Ghana 2016 National Education Assessment

Report of Findings

Ministry of Education
Ghana Education Service
National Education Assessment Unit

Ghana 2016 National Education Assessment

Report of Findings

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 sole responsibility of RTI International and the National Education Assessment Unit of Ghana, and do not necessarily reflect the views of USAID or the United States Government.

Prepared under the Education Data for Decision Making (EdData II) project, Task Order Number AID-641-BC-13-00001.

Table of Contents

List of Figures

List of Tables

Acknowledgements

The authors wish to acknowledge the staff of the National Education Assessment Unit of Ghana for its leadership in the 2016 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 2016. These staff are Mr Antwi Aning, Mr Anthony Sarpong, Mr Joachim Honu, and Mrs Joana Vanderpuije. We would also like to acknowledge the education staff from USAID and in particular Sarah Banashek, 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 2016 NEA.

Crucial support throughout and at key points along the way was provided by the RTI International Partnership for Education: *Testing* team: Emmanuel Sam-Bossman, Chief of Party; Elizabeth Randolph, International Team Leader; Megan McCune, Project Coordinator; Emily Kochetkova, Project Manager; Jamie Freedman; Survey Methodologist; and Erin Newton, Senior Editor. We are also grateful to Pierre Varly, independent statistician and mathematics expert, and Michael Fast and Mauricio Estrada, assessment specialists and psychometricians at the American Institutes for Research. We thank our partner the Education Assessment and Research Centre (EARC) for their support during data analysis and report writing.

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; nongovernment 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.

Abbreviations

BECE	Basic Education Certificate Examination
CRDD	Curriculum Research and Development Division
EARC	Education Assessment and Research Centre
EdData II	Education Data for Decision Making (USAID project)
EGMA	Early Grade Mathematics Assessment
EGRA	Early Grade Reading Assessment
FCUBE	Free and Compulsory Universal Basic Education
G2G	Government to Government
GDP	gross domestic product
GER	gross enrolment ratio
GES	Ghana Education Service
GHS	Ghanaian cedi
NAR	net admission ratio
NEA	National Education Assessment
NEAU	National Education Assessment Unit (formerly Assessment Services Unit)
NER	net enrolment ratio
P1, P2, P3, P4, P6	primary grades 1, 2, 3, 4, 6
RTI	RTI International (trade name of Research Triangle Institute)
UNICEF	United Nations Children's Fund
USAID	United States Agency for International Development

Executive Summary

This report presents the findings from the 2016 administration of the Ghana National Education Assessment (NEA), carried out by the National Education Assessment Unit (NEAU) within the Ghana Education Service (GES). The NEA is a biennial nationally and regionally representative measure of pupil competency in mathematics and English in primary classes 4 and 6 (P4 and P6). In 2016, the NEA was revised to assess P4 rather than P3 pupils to better align with Ghana's current language-of-instruction policy.

The 2016 edition was the sixth application of the NEA, and it covered all 10 regions of Ghana, sampling 550 schools and testing 35,996 pupils over the course of three days in July 2016. The sample size was designed to be representative at the national and regional levels, but not at the district level.

The 2016 NEA was a classroom-based multiple-choice written test whose content was based on the national curricula. The P4 test contained 40 items each for mathematics and for English, and the P6 test contained 45 items for each subject. Test items covered skills and knowledge across the following domains:

English	**Mathematics**
Listening Comprehension	Operations
Grammar	Numbers
Reading	Measurement
	Shape and Space
	Data and Chance

For the past three administrations (2011, 2013, and 2016), the United States Agency for International Development (USAID) has sponsored technical assistance to the NEAU through the Education Data for Decision Making (EdData II) contract. Under this contract, test content was improved, comparability of test results between 2013 and 2016 was ensured and a test item bank was developed to facilitate future test development and administration. Pilot testing ensured that individual items within the test would measure the intended cognitive skills with reliability and an appropriate range of difficulty.

Test Results

All four subject-area components (P4 mathematics, P4 English, P6 mathematics and P6 English) used the same test score cut-points to indicate that a pupil had achieved the *minimum competency level* and the *proficiency level*. These cut-points were established by the GES in 2005. Pupils who scored 35% correct were defined as having reached minimum competency and pupils scoring 55% or better were defined as having reached proficiency.[1]

[1] The authors wish to note that international standards for 'proficiency' are generally set at a higher cut-point. That is, a more common requirement for 'proficiency' is to correctly answer least 70% of the questions correctly. The NEA's criterion for 'proficiency', reported here, was established in 2005 with the first NEA, and is based on answering just over half of the items correctly (i.e., ≤ 55%) and thus does not effectively identify pupils who have a full grasp of the curriculum – that is, who are truly proficient in the subject area.

Table ES1 provides information on pupil performance according to the two nationally defined cut-points and includes the proportion of pupils who failed to achieve minimum competency. Less than 25% of the pupils met the proficiency cut-point in P4 and P6 mathematics and less than 40% of the pupils achieved proficiency in P4 and P6 English. A range of pupils from 28% to 45% 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 pupils meeting criteria for minimum competency and proficiency, by subject and grade

	Percentage of pupils in competency range, by grade and subject tested			
	P4		P6	
Competency levels	Mathematics	English	Mathematics	English
Below Minimum Competency	45.2	29.3	29.2	28.4
Minimum Competency	32.8	33.5	45.9	33.7
Proficiency	22.0	37.2	24.9	37.9
Total	100.0	100.0	100.0	100.0

Performance According to Subject Domains

Analyses of pupil performance across the various subject domains within English and mathematics revealed some noteworthy patterns. In English, the Reading domain presented the greatest challenge to pupils, in both P4 (44% correct, on average) and P6 (43% correct, on average). The highest score was in the P4 Listening Comprehension domain, where pupils scored 70% on average. In mathematics, both P4 and P6 pupils had difficulty with the higher-order cognitive tasks involving Measurement (34% and 29% correct, on average, in P4 and P6 respectively) and Shapes and Space (38% and 39% correct). The highest score was in the P6 Data and Chance domain, where pupils scored 53% on average.

Performance Across Subpopulations

The NEA data were also examined according to several subgroups (see ***Table ES2***). Average performance for males and females was the nearly the same for P4 mathematics (41.9% correct for males and 41.5% correct for females) and similar for P6 English (47.6% correct for males and 48.1% correct for females). There were small but statistically significant differences between male and female pupils' performance in P4 English (49.8% correct for males and 52.0% correct for females) and P6 mathematics (44.9% correct for males and 42.8% correct for females). Females outperformed males in P4 English and males outperformed females in P6 mathematics.

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. Average performance among pupils in urban areas was significantly higher than for pupils residing in rural areas, although the highest average score was still below 60% correct (in P4 English). The disparities were similar for pupils residing in deprived versus non-deprived districts. Not surprisingly, the performance of pupils residing in the three regions of northern Ghana (Northern, Upper East,

Upper West) – where the majority of pupils sampled were residing in a deprived district – was poorest.

Performance differences between pupils attending public schools as compared to private schools were even greater, with a 23.5 percentage point difference between the highest average scores of 69.9% correct in P4 English for private school pupils and 46.1% correct for public school pupils.

Table ES2: Pupil performance by sex, location, type of school, and type of district

Subject and grade	Sex		School location		School type		District type	
	Male	Female	Rural	Urban	Public	Private	Non-deprived	Deprived
Mathematics								
P4	41.9%	41.5%	37.9%^	47.0%***	38.1%^	55.6%***	43.5%^	35.1%***
P6	44.9%***	42.8%^	40.8%^	47.8%***	41.6%^	53.2%***	45.4%^	37.9%***
English								
P4	49.8%^	52.0%***	45.2%^	59.0%***	46.1%^	69.6%***	53.3%^	42.1%***
P6	47.6%	48.1%	41.6%^	56.0%***	43.9%^	64.6%***	50.3%^	38.4%***

^ = reference; ***p = 0.000.

Recommendations

The results from the 2016 NEA are similar to the 2013 NEA findings. There has been no significant or substantive change in pupil performance since the 2013 NEA.[2] Large numbers of pupils are struggling to master the P4 and P6 curricular content. This finding is not surprising given that the both the 2013 and 2015 National Early Grade Reading and Mathematics Assessments (EGRA and EGMA) indicated that the majority of children in P2 lacked the foundational skills that they would need to succeed at the P4 and P6 level. In order for pupils to perform in the upper primary classes, P4 to P6, they must develop and apply the critical foundational concepts and skills taught in the early primary curriculum. Specifically, pupils need to learn basic mathematics concepts and skills, and apply this knowledge to solve mathematical problems of a more conceptual nature. Similarly, it is critical that pupils develop important pre-reading skills such as letter sound knowledge and word decoding strategies in order to read new words. In order to perform in P4 English and beyond, pupils must learn to read with fluency and comprehension. Thus it is recommended that the focus in Ghana primary education now and in the years to come be enhancing instruction in the early primary grades to ensure that these pupils have the foundational skills needed to succeed in school.

Second, it is recommended that the Ghanaian education sector continue to work strategically towards reducing the disparities in primary education programs that exist across rural and urban settings. The stark differences in learning outcomes among pupils attending schools in urban and peri-urban locations compared to those of pupils attending schools in rural

[2] The comparison with 2013 is only for P6, given that P4 pupils were not tested before 2016.

locations has been a consistent finding of the NEA since its first administration in 2005. This is a challenge that continues to demand attention in order to ensure that children across Ghana have access to high-quality education, leading to successful achievement of the knowledge and skills that are supported by Ghana's primary school curriculum.

I. Introduction

Purpose, History and Transition of the National Education Assessment

This report presents the findings from the 2016 administration of the Ghana National Education Assessment (NEA), a curriculum-based measure of pupil competency in mathematics and English in Primary Class 4 and 6 (i.e., P4 and P6). The NEA was carried out by the National Education Assessment Unit (NEAU)[3] of the Ghana Education Service (GES).

In addition, the report reviews a number of enhancements to the NEA that took place before the 2016 NEA was administered. One of the key enhancements in the NEA in 2016 was the development of the P4 assessment instruments. After consideration of the national curriculum in language and the language-of-instruction policy, the GES recommended that it would be best to test competency in English only after the pupils fully transition to English as the medium of instruction, which takes place in P4. Thus, new curriculum-based mathematics and English tests were developed prior to the 2016 NEA administrations, along with an item bank that will be used to develop tests in subsequent years. Technical assistance in support of the 2016 NEA was provided under the United States Agency for International Development (USAID) Ghana Partnership for Education: *Testing* activity.[4]

The 2016 NEA was the sixth round of the biennial NEA and covered all 10 regions of Ghana, sampling 550 schools and 35,996 pupils in the course of three days. As part of the analysis prepared for this report, the results were disaggregated on the following: sex, location (urban/rural), type of school (public vs. private), and whether schools were within a deprived district or not.

Over the years, as part of the *Testing* activity, technical assistance has involved enhancements to the test content as well as changes to conform to policy, to improve comparability for detecting any historical trends, and to ensure that individual items within the test are measuring the intended cognitive skills with reliability and an appropriate range of difficulty.

Background on Ghana's Education Sector

This section presents some background on the state of Ghana's education system and culture, as context for the NEA from its inception in 2005 through the current administration.

Education Expenditures

Information on education expenditure in Ghana was sourced from the 2015 *Education Sector Performance Report.*[5] Since 2011 there has been a steady increase in absolute expenditure on education: an 84% increase from GHS 3.6 billion in 2011 to GHS 6.6 billion in 2014 and a 15.2% increase in the past year (from GHS 5.7 billion in 2013). However, the percentage of education expenditure as a percentage of gross domestic product (GDP) decreased between 2011 and 2014, from 25.8% in 2011 (2011 GDP = GHS 57.0 billion) to 20.5% in 2014 (2014 GDP = GHS 113.4 billion). Between 2011 and 2014, education expenditure as a percentage of GDP decreased only slightly, from 20.7% to 20.5%. Although allocations to primary

[3] The National Education Assessment Unit was formerly known as the Assessment Services Unit.
[4] 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, and the Ghana Education Service, called the *USAID Partnership for Education Program.*
[5] Ghana Ministry of Education and Sports. 2015. *Education Sector Performance Report.*

education as a proportion of all education expenditure decreased from 2011 to 2014, from 34.6% in 2011 to 22.0% in 2014, absolute spending increased, from GHS 1.2 billion to GHS 1.4 billion in 2014.

Trends in education expenditure also reflect the growing emphasis in Ghana on the importance of pre-primary education. Overall funding for preschool has increased substantially, in terms of both the share of all education expenditure and absolute spending. In 2014, 7.6% of education expenditure was allocated to pre-primary education (GHS 501.4 million) compared to 2.9% in 2011 (GHS 103.4 million).

Trends in nongovernment sources of education finance in Ghana reflect increases in funding to the education sector as well. Between 2011 and 2014, there was a 120% increase in education spending from Internally Generated Funds, with an 11.3% increase from 2013 to 2014. Donor funding also increased substantially, by 150% from 2011 to 2014, with a 19.7% increase in funding from 2013 to 2014.

The majority of government spending on education is used for salaries and other personnel costs (e.g., travel, allowances). In 2014, this figure amounted to 97.7% for compensation, a slight increase from 95.6% in 2011. For primary education, in 2014 and previously, almost all government spending (99.2% in 2014 as compared to 99.4% in 2011) was allocated for salaries and other personnel costs, leaving a budget allocation of 7.7% in 2014 for goods and services. This disparity in funding for compensation (i.e., salaries and other personnel costs) versus goods and services is exacerbated whenever personnel costs run over budget. Even so, this situation represents an improvement from 2011 (when compensation costs were 172% of the budgeted amount); in 2014, compensation costs (i.e., salaries and other personnel expenses) were 116.6% of the budgeted amount. This, in turn, further squeezed expenditures for services, which tended to run under budget.

Access, Retention, Quality, and Equity

Primary school enrolments have almost doubled since the introduction of Free and Compulsory Universal Basic Education, or 'FCUBE' (e.g., an increase from 2.5 million in 1999/2000 to 4.3 million in 2013/2014). Both the gross enrolment ratio (GER) [6] and net enrolment ratio (NER) [7] have grown in the past decade, with a 33 percentage-point increase, from 58% in 2003/2004 to 91% in 2014/2015. 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. The NER for children residing in deprived districts, 93.5% in the 2014/2015, was slightly higher than that of the national NER for Ghana's primary schools (91.0%).

