17.0	**24.7**	22.7	26.8
Proficiency	**0.2**	0.1	0.3	**0.8**	0.6	1.1	**9.9**	8.3	11.6

CI = confidence interval about the percentage estimate

Table F10: Distribution of pupils across performance categories, by gender

Grade, subject, and competency level	Male			Female		
	%	CI		%	CI	
P3 maths						
Below minimum competency	**42.2**	39.9	44.5	**43.6**	41.1	46.1
Minimum competency	**36.0**	34.5	37.5	**33.9**	32.5	35.3
Proficiency	**21.8**	19.6	24.2	**22.5**	20.2	25.0
P3 English						
Below minimum competency	**42.2**	39.7	44.6	**41.4**	38.8	44.1
Minimum competency	**30.8**	29.4	32.3	**28.6**	27.2	30.1
Proficiency	**27.0**	24.3	29.9	**29.9**	26.9	33.1
P6 maths						
Below minimum competency	**37.3**	35.0	39.7	**41.0**	38.3	43.8
Minimum competency	**51.1**	49.4	52.8	**48.8**	46.8	50.8
Proficiency	**11.6**	9.9	13.4	**10.2**	8.5	12.1
P6 English						
Below minimum competency	**30.7**	28.4	33.1	**31.9**	29.2	34.7
Minimum competency	**30.9**	29.2	32.7	**28.5**	26.8	30.4
Proficiency	**38.4**	35.3	41.6	**39.6**	35.9	43.3

CI = confidence interval.

Table F11: Distribution of pupils across performance categories, by school location

Grade, subject, and competency level	Rural			Urban		
	%	CI		%	CI	
P3 Maths						
Below minimum competency	**48.7**	45.9	51.5	**33.0**	28.3	38.1
Minimum competency	**34.8**	33.4	36.3	**34.0**	30.9	37.3
Proficiency	**16.5**	14.2	19.2	**33.0**	27.6	38.9
P3 English						
Below minimum competency	**49**	46.1	51.9	**29.6**	24.5	35.1
Minimum competency	**30.6**	29.2	32.0	**26.7**	23.9	29.7
Proficiency	**20.5**	17.5	23.8	**43.7**	36.8	50.9
P6 Maths						

Grade, subject, and competency level	Rural			Urban		
	%	CI		%	CI	
Below minimum competency	**46.1**	43.3	48.9	**27.7**	22.4	33.8
Minimum competency	**46.6**	44.7	48.4	**54.5**	50.8	58.1
Proficiency	**7.3**	5.6	9.6	**17.7**	14.0	22.3
P6 English						
Below minimum competency	**39.4**	36.2	42.6	**17.8**	13.7	22.7
Minimum competency	**33.6**	31.8	35.4	**23.1**	19.6	26.9
Proficiency	**27.1**	23.6	30.8	**59.2**	51.5	66.4

CI = confidence interval.

Table F12: Distribution of pupils across performance categories, by school type

Grade, subject, and competency level	Public			Private		
	%	CI		%	CI	
P3 maths						
Below minimum competency	**49.1**	46.6	51.5	**19.5**	16.6	22.7
Minimum competency	**35.7**	34.3	37.1	**32.3**	29.4	35.4
Proficiency	**15.3**	13.2	17.6	**48.2**	43.4	53.0
P3 English						
Below minimum competency	**49.0**	46.4	51.5	**14.9**	12.4	17.8
Minimum competency	**31.9**	30.6	33.1	**21.7**	19.1	24.6
Proficiency	**19.2**	16.6	22.1	**63.4**	58.6	67.9
P6 maths						
Below minimum competency	**44.5**	41.8	47.2	**18.9**	15.9	22.2
Minimum competency	**48.0**	46.2	49.7	**57.6**	54.5	60.7
Proficiency	**7.6**	6.1	9.3	**23.5**	19.4	28.3
P6 English						
Below minimum competency	**37.2**	34.5	40.0	**8.8**	6.7	11.5
Minimum competency	**33.1**	31.5	34.8	**17.0**	14.3	20.1
Proficiency	**29.7**	26.5	33.2	**74.2**	69.0	78.7

CI = confidence interval.

Table F13: Distribution of pupils across performance categories, by schools located in deprived versus non-deprived districts

Subject and competency level	Public schools:					
	Non-deprived district			Deprived district		
	%	CI		%	CI	
P3 maths						
Below minimum competency	**45.7**	42.7	48.8	**57.5**	54.2	60.7
Minimum competency	**36.9**	35.2	38.7	**32.5**	30.5	34.6
Proficiency	**17.3**	14.7	20.4	**10.0**	8.0	12.5
P3 English						
Below minimum competency	**44.9**	41.8	48.1	**59.3**	55.3	63.3
Minimum competency	**33.0**	31.5	34.6	**28.9**	26.9	30.9
Proficiency	**22.1**	18.7	25.9	**11.8**	8.9	15.5
P6 maths						
Below minimum competency	**40.6**	37.4	43.9	**55.0**	50.9	59.1
Minimum competency	**50.7**	48.6	52.8	**40.5**	37.4	43.6
Proficiency	**8.7**	6.9	10.9	**4.5**	2.8	7.1
P6 English						
Below minimum competency	**32.8**	29.7	36.1	**49.2**	44.6	53.7
Minimum competency	**32.7**	30.7	34.8	**34.2**	31.8	36.7
Proficiency	**34.5**	30.5	38.7	**16.6**	12.8	21.2

CI = confidence interval.