In spite of overall gains in enrolment, late entry into primary school and irregular attendance continue to present barriers to learning in primary school. Even though there have been

[6] Gross enrollment ratio for primary school is calculated as the number of children enrolled in primary school divided by the number of children in the population who are of school-going age and should be in school. GER is often observed to be more than 100%. That is, often children who attend primary school are under age, because parents may enroll their children before they turn six; or over age, because children may start school late or frequently repeat classes, and as a result may still be attending primary school when they are over 12 years of age.

[7] Net enrolment ratio is calculated as the number of children enrolled who are of school-going age divided by the number of children in the population who are of school age and should be in school.

overall improvements in access, the net admission ratio (NAR)[8] highlights the fact that large proportions of children of school-going age still are not in school. The NAR for Ghana in 2014/2015 was 79.6%, which suggests that 20.4% of children entering P1 were not six years of age.

The percentage of primary school pupils attending private schools has also increased, from 18.6% in 2009/2010 to 25.3% in 2014/2015. According to the 2014/2015 data, the percentage of children in deprived districts who were attending private schools was about half the national average, at 12.6%. Even though children from rural areas, particularly in the north, depend heavily on public education, the distribution of resources, particularly trained teachers, favours the urban and wealthier districts and thus is an important factor in inequities observed in learning outcomes between pupils attending school in urban versus rural locations. The overall percentage of trained teachers in Ghana's primary schools was 61.7% in 2014/2015 (an increase from 32% in 2009/2010) compared to the percentage of trained teachers assigned to schools in the deprived districts, which was 47% according to the 2014/2015 educational statistics.

II. Test Structure and Administration

Test Content and Structure

The National Education Assessment tests are based on national curricula and are made up of 40 multiple-choice questions for P4 and 45 for P6. In 2016, ten forms were developed for the P4 and P6 mathematics and English tests. This served to maximise test integrity, but also provided an opportunity to pilot test items that are planned for the 2018 NEA. Each test form was made up of operational test items (i.e., items developed for the 2016 NEA), anchor items (i.e., items from the 2013 NEA that were also included in the 2016 NEA), and five new items that were included for the purpose of piloting them for use in the 2018 NEA. See ***Annex A*** for a detailed explanation of the test development process.

The subdomains tested for mathematics are presented in ***Table 1,*** and those for English appear in ***Table 2***. This was the first year that pupils were administered the P4 tests, and thus the distribution of items across domains for P4 is given only for 2016. Changes in the item specifications across domains for the two subject areas were, in part, due to the differences in the curriculum from which the tests in 2013 and 2016 were based. The most recent primary school curriculum was developed in 2012, yet this curriculum had not been fully disseminated by the time of the 2013 NEA. Thus, the 2013 NEA was based on the previous curriculum. The 2016 NEA was based on the 2012 curriculum for the first time.

The distributions of items across subject domains given in Table 1 and Table 2 reflect, in part, differences in the curricula from which the tests were based. In both subjects, the numbers of items tested in each domain were sufficient to produce accurate average pupil scores by domain.

[8] Net admission ratio is the number of children six years of age who are admitted into Primary Class 1.

Table 1: Distribution of mathematics items by content domain

Subject and domain	No. of P4 items		No. of P6 items	
	2013	2016	2013	2016
Maths				
Basic Operations	n/a	16	15	16
Numbers	n/a	5	7	4
Measurement	n/a	6	n/a[a]	8
Shape and Space	n/a	3	12	6
Data and Chance	n/a	5	6	6
Pilot items for 2018[b]	n/a	5	5	5
Total	**n/a**	**40**	**40**	**45**

n/a = not applicable.

[a] Measurement and Shape-Space were combined into one domain in 2013.

[b] Pilot items were allocated differently by domain across the test forms.

Table 2: Distribution of English items by content domain

Subject and domain	No. of P4 items		No. of P6 items	
	2013	2016	2013	2016
English				
Listening Comprehension	n/a	8	8	10
Grammar	n/a	11	16	14
Reading	n/a	16	16	16
Pilot items for 2018[a]	n/a	5	5	5
Total	**n/a**	**40**	**40**	**45**

n/a = not applicable.

[a] Pilot items were allocated differently by domain across the test forms.

In addition to developing the P4 tests and ensuring alignment with the 2012 curriculum, the research team developed operational items for the 2016 NEA which extended the scope of skills tested. The 2016 set included a balance of items that tested pupils' ability to perform tasks involving application and critical thinking cognitive abilities, and items that involved lower levels of cognitive ability, such as knowledge and understanding.

Another enhancement to the 2016 NEA was the development of an electronic item bank. This was a noteworthy addition in that it provides a comprehensive profile of each item that could be selected for developing operational forms for the NEA tests in the future.

Test and Item Analyses

The NEA tests for all subjects and both grades were taken through all stages of test development. After the test items were developed and pilot tested, the instruments were subjected to a number of conventional psychometric analyses to finalise them. These

included: item difficulty analysis, distractor analysis, differential item function, analysis of reliability, and use of item response theory and Rasch methods.

The results of psychometric analyses conducted on the final set of items revealed some improvements in the 2016 tests over the 2013 versions. First, in comparison to the 2013 tests, the 2016 tests had less measurement redundancy (i.e., more diverse content within the same number of items) for both subjects and grades. Second, the mathematics and English tests for the 2016 NEA demonstrated better alignment with the curricula than was achieved in 2013. Third, efforts were made to reduce the amount of text and written instructions as part of the mathematics operations in order to lower the literacy burden associated with the mathematics tests. However, it was not possible to eliminate written instructions altogether, considering the need for word problems to test mathematical reasoning in P6. As a result, for some test items, literacy may have been a confounding variable for some children. Finally, test reliability across the 2016 instruments and forms met or exceeded the conventional acceptable value for tests of this nature (Cronbach's alpha $\geq$ .80).

Additional analyses were conducted to evaluate the relative pupil performance on the P6 tests from 2013 to 2016. Findings from this equating exercise indicated that there were no significant performance differences from 2013 to 2016, in either English or mathematics.

Sampling

The 2016 NEA sample was drawn from a sampling frame based on the 2014/2015 EMIS data that contained a census of all primary schools. As with previous administrations, after the exclusion of schools that contained a P4 or P6 pupil enrolment of less than 10 pupils (n = 4,022 schools), 15,754 schools remained in the sample frame. Schools were stratified by region and by district, locality (urban or rural), school type (public or private), deprived and non-deprived districts and enrolment size, to ensure representation from each of these categories. For each region, 55 schools were randomly sampled with equal probability, for a total of 550 schools. All P4 and P6 pupils attending selected schools on the day the NEA was administered (11 July 2016) were automatically selected to take the test. A total of 18,915 P4 pupils and 17,081 P6 pupils participated in the 2016 NEA administration. More details on the sampling methodology are provided in ***Annex B***.

Training of Test Administrators and Monitors

The 2016 NEA was administered by a team of trained test administrators and monitors, who were first trained by a team of four master trainers. NEAU staff conducted a training-of-trainers workshop for the 2016 NEA master trainers. During the workshop, which started on May 11 and ended on May 13, 12 GES staff from three District Education Offices and the Curriculum Research and Development Division (CRDD) were trained. These individuals conducted 10 regional trainings at five training centres from late May through mid-June.

Over 900 test administrators and monitors participated in a three-day training, which focused on standard administration of the NEA, with ample opportunity for review and discussion of the instruments and of administration procedures and practice.

Based on lessons learned from the 2013 training, the goals of the 2016 training were extended and enhanced to ensure that participants:

- Developed the ability to present a standardised training programme to instruct test administrators and test monitors, and obtained all instructions needed to oversee the test monitors and administrators throughout the NEA implementation.
- Demonstrated a thorough understanding of the details of the 2016 implementation plan, protocol, and procedures.
- Demonstrated that they were conversant in the reading of instructions, and learned how to present the training materials and to train test administrators to follow the instructions for administering the test in a standardised manner.
- Developed the skill of creating, modifying and using PowerPoint presentations for training.

Interactive approaches were applied to the 2016 test administrator training, which was an improvement from the 2013 workshops. There were important guided role-playing activities, peer learning experiences, and self-evaluation facilitated by the trainers. These activities enabled all participants to seek clarification on both the procedural and the material changes made to the 2016 NEA test administration. Unlike during the 2013 workshop, test administrators and monitors were encouraged to take notes in the training manuals given to them.

Preparation of Testing Materials

During the NEA regional trainings, the NEAU collected school enrolment figures from the District Education Offices.

District enrolment figures collected from test administrator training participants were used to create packing allocation (materials) forms to guide the packing of schools-based materials, which included instruments, 'bubble' answer sheets (see ***Figure 1***), and test monitoring forms.

Over the course of three weeks, 35 packers, including 4 CRDD staff, 4 NEAU staff, and 4 RTI staff, packed over 50,000 sets of testing materials, and controlled the quality of each school material package, after which the packaged materials were put under lock and key until the day of the test. The packaged material was distributed by truck in 10 days.

2016 NEA test forms and storage bags

In addition to the testing materials, districts received an envelope containing the following:

- List identifying the sampled schools
- Letter to each school, informing them of the test dates
- Monitoring form

- Note to the test monitors listing the contents of the district envelope and instructions on how and when to distribute the contents.

Data Collection and Management

Data collection for the 2016 NEA took place July 11–13, with test administration completed in approximately 60% of the schools on the first day. By July 13, all schools had completed the tests. The test booklets were collected from the Regional Centres July 21–28. The data cleaning and scanning of the answer sheets followed, with the scanning completed by 9 August 2016.

Based on lessons learned from the 2013 data collection, the following approaches were adopted to increase the efficiency of the NEA 2016 data collection:

- The two-sided bubble sheets for P4 and P6 were distinguished by colour. The P4 bubble sheets for optical scanning were pink while the P6 sheets were blue (see Figure 1).
- The bubble sheets were modified such that number of items on bubble sheets aligned with the number of items on the test.

Figure 1: Sample pupil answer sheet ('bubble sheet'), front and back: Colour-coded for P4 (pink) and P6 (blue)

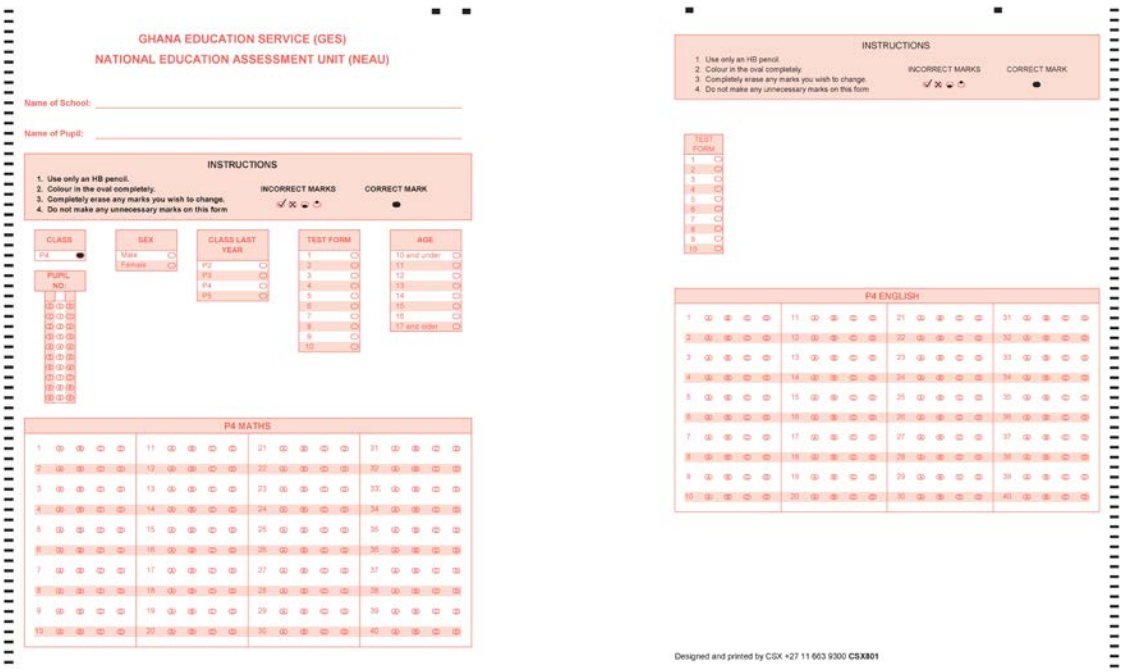
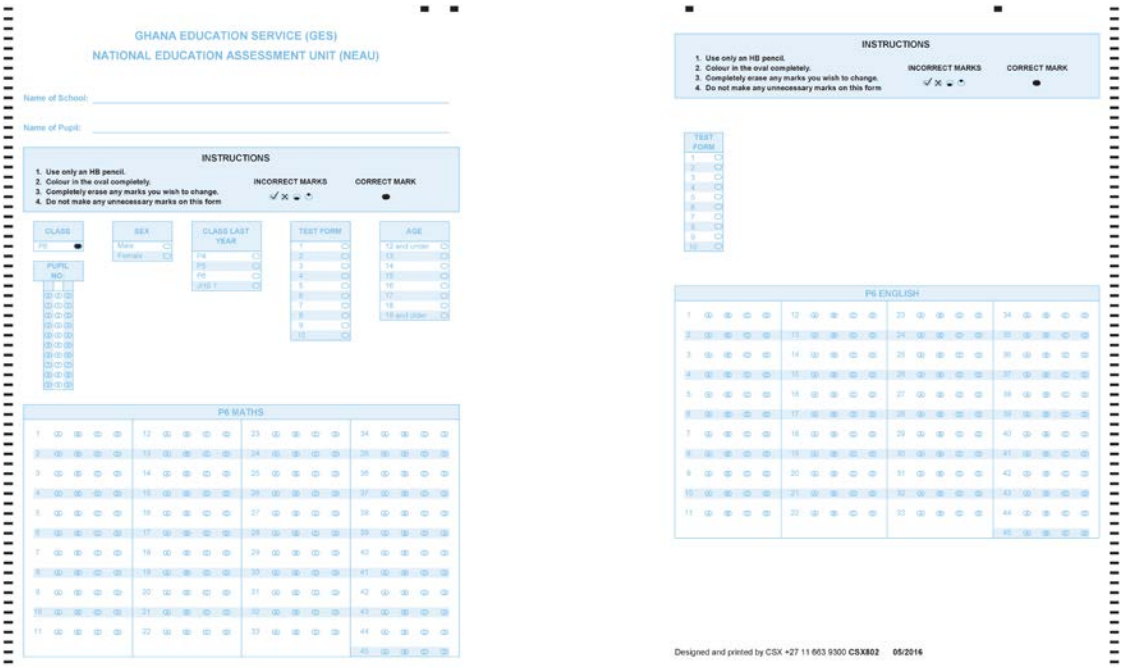


Figure 1, continued



III. 2016 NEA Results

All four tests (i.e., P4 mathematics, P4 English, P6 mathematics, and P6 English) used the same test score cut-points to determine a performance level for each pupil. According to the criteria set by the NEAU (then, the Assessment Services Unit) and the GES in 2005, pupils correctly answering at least 35% of the items on a test were considered to have achieved minimum competency in the subject. Pupils who correctly answered 55% or more of the items were considered to have achieved proficiency in the subject matter.

The authors wish to note that international standards for 'proficiency' are generally set at a higher cut-point. That is, a more common requirement for 'proficiency' is to correctly answer least 70% of the questions correctly. The NEA's criterion for 'proficiency', reported here, is based on answering just over half of the items correctly (i.e., $\leq$ 55%) and thus does not effectively identify pupils who have a full grasp of the curriculum – that is, who are truly proficient in the subject area. In keeping with past NEA reporting convention, pupil performance by cut-point is presented. However, mean scores are presented as well, which give a more accurate picture of pupil performance.

Pupils Reaching Minimum Competency and Proficiency

Figure 2 illustrates the proportions of pupils achieving minimum competency and proficiency, as defined by the cut-points described above. A third category presents the percentage of pupils whose performance fell below the minimum competency level, or having less than 35% of the items correct.

The NEA findings indicated that primary school pupils were challenged by both English and mathematics, with no more than 37% of pupils achieving proficiency levels in any grade or subject. Performance was noticeably lower for mathematics than for English, with only 22%

of P6 pupils and 25% of P6 pupils achieving proficiency in mathematics compared to 37% of P4 pupils and 36% of P6 pupils achieving proficiency in English. It is also important to highlight that for both grades and for English and mathematics, at least 29% of the pupils failed to correctly answer 35% of the questions correctly, the cut-point for minimum competency. That is, 29% of the P4 English pupils and P6 mathematics pupils, 30% of P6 English pupils, and 45% of P4 mathematics pupils performed below the minimum competency level.

Figure 2: 2016 minimum competency and proficiency results, English and mathematics, P4 and P6

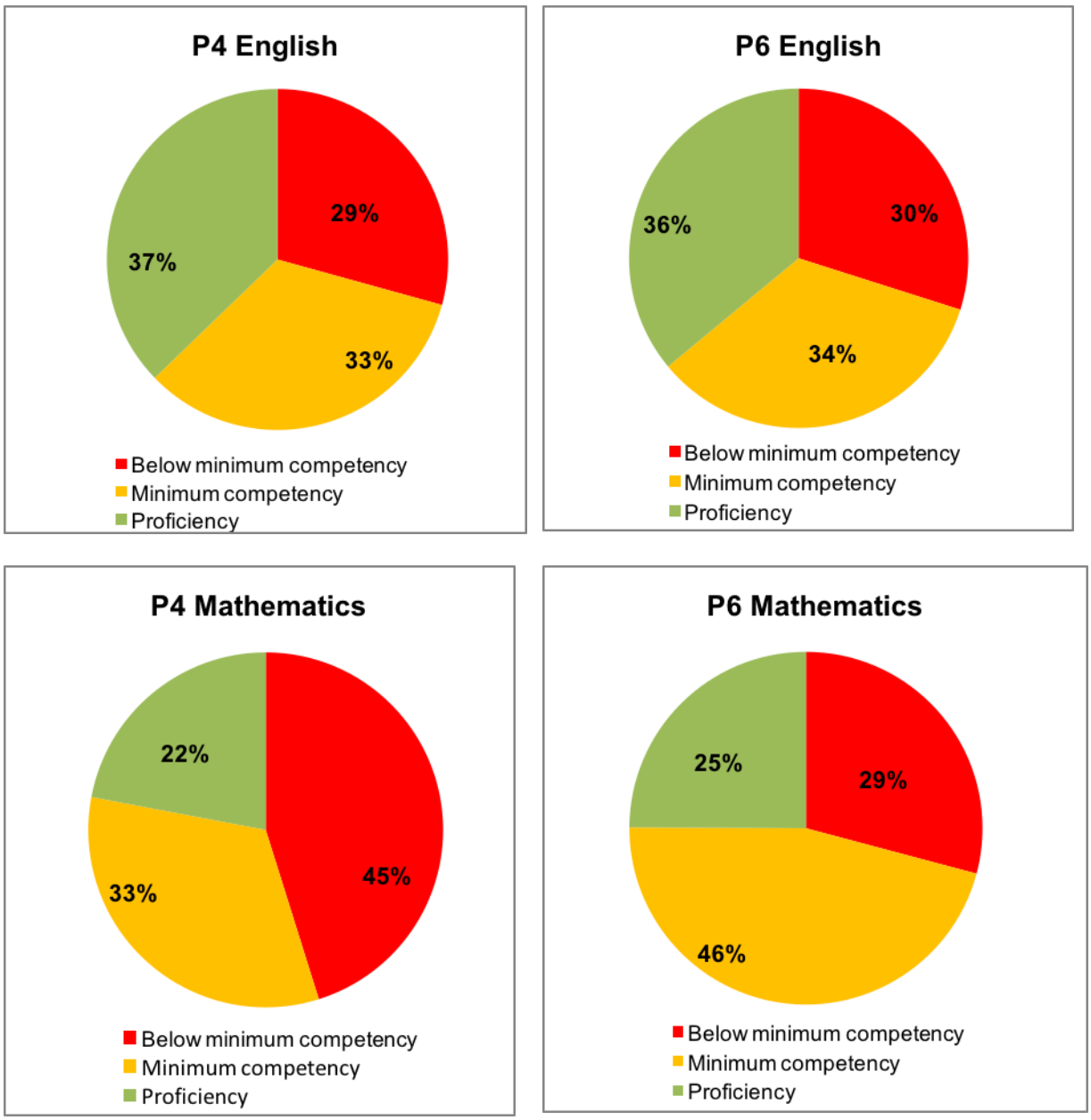


Table 3 shows the national means, based on percentage correct scores, by grade and subject. As demonstrated above, mathematics seems to have presented slightly more challenges to P4 and P6 pupils than English.

Table 3: Overall average percent correct, by grade and subject

Subject	Mean scores by grade	
	P4 (95% confidence interval)	P6 (95% confidence interval)
Mathematics	41.7 (40.5 – 42.8)	43.8 (43.0 – 44.7)
English	50.9 (49.4 – 52.3)	47.8 (46.4 – 49.2)

Results by Domain

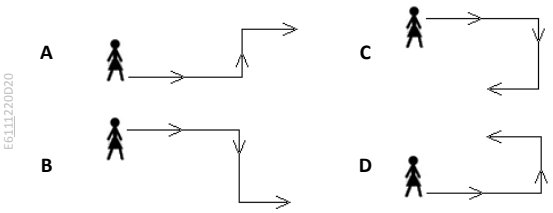
As noted at the beginning of Section II, the questions in mathematics and English covered multiple domains, providing an opportunity to deepen understanding of pupils' relative strengths across domains and to help identify potential gaps in performance according to the different domains tested. The following provides information on pupil performances in relation to the subject domains.

English Outcomes According to Domain

The English subject domains tested included (1) Listening Comprehension, (2) Grammar, and (3) Reading. These are briefly described below, alongside an example test question for each domain.

Questions from the Listening Comprehension domain required children to listen to a few sentences that the test administrator presented orally (the administrators presented the sentences twice) and to answer a question asked about the sentence by selecting the best response from four multiple-choice questions (see sample in ***Figure 3***).

Figure 3: Sample Listening Comprehension questions, P4 and P6 English

Primary 4. Question 1: Here is the text. Please listen carefully. You will hear the text twice. Draw a triangle. Then draw a straight line inside the triangle. Now write the letter C inside the triangle at the top.	**Primary 6. Question 1: Here is the text. Please listen carefully. You will hear the text twice. To get to the market, Adwoa walks straight from her house, turns left, and then turns right. The market is right in front of her.**
1. What should you draw? 	**1. How does Adwoa get to the market?** 

The **Grammar** domain assessed the pupils' ability to select the correct word in a sentence that would ensure the sentence was correct in content and grammatical structure ***(Figure 4).*** The pupils' knowledge of certain grammatical morphemes or 'functions' words, such as prepositions and pronouns, were assessed along with other grammatical structures such as plurality and tense.

Figure 4: Sample Grammar questions, P4 and P6 English

Section B: Grammar

Choose the correct word or words that complete each of the following sentences.

Primary 4	Primary 6
11. The School Prefect _____ all the fruits in the basket. **A has eating** **B have eaten** **C have eating** **D has eaten** E42111000107	**13. The goalkeeper is the _____ boy in the team.** **A taller** **B tall** **C more tall** **D tallest** E62111000054

The **Reading** domain assessed pupils' ability to read a short passage silently and then answer a series of comprehension questions based on the passage, using a multiple-choice format ***(Figure 5).***

Figure 5: Sample Reading questions, P4 and P6 English

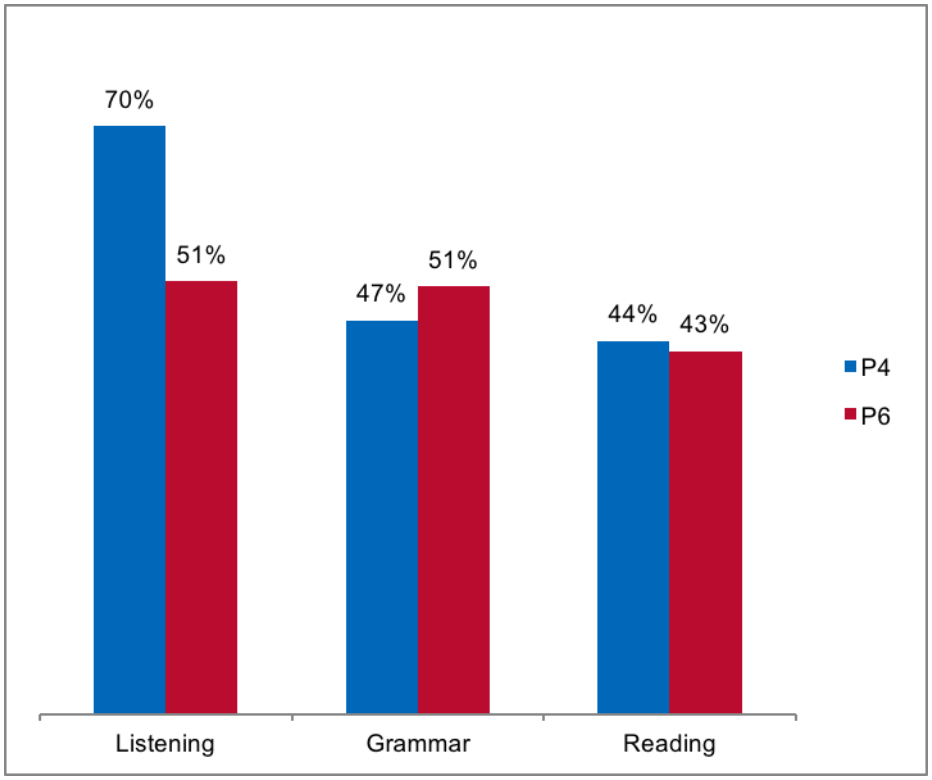
Section C: Reading

First read the text, then answer the questions that follow.

Primary 4	Primary 6
My name is Umar. I live on a farm with my mother, father and sister Aisha. Every year the land gets very dry before the rains come. We watch the sky and wait. One afternoon as I sat outside, I saw dark clouds. Then something hit my head, lightly at first and then harder. I jumped up and ran towards the house. The rains had come at last. **25. Where does Umar live?** E4312114N086 **A In the city.** **B Next to a shop.** **C On a farm.** **D In the north.**	**A farmer went out one day to search for a lost calf. The herdsmen had returned without it the day before. And that night there had been a terrible storm. The farmer went to the valley and searched by the river banks, among the weeds, behind the rocks and in the rushing water. He climbed the slopes of the hill with its rocky cliffs. He looked behind a large rock in case the calf was hiding there from the storm.** **There, on a rock, was a most unusual sight. An eagle chick had hatched from its egg a day or two earlier, and it had been blown from its nest by the storm. The farmer took the chick and held it in both hands. He would take it home and care for it.** **He was almost home when his children ran out to meet him. "The calf came back by itself!" they shouted.** **30. Why did the farmer go out one day?** E6311205N032 **A To look for some of his herdsmen.** **B To rescue an injured eagle chick.** **C To protect his animals from the storm.** **D To find a calf that had got lost.**

The average percentage correct scores according to each domain in English are presented in ***Figure 6.*** It can be seen that, with the exception of Listening Comprehension, performance across these domains was similar for P4 and P6, with performance on tasks in the Grammar domain very slightly stronger than on the tasks in the Reading domain. The Listening Comprehension tasks for P4 were the strongest for P4, with an average percentage correct score of 70% compared to the average percentage correct on Listening Comprehension for P6 pupils, which was 51%. Performance on questions in the Listening Comprehension domain for P6 was not as strong as one would have expected.

Figure 6: Average (% correct) scores by domain—P4 and P6 English



Mathematics Outcomes According to Domain

The 2016 NEA tested performance in five mathematics subject domains: (1) Operations, (2) Numbers, (3) Measurements, (4) Shape and Space, and (5) Data and Chance. These are briefly described below, with an example test question from the P4 and P6 mathematics test given for each domain.