Table F14: Distribution of pupils across performance categories in English, by school location and school type

Grade, school type, and location		Below minimum competency	CI		Minimum competency	CI		Proficiency	CI	
P3										
Rural	Public	55.0	52.2	57.7	31.7	30.3	33.2	13.3	11.0	16.1
Rural	Private	18.4	14.1	23.7	24.6	21.1	28.5	57.0	49.6	64.1
Urban	Public	36.0	30.0	42.4	30.3	27.3	33.5	33.8	26.5	41.8
Urban	Private	10.4	7.3	14.6	16.0	11.7	21.5	73.6	65.3	80.5
P6										
Rural	Public	44.2	40.9	47.6	35.4	33.7	37.1	20.4	17.5	23.6
Rural	Private	13.4	9.3	19.0	23.8	18.9	29.4	62.8	53.3	71.3
Urban	Public	21.9	17.0	27.7	26.9	23.1	31.2	51.2	42.7	59.6
Urban	Private	4.6	3.0	7.0	10.7	7.5	15.2	84.7	78.8	89.2

CI = confidence interval.

Annex G: Results of Methods Used to Cross-Check Non-reader Estimates

Section 2.9 of this report described four methods for estimating the percentage of non-readers using item data:

1. Pupils having ≤ 25% correct answers on the Knowledge and Understanding reading items, excluding Critical Thinking items
2. Pupils having zero scores on the 4 items assessing lower cognitive skills (Knowledge)
3. Pupils having zero scores on the 4 EGRA items
4. Pupils in a particular grade that evidenced a lower probability of responding correctly to the reading items (using latent class analysis, or LCA).

Each method yielded different estimates of the proportion of non-readers. Methods 2 and 3—the scales using zero scores on 4 items (EGRA and Knowledge level), which were more restrictive in their definition—resulted in a small proportion of pupils being classified as non-readers using the two zero-scores methods (see ***Table G1***).

Table G1: Estimation of the percentages of non-readers using various methods

Grade and estimation method	Definition	Proportion (95% confidence interval)
P3		
Method 1	Score on reading Knowledge and Understanding <25%	44.4 (42.5 – 46.4)
Method 2	Zero score on 4 Knowledge items	21.0 (20.0 – 22.1)
Method 3	Zero score on 4 EGRA items	23.4 (22.3 – 24.6)
Method 4	Latent class analysis	49.70 (not applicable)
P6		
Method 1	Score on reading Knowledge and Understanding <25%	35.2 (33.0 – 37.5)
Method 2	Zero score on 4 Knowledge items	15.0 (13.9 – 16.2)
Method 3	Zero score on 4 EGRA items	not applicable (no EGRA-like items in P6)
Method 4	Latent class analysis	not applicable

Method 4, latent class analysis, used more sophisticated tools. These methods were applied for P3 only. The optimal number of classes (or categories)—using the Bayesian information criterion (BIC) and Akaike information criterion (AIC)—is five; see ***Table G2.***

Table G2: Criteria for model selection

Model	ddl	G²	AIC	BIC
2 classes	492	1746	1784	1934
3 classes	482	1144	1202	1431
4 classes	472	845	923	1230
5 classes	462	743	841	**1227**
6 classes	452	662	780	1245

The probability of each item response, by class, is provided in ***Table G3.*** As noted, latent class analysis allows the definition of five classes.

Table G3: Results from latent class analysis and interpretation

Class		1	2	3	4	5
Prevalence of each class		**12.8%**	**13.7%**	**7.7%**	**49.5%**	**16.2%**
Cognitive levels	ref. item					
Knowledge	item23	36.3%	32.7%	59.5%	**16.5%**	96.8%
Knowledge	item24	47.1%	34.8%	49.6%	**25.6%**	93.4%
Knowledge	item27	34.0%	18.5%	58.0%	**27.6%**	96.8%
Knowledge	item20	33.3%	14.3%	84.1%	**30.3%**	96.8%
Understanding	item19	36.1%	19.5%	89.4%	**20.5%**	97.2%
Understanding	item28	55.6%	0.0%	23.2%	**27.3%**	76.4%
Understanding	item29	40.2%	16.0%	21.0%	**14.7%**	83.2%
Critical Thinking	item22	54.1%	46.8%	57.8%	**25.4%**	76.2%
Critical Thinking	item30	66.4%	42.0%	45.1%	**22.2%**	63.0%
Average score per class		**44.8%**	**25.0%**	**54.2%**	**23.3%**	**86.7%**
Definition of class		Average performers	Low performers but manage some items	Average to high performers	Non-readers (all probabilities around 25%)	High performers (manage all items)

Class 1 was characterized by pupils whose probability of answering each item correctly was around or below 25%. They can be considered as having responded randomly and were likely to be non-readers. The proportion in this class was estimated to be 49.7%, or very close to the estimate using Method 1 (44.4%).

When the two classifications using Method 1 (score under 25%) and Method 4 (LCA) were cross-tabulated, 74.8%[36] of the pupils were classified into the same category in both ***(Table G4).***

[36] A similar figure was found in the cross-tabulations for a paper-and-pencil and fluency test in Cameroon.

Table G4: Cross classification of pupils as readers/non-readers (% of pupils) using Methods 1 and 4

		Method 1 (Score < 25%)	
		Non-reader	Reader
Method 4 (LCA)	Non-reader	**38.8%**	17.9%
	Reader	7.3%	**36.0%**

Only 7.3% of the pupils classified as readers by LCA were also non-readers according to the other methods.

Finally, the research team concluded that the proportion of non-readers at P3 could be estimated between 44.4% and 49.7% using methods involving several items, and between 21% and 23.4% based on zero scores on four low-ability-level items. At P6, the estimation ranged from 15% (zero scores) to 35% (score under 25%).