The **Operations** domain involved having pupils compute basic mathematical operations involving addition, subtraction, multiplication, and division, such as those in ***Figure 7:***

Figure 7: Sample Operations questions, P4 and P6 mathematics

Primary 4	Primary 6
12. 640 – 280 = ☐ M4M23012001A **A. 920** **B. 360** **C. 420** **D. 680**	**7. 12,231 + 4,222 = ☐** M6M21011009A **A. 15,453** **B. 16,453** **C. 16,433** **D. 16,463**

The **Numbers** domain assessed how well pupils understood basic numerical expressions, such as place value, numerical symbols, and the use of a number line ***(Figure 8).***

Figure 8: Sample Numbers questions, P4 and P6 mathematics

Primary 4	Primary 6
1. What is the value of 8 in 4,870? **A. Eight hundred** **B. Eighty** **C. Eight** **D. Eight thousand** M4M11011003A	**3. What is the expanded form of five thousand eight hundred and eleven?** **A. 50 + 800 + 10 + 1** **B. 500 + 800 + 11** **C. 5,000 + 800 + 10 + 1** **D. 500 + 800 + 11** M6M13011002A

The **Measurement** domain involved understanding basic measurement and applying measurement skills ***(Figure 9).***

Figure 9: Sample Measurement questions, P4 and P6 mathematics

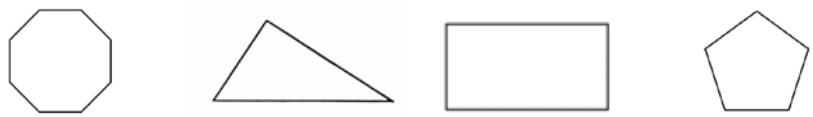
Primary 4	Primary 6
33. Aku has a 45kg bag of sugar. Afia has a 35kg bag of sugar. How much do the 2 bags weigh in kg? **A. 80kg** **B. 53kg** **C. 54kg** **D. 10kg** M4M42013002A	**18. Kofi uses 8cm of wire to make a square and 6cm to make a triangle. How many centimetres of wire will he use to make 2 squares and 1 triangle?** **A. 14cm** **B. 17cm** **C. 20cm** **D. 22cm** M6M23014001A

The **Shape and Space** domain ***(Figure 10)*** involved understanding the basic properties of plane and solid shapes and using the skills to evaluate the relative size of shapes and spaces.

Figure 10: Sample Shape and Space questions, P4 and P6 mathematics

Primary 4

26. Which of these shapes has only four lines?



P **Q** **R** **S**

M4M31011002B

A. P

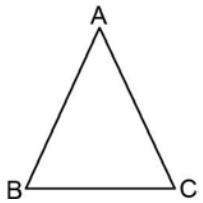
B. Q

C. R

D. S

Primary 6

30. In the triangle, if only the angles at B and C are equal, which two lines of the triangle are equal?



M6M31021002A

A. AB and AC

B. AB and BC

C. BC and AC

D. BA and BC

The **Data and Chance** domain required applying mathematics operations to data to perform 'real life' mathematics problems and finding out how certain real life events occur. For example, see ***Figure 11:***

Figure 11: Sample Data and Chance questions, P4 and P6 mathematics

Primary 4

40. If 🧍 stands for one boy, which picture graph matches the table?

Class	KG	P1	P2	P3
Number of boys	4	3	4	2

M4M51012003B

A.

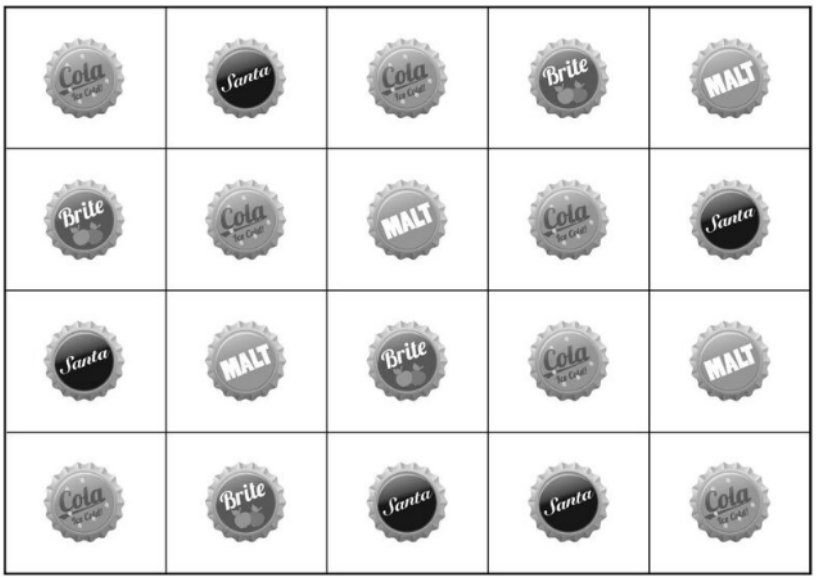
B.

C.

D.

Primary 6

43. What is the chance that a bottle top picked is Cola?



M6M53013001A

A. $\frac{3}{20}$

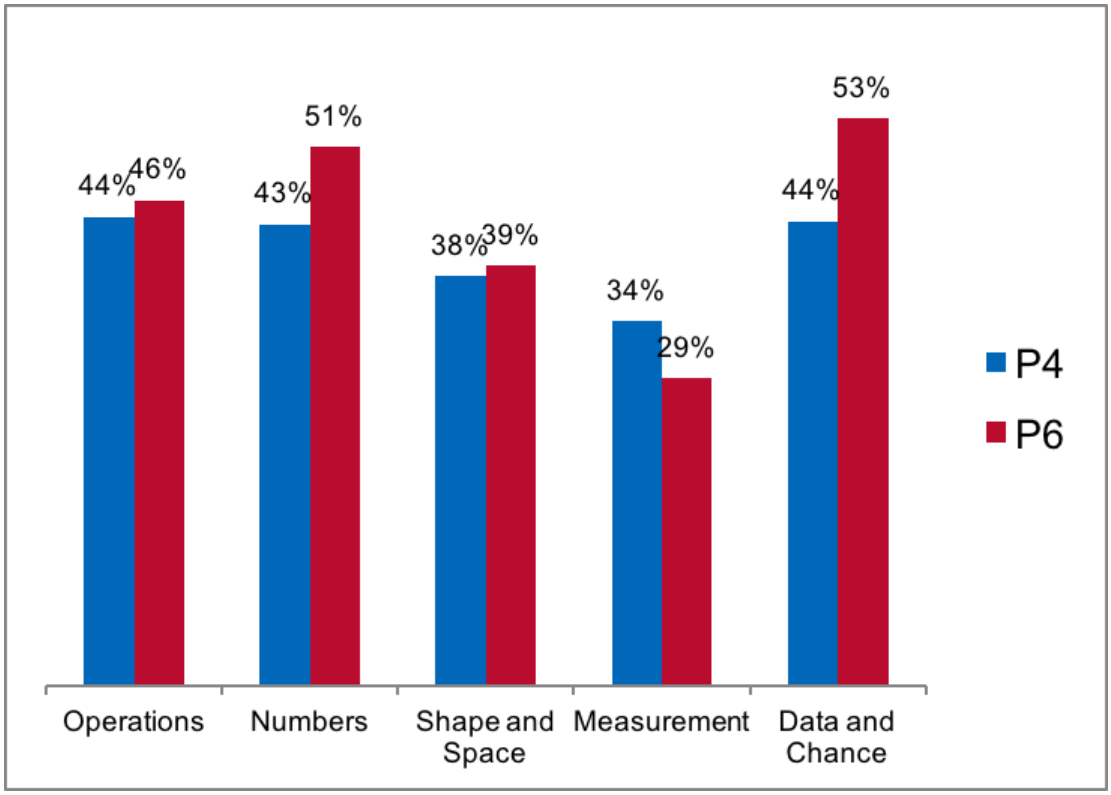
B. $\frac{4}{20}$

C. $\frac{5}{20}$

D. $\frac{7}{20}$

The average performance scores of pupils in each of the mathematics domains, based on percentage correct scores, are given in ***Figure 12.*** Pupils demonstrated having the most difficulty with tasks in the Measurement domain and the Shape and Space domain, with Measurement tasks being particularly challenging (34% for P4 and 29% for P6) compared to Shape and Space (38% for P4 and 39% for P6).

Figure 12: Average (% correct) scores by domain—Mathematics



Results by Core Demographic Variables

Sex

As shown in Figure 2 above, mathematics was quite challenging for both P4 and P6 pupils. This held true for boys and girls***. Table 4*** contains the mean percentages of items correct by male and female, including confidence intervals and significance levels. It can be seen that boys and girls performed similarly in P4 mathematics. Although girls and boys also struggled with P6 mathematics, males outperformed females by two percentage points, a difference that, while not substantive, is statistically significant. Girls and boys performed similarly in P6 English. For P4 English, the reverse was true. Female pupils outperformed males in P4 English by about two percentage points, a difference which is, again, statistically significant although not large. These findings do not indicate evidence of a great disparity in instruction by sex. Both males and females struggle with mathematics.

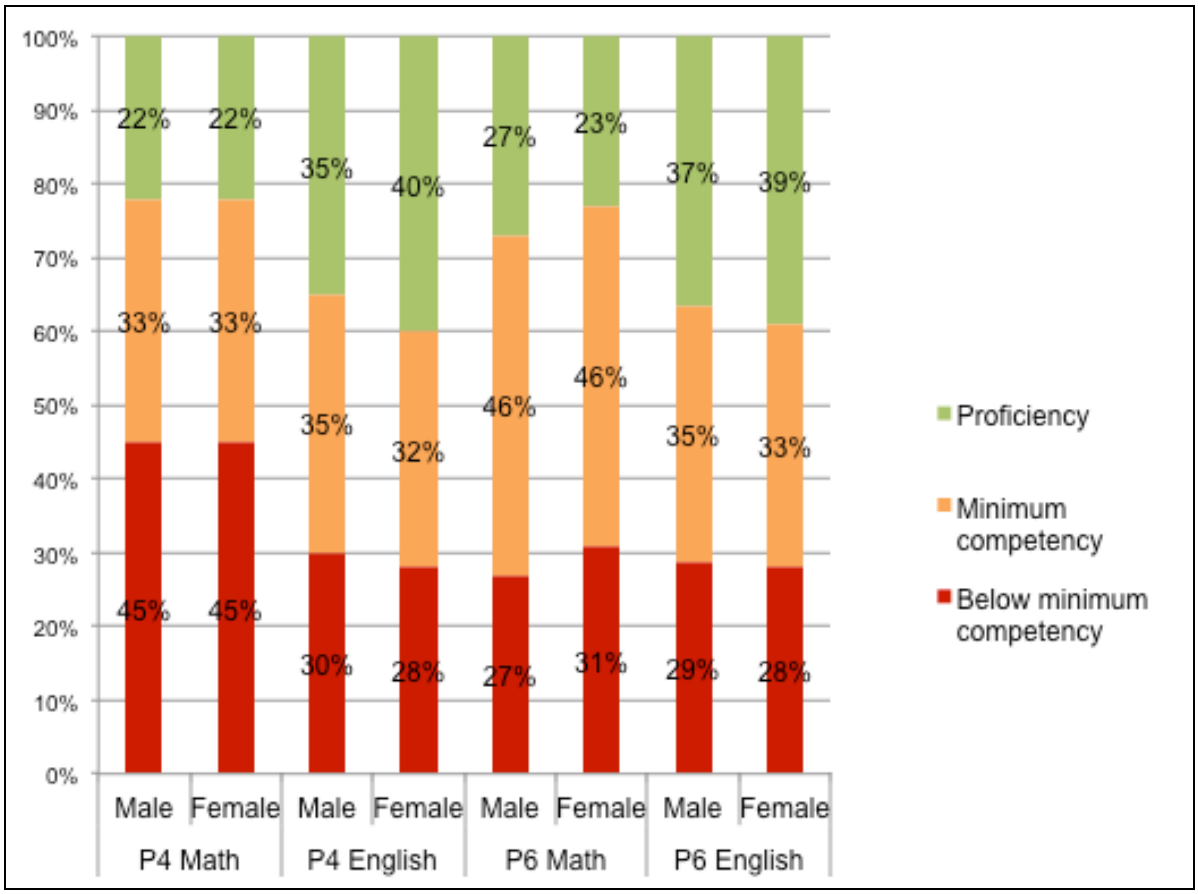
Table 4: Mean percentage of items correct, by sex

Subject and grade	Male		Female	
	Mean % correct	95% confidence interval	Mean % correct	95% confidence interval
Mathematics				
P4	41.9%	(40.69, 43.01)	41.5%	(40.36, 42.73)
P6	44.9%***	(43.93, 45.80)	42.8%^	(41.89, 43.73)
English				
P4	49.8%^	(48.34, 51.28)	52.0%***	(50.47, 53.50)
P6	47.6%	(46.15, 49.0)	48.1%	(46.64, 49.59)

^ = reference value; ***$p < 0.001$.

The percentages of pupils achieving minimum competency and proficiency levels according to sex are given in ***Figure 13.***

Figure 13: Percentages of pupils achieving minimum competency and proficiency levels, by sex



Urban vs. Rural

Performance gaps between children from rural versus urban areas were considerable and statistically significant for both P4 and P6 and for English and mathematics. For all tests, children from urban areas outperformed children from rural areas. ***Table 5*** breaks down the overall means of items correct by school location. There was no significant change in the performance gap between pupils in urban vs. rural schools from P4 to P6 in mathematics (the gap narrowed by 2.2 percentage points) or from P4 to P6 in English (the gap widened by half a percentage point [0.5]).

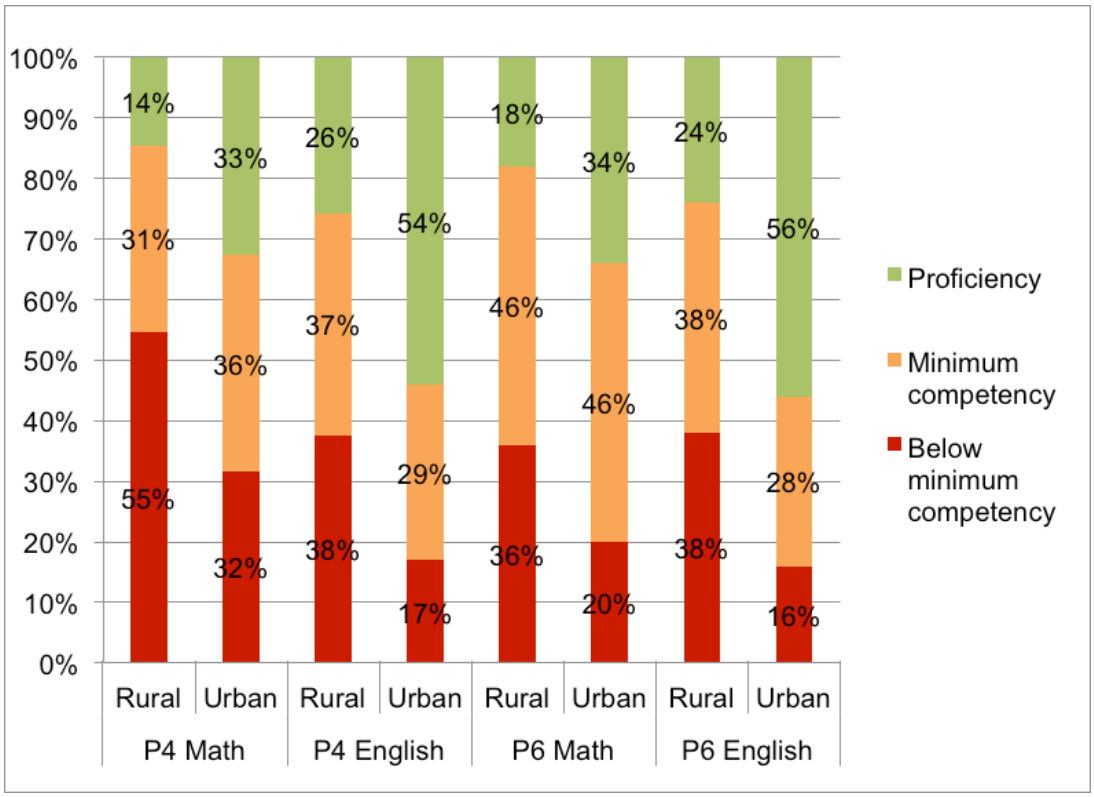
Table 5: Mean percentage of items correct, by school location

Subject and grade	Rural		Urban	
	Mean % correct	95% confidence interval	Mean % correct	95% confidence interval
Mathematics				
P4	37.92%^	(36.53, 39.31)	47.04%***	(45.26, 48.83)
P6	40.81%^	(39.75, 41.88)	47.77%***	(46.42, 49.13)
English				
P4	45.17%^	(43.40, 46.94)	59.04%***	(56.77, 61.31)
P6	41.59%^	(39.99, 43.18)	55.97%***	(53.75, 58.20)

^ = reference value; ***$p < 0.001$.

Figure 14 presents the percentage of pupils achieving the NEA performance levels across rural and urban locations. The gap between rural and urban performance for both P4 and P6 English was particularly large. More than half the pupils in P4 and P6 (54% and 56% respectively) in urban areas achieved proficiency, versus less than 27% in rural areas (24% for P6 and 26% for P4). Only 14% of P4 pupils and 18% of P6 pupils from rural areas achieved proficiency in mathematics. The percentage of children who could not answer 35% or more of the items – that is, the percentage who did not achieve minimum competency – was consistently higher among pupils from rural areas than among pupils from urban areas.

Figure 14: Percentages of pupils achieving minimum competency and proficiency levels, by school location



School Type

The NEA is administered in both private and public schools in Ghana. A consistent trend is that pupils attending private schools have much better results than pupils attending public schools. Still, as can be seen in ***Table 6,*** even private school average scores fell below 70% correct. There was a slight change in the performance differences between pupils in public vs. private schools from P4 to P6 in English (the performance gap narrowed by 2.8 percentage points). The change in difference was greater in mathematics, where the performance gap between school types narrowed by 5.9 percentage points from P4 to P6. The reason for this is not revealed by the NEA data.

Table 6: Mean percentage of items correct, by school type

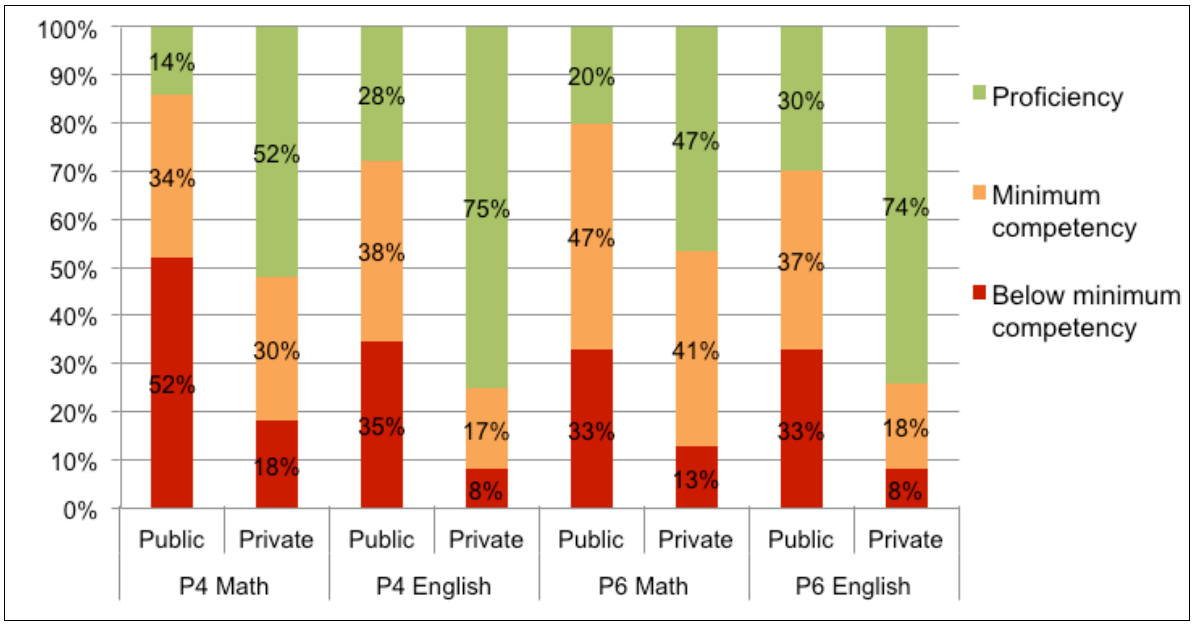
Subject and grade	Public		Private	
	Mean % correct	95% confidence interval	Mean % correct	95% confidence interval
Mathematics				
P4	38.14%^	(37.22, 39.06)	55.63%***	(53.10, 58.16)
P6	41.63%^	(40.75, 42.51)	53.17%***	(51.15, 55.20)
English				
P4	46.12%^	(44.74, 47.50)	69.64%***	(66.76, 72.52)
P6	43.87%^	(42.43, 45.32)	64.56%***	(61.81, 67.32)

^ = reference value; ***$p < 0.001$.

With the exception of P6 mathematics, the majority of pupils in private schools in 2016 reached the proficiency level in both grades and subjects ***(Figure 15***). In P4 mathematics, 52% of pupils in private schools achieved proficiency in 2016, as opposed to 14% in public schools. In P6 mathematics, 47% of pupils in private schools achieved proficiency as opposed to 20% in public schools. In P4 English, 75% of pupils in private schools achieved proficiency as opposed to 28% in public schools. In P6 English, 74% of pupils in private schools achieved proficiency as opposed to 30% in public schools.

The contrast in public and private school performance is even more striking in comparisons of the relative proportion of pupils achieving at least minimum competency – that is, when the pupils reaching proficiency levels and minimum competency levels are combined. It can be seen in Figure 15 that, unlike pupils attending public schools, more than 80% of the pupils in private schools reached at least the minimum competency level in both grades and subjects. The percentage of private school pupils achieving at least minimum competency in English was 92% for both P4 and P6, compared to the performance of public school pupils, where only 66% of the P4 pupils and 67% of the P6 pupils achieved at least minimum competency. While the percentage of private school pupils achieving at least minimum competency in mathematics was 82% for P4 and 88% for P6, only 48% of the public school pupils in P4 and 67% of the public school pupils in P6 achieved at least minimum competency in mathematics. The proportion who were not able to answer at least 35% of the questions (i.e., achieve minimum competency) was high for public school pupils, with a third or more of the pupils falling below the minimum competency cut-off. Over half of the public school pupils (52%) were unable to answer at least 35% of the questions in P4 mathematics. It can be seen in Figure 15 that this result contrasted strongly with the percentages of private school pupils who did not reach the minimum competency levels.

Figure 15: Percentages of pupils achieving minimum competency and proficiency levels, by school type



Deprived Districts

Since 1999, Ghana's government has classified roughly one third of the districts as *deprived*, based on various education outcome and resource indicators, including GER in primary, gender parity, seats and core textbooks per pupil, share of schools needing major repairs, Basic Education Certificate Examination (BECE) pass rates in both English and mathematics, per pupil expenditure in primary, pupil–teacher ratio in primary, and the share of qualified primary teachers. The majority of districts in Ghana that are classified as 'deprived' are in one of the three northern regions (Northern, Upper East, Upper West).

The performance of pupils attending public schools in deprived districts and districts that are not deprived is presented in ***Table 7.*** As expected, performance among pupils attending schools in the deprived districts was much lower than that of pupils attending schools in not-deprived districts, and these differences were statistically significant. Differences in performance between pupils in deprived vs. non-deprived districts did not significantly change from P4 to P6. There was a very slight change in performance from P4 to P6 in mathematics (the gap narrowed by less than one percentage point [0.9]). Similarly, in English the change was minor (the gap widened by less than one percentage point [0.7]). Thus, the gap in performance remained essentially the same for both grades.

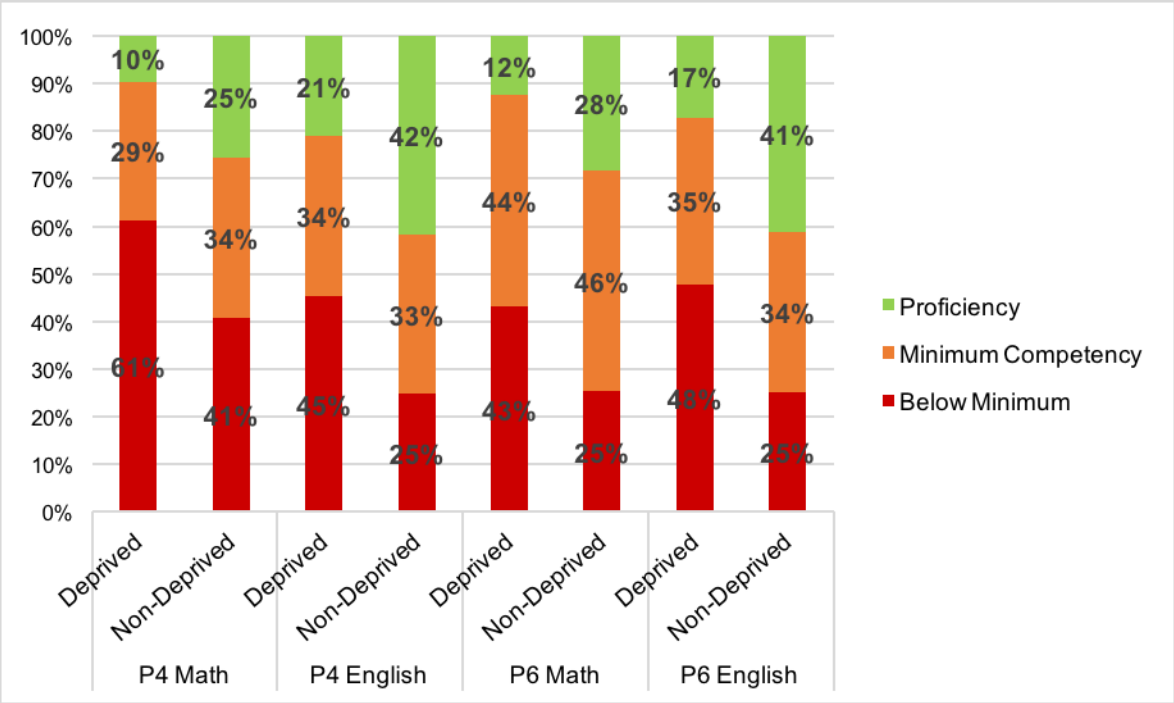
Table 7: Mean percentage of items correct, by deprived and non-deprived district status

Subject and grade	Non-deprived		Deprived	
	Mean % correct	95% confidence interval	Mean % correct	95% confidence interval
Mathematics				
P4	43.50%^	(42.16, 44.85)	35.06%***	(33.38, 36.74)
P6	45.42%^	(44.40, 46.43)	37.89%***	(36.58, 39.21)
English				
P4	53.31%^	(51.62, 55.00)	42.08%***	(39.29, 44.87)
P6	50.35%^	(48.76, 51.93)	38.37%***	(36.13, 40.62)

^ = reference value; ***$p < 0.001$.

For all grades and subjects, the proportion of pupils from deprived districts achieving proficiency, answering at least 55% or more of the questions correctly, was half that of pupils from non-deprived districts (Figure 16). The most striking comparison was that of P6 English. The percentage of children attending schools in deprived districts that reached or surpassed the cut-off for proficiency was 17%, in comparison to 41% for pupils attending schools in the non-deprived districts. It is also important to highlight the relatively large proportion of pupils from deprived districts who fell below the cut-point for minimum competency compared to pupils from non-deprived districts. In both grades and subjects, approximately 20 percentage points more of the pupils from deprived districts fell below the cut-point for minimum competency (i.e., less than 35% correct) than pupils from non-deprived districts.

Figure 16: Percentages of pupils achieving minimum competency and proficiency levels, by deprived and non-deprived district status



Regions

Table 8 presents the mean pupil performance by region. With the exception of performance in English in Greater Accra region, the average score was below 55% correct in all regions and subjects. The higher performance in English in Greater Accra region is not altogether surprising given the greater exposure to English that pupils in this metropolitan area have, compared to pupils in other regions. This is borne out by the findings of the 2015 National EGRA, in which public school pupils in P2 also performed better in English in this region, although reading ability was still very low overall. These regional findings clearly show that pupils across Ghana are struggling to perform grade-level tasks.

Table 8: Mean percentage of items correct, by region

Region and subject	Mean score (%)	95% confidence interval	Region and subject	Mean score (%)	95% confidence interval
P4 mathematics			**P6 mathematics**		
Ashanti***	42.9	(39.2, 46.5)	Ashanti***	45.2	(42.6, 47.7)
Brong Ahafo***	40.6	(37.2, 44)	Brong Ahafo***	42.7	(39.9, 45.4)
Central***	38.7	(36.1, 41.3)	Central***	43.3	(41.5, 45.1)
Eastern***	41.5	(38.9, 44.2)	Eastern***	42.9	(40.8, 45)
Greater Accra^	53.5	(51.2, 55.9)	Greater Accra^	53.0	(51.1, 55)
Northern***	35.7	(32.4, 38.9)	Northern***	36.8	(34.2, 39.4)
Upper East***	36.6	(32.8, 40.3)	Upper East***	42.2	(39, 45.5)
Upper West***	31.6	(30.4, 32.8)	Upper West***	36.6	(35.3, 38)
Volta***	42.2	(38.8, 45.6)	Volta***	44.0	(41.1, 46.8)
Western***	41.1	(36.9, 45.3)	Western***	42.3	(38.9, 45.7)
P4 English			**P6 English**		
Ashanti***	51.9	(47.6, 56.2)	Ashanti***	48.7	(44.7, 52.6)
Brong Ahafo***	46.7	(41.8, 51.6)	Brong Ahafo***	43.9	(39.2, 48.5)
Central***	48.2	(44.6, 51.9)	Central***	45.9	(42.7, 49.1)
Eastern***	50.2	(46.5, 53.9)	Eastern***	46.4	(43.2, 49.6)
Greater Accra^	70.3	(67.2, 73.4)	Greater Accra^	66.7	(64, 69.5)
Northern***	46.2	(41.2, 51.2)	Northern***	40.6	(36.2, 45)
Upper East***	40.5	(35.9, 45.2)	Upper East***	40.6	(35.6, 45.6)
Upper West***	36.4	(34.2, 38.6)	Upper West***	37.0	(35.1, 39)
Volta***	50.3	(45.7, 54.9)	Volta***	46.8	(43, 50.6)
Western***	48.5	(43.5, 53.6)	Western***	44.9	(39.3, 50.5)

^ = reference value; ***p = 0.001.

Table 9 presents the regional performance, as described by the distribution of pupils in each region across the three performance categories: below minimum competency, minimum competency and proficiency. Regional performance is presented for each grade and subject. There was considerable variability in performance across regions, with distinctly lower performance for pupils attending schools in the three regions of northern Ghana (shaded in blue) compared to pupils attending schools in other regions of the country. As in previous NEA administrations, in general, pupils attending schools in Greater Accra were shown to outperform those in the other regions of the country. The relative proportion of pupils who achieved proficiency was highest in Greater Accra, for both grades and both subjects.

Table 9: Percentages of pupils achieving minimum competency and proficiency levels, NEA 2016, by region

Proficiency level by grade and subject	Regions									
	Ashanti	Brong Ahafo	Central	Eastern	Greater Accra	Northern	Upper East	Upper West	Volta	Western
P4 Mathematics										
Below Minimum Competency	42.9	47.2	49.7	43.2	18.0	61.0	58.5	71.5	44.2	46.2
Minimum Competency	33.9	32.8	35.1	36.6	35.2	26.6	28.4	23.5	32.3	34.2
Proficiency	23.3	20.0	15.2	20.2	46.8	12.4	13.2	5.0	23.4	19.5
P4 English										
Below Minimum Competency	26.2	36.9	28.6	28.8	5.7	34.7	50.4	56.1	29.7	31.1
Minimum Competency	34.6	33.3	39.6	36.3	19.7	37.5	33.1	33.1	33.8	35.5
Proficiency	39.2	29.8	31.8	35.0	74.6	27.8	16.5	10.8	36.5	33.4
P6 Mathematics										
Below Minimum Competency	25.3	31.5	28.8	30.3	11.7	46.4	32.1	48.5	29.2	30.5
Minimum Competency	48.0	46.6	48.8	47.2	39.9	43.0	47.7	41.1	45.1	48.8
Proficiency	26.7	21.9	22.4	22.5	48.3	10.7	20.2	10.5	25.7	20.7
P6 English										
Below Minimum Competency	24.4	34.6	28.7	28.3	5.1	40.0	44.1	46.2	29.6	31.9
Minimum Competency	36.6	35.6	37.8	36.1	16.3	37.2	33.8	39.7	33.0	37.2
Proficiency	39.0	29.8	33.5	35.6	78.6	22.8	22.1	14.1	37.4	30.9

Although performance in all three of the northern regions was characteristically low in comparison to the other regions of the country, it is clear that pupils attending schools in Upper West were at an even greater disadvantage. The percentage of pupils attending schools in Upper West who reached the proficiency level (5%) was approximately half that of those from Northern and Upper East. The one exception to this was in the subject area of mathematics for P6 pupils. In P6 mathematics, both Northern Region (10.7%) and Upper West (10.5%) showed very low percentages of pupils to achieve proficiency, as compared to Upper East (20.2%).

IV. Key Findings and Conclusions

Key Findings

The results of the 2016 NEA showed clearly that the performance of P4 and P6 pupils was generally low. There has been no significant or substantive change in pupil performance since the 2013 NEA.[9] The highest overall mean score was 50.9% correct, in P4 English. In public schools, the proficiency rates (e.g., percentage of pupils achieving proficiency, or answering at least 55% of the questions correctly) achieved across grades and subjects ranged from 14% for P4 mathematics to 30% for P6 English. One-third of the public school pupils or more (ranging from 33% in P6 English to 52% in P4 mathematics) did not achieve a minimum competency level of performance, which required answering at least 35% of the questions correctly.

Although the scores were low for both English and mathematics, mathematics seemed to present a greater challenge to Ghanaian pupils, in both public and private schools. In mathematics, pupils tended to have considerably more difficulty understanding the concepts of measurement and applying these to solving mathematical problems. A second area that challenged pupils was the understanding of the properties of two- and three-dimensional shapes and using these concepts to evaluate the relative size of shapes and spaces.

Measurement and Shape and Space are two domains of mathematics learning that are included in the Ghanaian primary school curriculum. The EGMA provided some clues as to why P4 and P6 pupils were having difficulty with these skills. Findings from the national EGMA studies in 2013 and 2015 showed that pupils' best performances were on basic mathematical operations that involved rote learning. Understanding and applying mathematical concepts was observed to be limited. In the later grades, it is important for pupils to understand fundamental mathematical concepts and to apply these concepts to complete higher-level operations and solve mathematical problems. Failing to grasp basic mathematical concepts in the early grades presents an obvious barrier to grasping the concepts and skills required of pupils in the later grades, such as those in the higher level domains tested in the NEA.

Delayed development of pre-reading and reading skills in the early grades is also likely to be impacting performance in later-grade English and mathematics. If pupils are not able to read when they take the NEA, both English and mathematics performance will be negatively impacted. The NEA is a paper-and-pencil task conducted in a group setting. Literacy is

[9] The comparison with 2013 is only for P6, given that P4 pupils have not been tested before 2016.

required on most of the NEA tests, including on the mathematics forms, for which many of the test items require some reading.

Conclusions

In conclusion, it is clear that a concerted effort needs to be made to ensure that when pupils reach P4, they have the foundational skills they need to complete the more advanced tasks of mid- and upper primary English and mathematics. In the early grades, pupils need to learn basic mathematics concepts and apply them as they attempt more complex operations and problem-solving tasks. The development of pre-reading skills that lead to reading with comprehension is required for pupils to independently perform well on the P4 and P6 English and mathematics tests that constitute the NEA. The 2015 National EGRA and EGMA findings revealed that most pupils across Ghana's public schools were finishing P2 without even basic literacy skills, let alone the ability to read with fluency and comprehension. In addition, pupils' mathematics skills at the end of P2 were lacking in the conceptual understanding needed to perform more difficult tasks at higher grades. These findings in the early grades combined with the NEA findings in P4 and P6 convey a need for better instruction in literacy and mathematics.

The Ministry of Education and GES recognise the disparities in the quality of education in rural versus urban settings and in particular the most deprived districts and the three northern regions of Ghana. The 2015 EGRA and EGMA findings also revealed significant lack of resources (reading and mathematics textbooks and exercise books), especially in these areas. Continued and more effective efforts to reduce these disparities are needed, with an aim to ensure that the foundational skills of the early primary grades are acquired by all pupils – pupils attending schools in the most rural and remote regions of the country, as well as pupils attending schools in urban locations.

Other key findings are based on comparisons across a variety of subpopulations. Boys and girls performed similarly in P4 mathematics and P6 English. However, a slightly greater percentage of boys than girls achieved proficiency for P6 mathematics. For P4 English, the reverse was true. Girls outperformed boys in P4 English.

For all grades and subjects, pupils attending schools in urban regions performed significantly higher than pupils attending schools in a rural area. This finding was substantiated in the observations of regional differences. The proportion of pupils attending schools in the three northern regions of Ghana who achieved proficiency on the NEA tests was consistently and markedly less than the proportion of pupils residing in any other region of the country. Proficiency rates in the three northern regions were much lower than in the other regions of the country, for both grades and both subject areas. As in previous NEAs, the performance of pupils attending schools in the Greater Accra region, the most urban region of the country, outperformed pupils attending schools in any other region. For both grades and subjects, the proportion of pupils in the Greater Accra region achieving proficiency was approximately twice the proportion achieving proficiency in other regions of the country. Furthermore, the proficiency rates for the Greater Accra region were approximately three times the proficiency rates for Northern Region and Upper East and four times the proficiency rates for Upper West. The contrast in performance of pupils attending schools in the deprived versus non-

deprived districts also reinforces the finding that rural settings are plagued with very real barriers to learning.

Annex A: Technical Details on Test Item Development

Test Frameworks, Item Development, and Pilot Test Assembly Meetings for Ghana National Education Assessment (NEA) 2016 Report of Meetings held in Accra

Monday March 28th to Friday April 16th 2016

Introduction

The workshop described in this report was the second of two workshops designed to support the Ghana National Education Assessment Unit (NEAU) in developing operational tests for the NEA 2016. The workshop was held at the Institute for Local Government Studies, Madina, Greater Accra. The earlier workshop had focused on test frameworks, test blueprints, item specifications for item development, item development and review, and pilot test forms assembly.

The pilot test itself was administered under NEA leadership in the interim period between the two workshops (December 2015 – March 2016). During this period, pilot data were collected, cleaned and submitted to RTI for analysis. Pilot test data analyses were conducted with the intention of explaining the methodology and outcomes to participants at the second workshop.

The second workshop was held between Monday March 28th and Friday March 16th, a 3-week period which was divided into 3 sections each with a clear major objective leading into the next section or phase of test development. The 3 objectives were:

1. Develop item cards for all pilot-tested items with complete psychometric and content information (required for operational test forms assembly);
2. Select items for all operational forms (required for operational forms layout);
3. Produce print-ready operational forms (required for delivery to the printing contractor).

The agenda for the second workshop is provided in Table 1 below.

Table 1: Workshop 2 Agenda

Week	Dates	Topic
1	March 30 – April 1	NEA 2016 agenda and timelines Goals of pilot testing Analyses on pilot test data Review structure of NEA 2016 operational tests Review NEA 2016 test blueprints and test maps Align 2013 blueprints/test maps with 2016 blueprints/test maps Review items cards and print
2	April 4 - 8	Present item bank/test assembler Calculate number of good items in bank Calculate number of pilot items needed for embedding in 2016 operational forms Calculate number of operational forms required Select items for NEA 2016 Select items for piloting

Week	Dates	Topic
3	April 11-15	Assemble electronic version of all 2016 operational forms Review forms and ensure they are print-ready Discuss remaining technical issues for operational administration

Goals of Pilot Testing

Pilot testing within the context of tests of the type conducted by the NEAU is frequently misunderstood as an activity providing information on pupil, class, school or indeed any group performance in the subject areas measured. Equally, test level information – difficulty, correlations, and others – is of little interpretable value for its own sake in a pilot test. The explanation is clear: pilot tests are made up of items the quality of which is unknown and it is inevitable that a test under development will be composed of a range of poor-functioning to well-functioning items. While analyses are made at the test, pupil and group levels, these analyses are used only to understand item performance, the main focus and objective of pilot testing. The NEA 2016 tests contained multiple choice items only, and the goal of pilot testing was to determine for each multiple choice test item piloted what the statistical or psychometric properties were for each option and for each item as a whole. Some of these analyses required that individual item performance was compared with whole test performance – in the cases of determining item correlation and item discrimination. The next section discusses the types of analyses that were conducted on pilot test data.

A second important goal of pilot testing is to determine if there are problems with test length and related timing. This was particularly important for the NEA 2016 because changes in test length were made to introduce “embedded field testing”. This design, intended to enable pilot-testing to take place during the operational administration (rather than in a more expensive and less efficient separate pilot testing event), resulted in the addition of 5 test items per test form, increasing the length from 35 to 40 items in P4 and 40 to 45 in P6, for both subject areas. Test time was correspondingly increased by 15 minutes (assuming that each test item takes on average between 2 and 2.5 minutes to respond to). Increased test length and test time were not viewed as having any significant effect on pupil performance; data analyses on item omits (the number of pupils who failed to respond to an item) for the last 5 test items do not show that there was any adverse effect of pupil performance on these items (omits ranged from about 1% of the total in language P4 to about 4% of the total in Math P6).

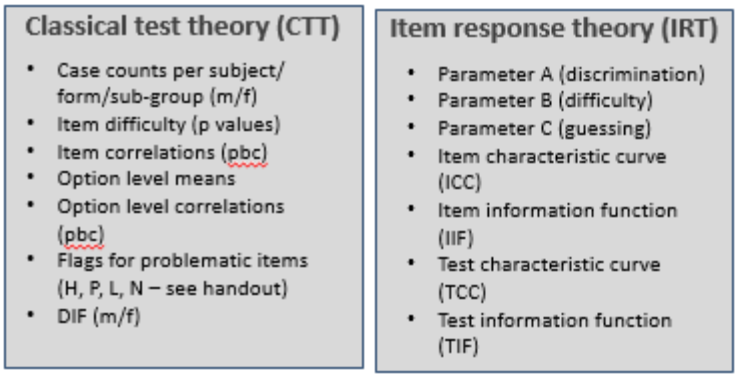
A third goal of pilot testing is a focus on logistical issues in test administration: use of the administration manual, administration of the oral listening component in the English language test, and completion of test booklets according to requirements described in the administration manual and pupil test booklets. Only one issue was reported as requiring significant attention: that of the administration of the listening component of the language test at both P4 and P6 levels. Of importance here was the change in test design from 2013 to 2016: in the 2013 operational administration, the oral listening prompt contained not only the stimulus but also the question posed about the stimulus (with only the item options to be found in the pupil booklet). In the 2016 pilot test, the oral listening prompt contained

only the stimulus and not the question posed about the stimulus (with the pupil booklet providing the item options as well as the question). The change from 2013 to 2016 was made on the grounds that a listening test should only be of the stimulus and not of the question posed of the stimulus; in a reading comprehension test the same holds – comprehension of the question is not the object of measurement, and this is achieved by ensuring that the language of the question is below the targeted level being measured. After the pilot-testing experience, it was agreed that the oral prompt should contain both stimulus and question, but that the question would also appear in a written form in the pupil booklet – thus rendering the true measure of listening of the stimulus only.

Analyses on Pilot Test Data

After pilot test administration and cleaning of all data files, the next step was to conduct pilot data analysis to determine the statistical quality of the items, which in turn would enable the team to determine which items could be included in the operational test form pool. During the data analysis process, Classical Test Theory (CTT) and Item Response Theory (IRT) were used to compute item difficulty and item discrimination indices. The analyses conducted under each methodology are listed in the following figure:

Figure 1: Classical and Item Response Theory Analyses Conducted on Data



Analysis under Classical Test Theory

Item Difficulty

Item difficulty represents the proportion of test takers who answered the item correctly. This index is also known as the item p-value. The possible range for the p-value is 0.0 to 1.0, where higher values indicate an easier item as a greater proportion of pupils answered it correctly. For an item to be considered acceptable, its difficulty should be between 0.3 and 0.9.

Item Discrimination

Item discrimination represents how well an item can distinguish between high-performing and low-performing test takers, in other words between those who know and those who do not know the content measured. The item-total correlation is used as a measure of item discrimination. The possible range for item discrimination is –1.0 to 1.0. If an item has a discrimination index below 0.0, the item may very well have a problem and should be scrutinized. A negative discrimination shows that high-performing pupils are getting the item wrong and low-performing pupils are getting it right. Often such a situation indicates to us that the item has likely been assigned the wrong key – we suspect that good pupils are

actually indicating what the correct key is, although this suspicion needs to be verified. It is important to consider that, if an item has a very high or very low difficulty index, its discrimination will be reduced – i.e., if most pupils get an item correct, that item fails to discriminate between different pupils; similarly if hardly any pupil gets an item correct, again the item is unable to discriminate between different pupil abilities. For an item to be considered acceptable, its discrimination should be greater than or equal to 0.25.

Option Analysis

For multiple choice items it is important to analyze option performance, i.e., that of the key and all distractors. For each option, analysts calculate the proportion of test takers that select a particular option and the point-biserial correlation for that option. The expected behavior is that the key is selected by the majority of test takers and that it has a positive point-biserial correlation. The distractors should be selected by fewer test takers and they should have negative point-biserial correlations.

During this analysis the following situations can be verified:

- **The key was not selected by most of the pupils:** this situation could indicate that the key might have been incorrectly identified.
- **Any of the options was selected by less than 2% of the test takers:** this situation suggests that the option is clearly incorrect even for low-performing pupils and therefore fails to discriminate between pupils of different levels of ability.
- **Positive correlation for any incorrect option:** this often suggests that the item has more than one correct option or that the key was incorrectly identified.
- **Negative correlation for the key:** this often suggests that the key was erroneously identified or that the item does not have a correct option.
- **The item was omitted by more than 20% of test takers:** this may indicate a problem with the readability of the item in the test booklet.

Analysis under Item Response Theory

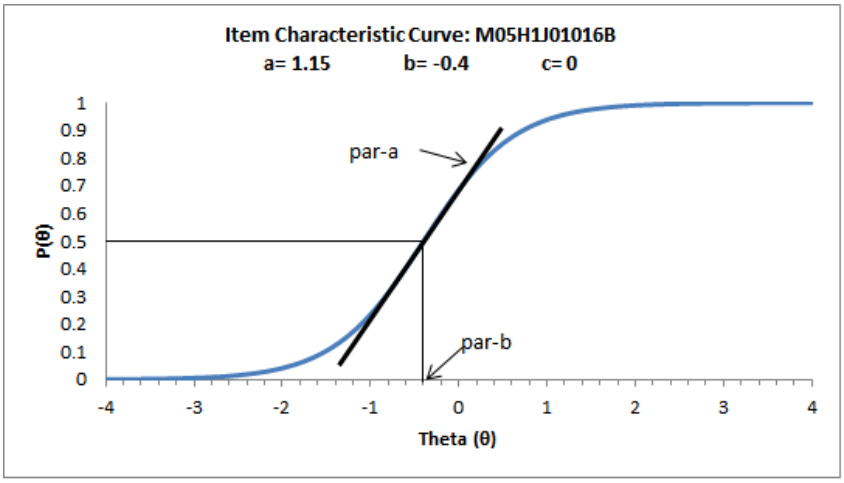
Item Response Theory (IRT) tries to establish a relationship between the latent trait that needs to be measured and the probability of responding to an item correctly. This relationship is represented in the Item Characteristics Curve (ICC), where the latent trait scale or ability, represented by θ on the x-axis, has practical values of -3.0 to 3.0, and the probability of responding to an item correctly, P(θ) on the y-axis, ranges from 0.0 to 1.0.

The shape of the ICC depends on the IRT parameters; a-parameter represents item discrimination, b-parameter represents item difficulty and c-parameter represents guessing. The IRT model used to analyze the items on the Ghana NEA 2016 pilot test was the 2 Parameter Logistic model where only the a-parameter (item discrimination) and b-parameter (item difficulty) are taken into consideration.

Item Difficulty (*b-parameter*)

The b-parameter indicates the item location on the ability scale, which represents a probability of 50% that a pupil will get the item correct. The higher the value of the b-parameter, the more difficult the item is considered to be. For an item to be considered acceptable, the b-parameter should be between –2.5 and 2.5.

Figure 2: Item Characteristic Curve (ICC)



Item Discrimination (*a-parameter*)

The a-parameter indicates the slope of the ICC, it shows how well an item differentiates between high-performing and low-performing test takers. The steeper the slope, the better the item is at separating test takers of different ability levels. For an item to be considered acceptable, the a-parameter should be greater than or equal to 0.5.

Differential Item Functioning (DIF)

DIF analysis helps to create fair assessments. The test takers are divided into groups (sex for example), matched by performance, and their performance on a particular item is compared. If the groups perform differently, there may be a real difference in ability between them or it may indicate that the item content is causing the difference to appear. On the basis of DIF analysis for a particular group (e.g., gender, home language, urban/rural, etc.), items are classified into one of three categories:

- Category A: little or no difference between groups.
- Category B: moderate difference between groups
- Category C: substantial difference between groups.

Items that show Category C DIF are not used in a test, unless there is some prevailing justification for why the item must be used (e.g., without the item, coverage of essential curriculum content is incomplete). Items in Category A and Category B can be used in test construction.

Item Banking and Test Assembly

The NEAU was provided with "Item Banking and Test Assembly" software to support the test construction process. The software has the following capabilities:

- **Item viewer**: allows test developers to select an item and see the item text, item statistics (classical and IRT), Item Characteristic Curve, and Item Information Function.
- **Test Assembler**: allows test developers to construct up to two test forms. The software also creates the Test Characteristic Curve and Test Information Function for

each test form. This functionality can be used to create equivalent test forms in terms of overall difficulty and information.

- **Item Banking**: allows test developers to store relevant statistics for each item such as difficulty, discrimination, option analysis, DIF, and item flags used to identify problematic items. The item flagging criteria are described in Table 3 below while the definitions of flag labels and desirable ranges are provided in Table 4:

Table 3: Item Flagging Criteria

	Bank file variable	Flag	Description
Item difficulty	pvalfl	PL	If P-value < .30 (PL)
		PH	If P-Value > .90 (PH)
	bparfl	BL	If b-parameter < -2.5 (BL)
		BH	If b-parameter > 2.5 (BH)
Item discrimination	itcfl	CL	If item-total correlation < 0.25 (CL)
	aparfl	AL	If a-parameter < 0.50 (AL)
Guessing	cparfl	CH	If c-parameter > .40 (CH)
Differential Item Functioning (DIF) ETS Categories	dif_MF	A	No or negligible DIF
		B	Moderate DIF
		C	High or substantial DIF
	fg_MF	M	Favoring Males
		F	Favoring Females
Option Analysis	optfl	H	If keyed option is NOT highest percentage (H)
		L	If any option ≤ 2% (L)
		P	If any non-keyed option pb-corr > 0.03 (P)
		N	If the keyed option pb-corr < 0 (N)
		B	If omit > 20% (B)

Table 4: Definitions of Flag Labels and Desirable Ranges

Flag label	Definition	Desirable range
PL	p-value low	p-value: between 0.3 and 0.90
PH	p-value high	
BL	b-parameter low	b-parameter: between -2.5 and 2.5
BH	b-parameter high	
CL	correlation low between item and total	a-parameter: greater than 0.5
CH	c-parameter high	
H	highest percentage is not a correct option	
L	lowest percentage of any option	
P	positive pb-correlation for any non-correct option	Item correlation: greater than 0.25
N	negative pb-correlation for the correct option	
B	blanks (omits) are over 20%	Omits: under 20%
A	no or negligible DIF	No DIF or B Category if balanced
B	moderate DIF	
C	substantial DIF	

Review Structure of NEA 2016 Operational Tests

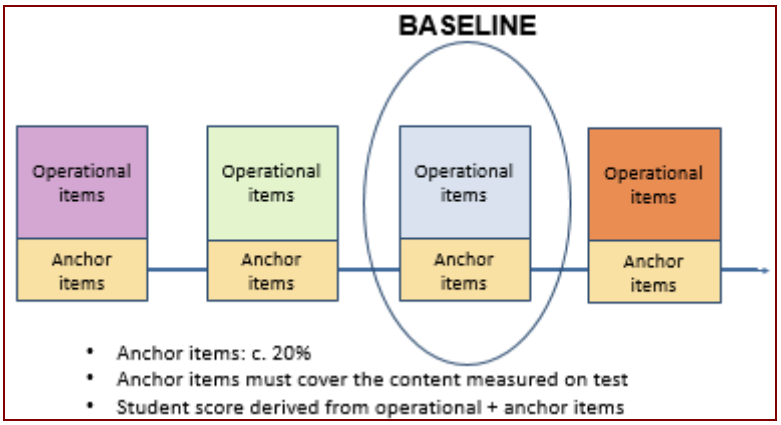
A number of new technical features of the NEA 2016 were presented to NEAU and associated partners in Workshop #1 held in Accra in November 2015. These technical features were intended to strengthen the design of the NEA tests and improve efficiency and quality from a number of respects. The features were approved by the NEAU and presented at the beginning of Workshop #2 to remind participants of what these features were and how they would have an impact on the structure of the tests being assembled for the NEA 2016 operational administration. The approved features, briefly discussed in the following section, are:

- Horizontal anchoring
- Embedded field test design
- Multiple forms per test
- Cognitive processing levels

Horizontal Anchoring**:** Tests, and particularly those that are sample-based assessments intended for national and sub-national educational policy design like the NEA, need to provide results that are reliably comparable from one year or administration event to the next. Comparison of test results cannot be derived from tests that repeatedly use exactly the same test items from one year to the next, not least because re-used test items do not provide an accurate measure of pupil performance. Therefore, test forms from one year to the next must use different items, and must be assembled to be parallel in both content measured and in terms of the psychometric properties of each item on the test. In addition, test forms are equated to enable all scores to be placed on the same scale, permitting accurate comparison of scores on one test compared with scores on another parallel test. Figure 3 below shows a test form for each of the past two administrations of the NEA in P6 (in 2011 and 2013), the current form being developed for 2016, and an upcoming form expected to be administered in 2018.

The bulk of the test items on each form, composed of operational test items, are coloured differently in the figure to signify that the items are a different set in each case, although

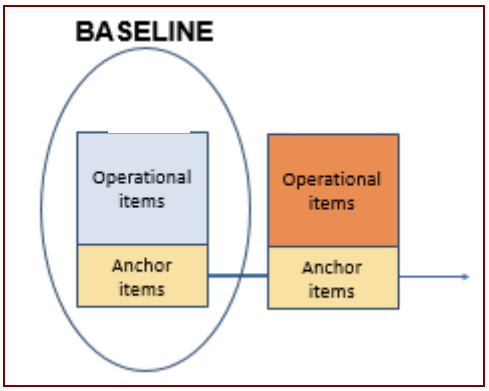
Figure 3: Horizontal Linking (Anchoring) for P6 (Language and Mathematics)



parallel in structure. Each form contains a sub-set of items, known as anchor items, which are common across all the forms from one year to the next. Anchor items are the basis for the methodology used to equate all forms and place test scores on the same scale so that valid comparisons can be made of test results. NEA 2016 was assembled in the manner described here to enable valid comparisons of test results from the current administrations with those of the two previous administrations (where possible), and with any upcoming administrations.

Horizontal anchoring for the P4 set of tests will not be possible with previously administered tests since in both 2011 and 2013 the grade of focus was P3 (see Figure 4 below). However, horizontal linking will be necessary between the 2016 forms to be assembled in P4 and those of any ensuing years of administration. Additionally, since multiple forms of P4 tests in both Mathematics and English language were used for the 2016 administration, these will also need to be anchored.

Figure 4: Horizontal Linking (Anchoring) for P4 (Language and Mathematics)



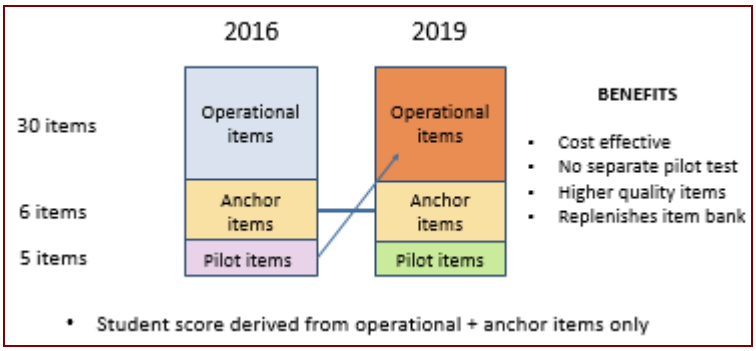
Embedded Field Test Design**:** In 2013, test items were developed at the beginning of the testing cycle and then pilot-tested, half-way through the school year, to determine their suitability for inclusion in operational test forms. The NEA is an end-of-grade test and the operational administration is conducted at the end of the targeted grades, i.e., typically in

the month of July. Pilot-testing as a separate event must therefore take place a number of months before the end of the school year to give time for pilot-test analysis, assembly of operational test forms and other related tasks. There are a number of problems that arise with pilot-testing conducted half way through the school year:

- In the middle of the school year teachers have not covered all of the content that is measured on the test and therefore pupils of a targeted grade are not suitable candidates for taking pilot tests;
- In order to avoid this problem, pilot-testing is conducted among pupils from a grade above that of the targeted grade; this is inconvenient from a number of respects not least of which is the fact that pilot test performance statistics are obtained from pupils who are not from the targeted grade, and who are not currently being instructed in the content that is measured on the test;
- Pilot-testing at a time in the year that does not correspond to the time of the year of application of the operational test, and with pupils who are not from the same grade, has a negative effect on item performance statistics, rendering statistical information unreliable for assembling operational test forms;
- Conducting pilot-testing as a separate event is a costly enterprise.

More common in assessment practice these days is the use of an approach known as "embedded field test design" which moves all pilot-testing to the operational forms administration event (see Figure 5 below). The methodology involves including a small number of test items among the operational items that are only used for item validation purposes (generating performance statistics). Pupil scores on the test forms in this design are derived from the operational items together with the anchor items but not from the pilot test items. Benefits from using the embedded field test design are two-fold: (a) significant cost reductions; and (b) higher quality item performance statistics.

Figure 5: Embedded Field Test Design

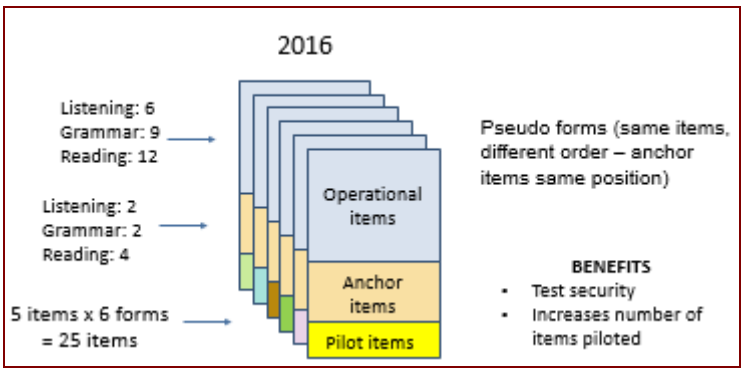


Multiple Forms per Test**:** The consequence of introducing an embedded field test design into the structure of the NEA test means that multiple forms of the test must be designed for the operational administration in order to absorb the number of items that have to be piloted. Having multiple forms allows for a sufficient number of items for piloting to be validated through the operational administration. The number of forms will depend on the number of items that need to be piloted tested, which in turn depends on the state of the item bank. Figure 6 below provides an example, for illustrative purposes only, of 6 forms each containing 5 new items for pilot-testing. One would probably NOT want to increase the size

of the test beyond approximately 5 new items because the length of the test would affect pupils' performance.

It is important to point out that the multiple forms required are what are known as "pseudo" forms, that is, forms that contain the same operational items but reorganized to give the appearance of being different. For the 2016 NEA, the anchor items were the same on each form and to the extent possible occupied the same position on the test. New items were completely different on each pseudo-form. Lastly, having multiple forms administered within a school or class significantly improved administration security by reducing instances of copying.

Figure 6: Multiple Forms per Test



Cognitive Processing Levels**:** Test items are developed to accurately measure 2 major issues: (1) content standards (i.e., what is expressed in the curriculum for a particular subject area and grade, and forms the "content" of what teachers should teach and pupils learn); and (2) cognitive ability, that is the skills that pupils develop in acquiring, processing, and using information. Content skills and knowledge and cognitive skills interact, and are measured in a controlled fashion by assessment items that are designed to measure a specific content standard within a specific cognitive ability. The purpose is to ensure that a test covers an appropriate range of content and range of easy to complex cognitive skills. In this way developers control the difficulty level of an item in a purposeful way. It is also possible to easily disaggregate test data to clearly defined sub-tests or item groupings.

During workshop #1 conducted in November 2015, agreements were made for definitions of cognitive processing levels which were applied to item development during the same workshop. As can be seen in Table 5 below, a discrepancy exists between the application of

Table 5: Cognitive Processing Levels Applied to the NEA Pilot Test of 2015

DOMAIN	Knowledge	Understanding	Application
		GENERAL	
	Recall or location of a fact, information, or procedure, without necessarily understanding the underlying concepts	Understanding of underlying concepts	Use of knowledge and concepts for the resolution of a problem

DOMAIN	Knowledge	Understanding	Application
ENGLISH LANGUAGE – P4 and P6			
Reading comprehension	At the knowledge level, there is a direct textual match between what is in the text and what the answer (key) is (i.e., is explicitly expressed in the text)	At the understanding level, the answer (key) is a paraphrase of what is in the text	At the application level, there is a need to piece together information from different sources within the text, or to interpret implicitly expressed information
Listening	Listening to directions and instructions at this level is based on listening to 1 piece of information or 1 step In listening to stories at this level, as with reading, there is a direct textual match between what is in the text and what the expected answer is	In listening to directions and instructions at this level, the text describes a 2-step process or 2 pieces of information In listening to stories at this level, as with reading, the answer (key) is a paraphrase of what is in the text	In listening to directions and instructions at this level, the text involves at least 3 steps or pieces of information In listening to stories at this level, as with reading, the listener must piece together information from different sources within the text, or interpret information that is implicit in the text
Grammar	At this level, involves recall of simple grammatical rules, without necessarily understanding the underlying concepts	At this level, involves demonstrating an understanding of underlying concepts	At this level, involves demonstrating an ability to apply underlying concepts to a specific context
MATHEMATICS – P4 and P6			
DOMAIN	**Knowledge and understanding**	**Application**	**Reasoning**
Numbers and numerals	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 for logical, systematic thinking; includes intuitive and inductive reasoning based on patterns and regularities that can be used to arrive at solutions to non-routine problems (IEA definition, TIMSS 2011 Framework)
Basic operations			
Collect and handle data			
Measurement			
Space and shape			

a cognitive scheme for defining levels of cognitive processing for English language and that used for Mathematics. It was recognized at the pilot test stage that the discrepancy did not represent a serious problem, but that it should be resolved before operational forms were assembled. In Workshop #2, this discrepancy was resolved such that the three levels of cognitive processing were common to both subject areas and both targeted grades. This was achieved through distinguishing between Mathematics items from both the P4 and P6 item pool that measured knowledge versus those that measured understanding and assigning these to the first two cognitive levels separately. All remaining items were categorized under the Application level, although two types of application were identified – reasoning versus non-reasoning. The validity of the distinction between these two types of Application items for Mathematics should be assessed in operational data analysis. The changes to cognitive processing levels for P4 and P6 Mathematics are reflected in Table 6 below while for English language they remained the same and do not appear in Table 6:

Table 6: Cognitive Processing Levels Applied to the NEA Operational Test Forms 2016 for Mathematics

DOMAIN	Knowledge	Understanding	Application
GENERAL			
	Recall or location of a fact, information, or procedure, without necessarily understanding the underlying concepts	Understanding of underlying concepts	Use of knowledge and concepts for the resolution of a problem
MATHEMATICS – P4 and P6			
DOMAIN	**Knowledge**	**Understanding**	**Application**
Numbers and numerals	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 for logical, systematic thinking; includes intuitive and inductive reasoning based on patterns and regularities that can be used to arrive at solutions to non-routine problems (IEA definition, TIMSS 2011 Framework)
Basic operations			
Collect and handle data			
Measurement			
Space and shape			

Table 7 below provides a comparison of desired item weights by cognitive levels which guided both pilot test assembly and operational test assembly. Note that no changes were observed in either the P4 or P6 weightings for English language. However, in Mathematics, after careful review of the items used for operational test assembly, the new weights show a shift to more complex items (Understanding + Application) on the P4 Mathematics operational forms (48% to 71%). On the P6 Mathematics operational forms, the relationship between cognitively less complex items (Knowledge) versus more complex items (Understanding and Application) remained approximately the same. It is important to bear in mind that using cognitive processing weights represented only an estimate of expected behaviour of items to guide the assembly of test forms.

Table 7: Percentages of Items at Each Cognitive Processing Level for Pilot Test Assembly Compared with Operational Test Assembly

	English language					
	Knowledge		Understanding		Application	
	Pilot	Operational	Pilot	Operational	Pilot	Operational
P4	44%	44%	32%	32%	24%	24%
P6	30%	30%	50%	50%	20%	20%
	Mathematics					
	Knowledge/ Understanding	Knowledge	Application	Understanding	Reasoning	Application
	Pilot	Operational	Pilot	Operational	Pilot	Operational
P4	51.4%	29%	34.3%	34%	14.3	37%
P6	42.5%	48%	40.0%	17%	17.5%	35%

Review NEA 2016 test blueprints and test maps

Test blueprints provide a tabular description of the structure of a test highlighting at least the following information: list of content standards intended to be measured on a test

crossed with the range of cognitive demand per standard; number of items per content standard (i.e., the weight that each content standard and cognitive level will receive on the test); item type for each item; and structure of the test from the point of view of broad functional categories of items (e.g., number of anchor items used for horizontal and/or vertical comparisons; number of operational items (non-anchor items); number of items used (for obtaining a test score).

A test map provides a tabular representation of the test listing each item on the test in sequence from first to last, together with key information that defines each item (typically standard measured, item type, cognitive level measured, function of item on test, scoring keys for multiple choice items, and score points per item). These documents are important for the assembly of test forms and facilitate consistent replication of test design from one administration to the next or from one form to another within the same administration. Once administration of test forms has been conducted, test maps typically get updated with the data obtained from administration.

Review of the test blueprints and test maps for the NEA 2016 was carried out ensuring that the features described in the Section on page A-8 of this document (*Review Structure of NEA 2016 Operational Tests*) were integrated into test structure, namely: Horizontal anchoring; embedded field test design; multiple forms per test; and, cognitive processing levels.

Review items cards and print

The final activity of week 1 of this workshop focused on producing item cards for all items that were piloted tested in January 2016 together with items in the item bank that derived from earlier test administrations (mostly 2013 anchor items). Item cards provide the following information, automatically generated from the item database:

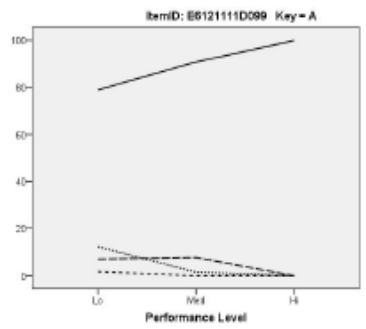
Figure 7: Example of an Item Card for a P6 English Language Item

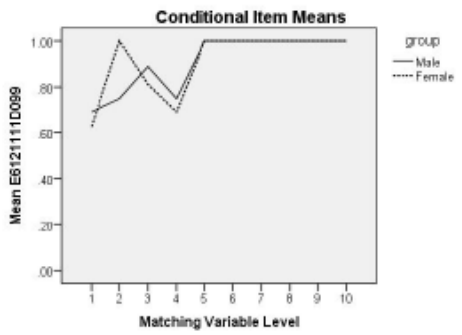
29/03/2016 11:02 a.m.

It is 4 o'clock in the afternoon and the people at Alcate Village are hurrying to the sea-shore to welcome the fishermen. They are returning from fishing, having left at dawn. Sometimes they return with no catch. Today the people in the village are expecting the fishermen to come back with a lot of fish.	Why are the people rushing to the sea shore? A To welcome the fishermen. B To go fishing. C To sell their fish. D To clean their fish.
COMMENTS	

English – EN06		Domain: 1	Year: 2016	Function: O N: 199	Accept O
ItemID: E6121111D099	Type: M	Strand: 2	Forms: 1		Revise O
TextID: E6121111D000	Key: A	SLO: 1	Position: 10		Reject O

Options analysis							Parameters		DIF	
Option	A	B	C	D	M	Omit	Difficulty	Discrimination	N Ma	94
Percent:	91	4	5	1		0	P-Value .91	PB-Corr .34	N Fe	105
Opt PB-Corr:	.34	-.26	-.18	-.14		.00	B-Par -1.13	A-Par .89	ETS Cat	
Flag	L						Flag PH	Flag	FavGroup	





- The complete item
- All item identifying information including: item ID code, text ID code (if relevant), item function on test form providing statistics, N size, item type, key, form and

position on test form, and information regarding content standard measured (domain, strand, and SLO)

- Statistics derived from option-level analyses (percentage selecting option and the option point-biserial correlation, with data on omits
- Statistics derived from item-level classical theory and IRT analyses (P-value, item point-biserial correlation, A- and B-parameter statistics
- All item flags including DIF flags
- Graphic representation of a) option performance indicating percentage of pupils choosing option across low, mid, and high achievement – a high discriminating item will exhibit lower to higher percentage across the 3 performance groups); and b) differential item functioning for groups assessed – in the case of the Ghana NEA pilot test, male versus female item performance

All item cards were printed and assembled in folders by subject area, grade, and domain. They were used to identify which items will be selected for operational forms, as described in the next section of this report.

Selection of Items for NEA 2016 Operational Forms

The second major goal of this workshop was to select items for all operational forms by subject/grade. The task was conducted during the second week of the workshop (April 4 through 8) and led to the production of completed test maps identifying operational, anchor and new items (for integrated pilot-testing) sequenced for each test form.

In order for item selection to be carried out for each of the required operational forms for the NEA 2016, it was necessary to conduct an audit of the item bank to determine the number of validated (i.e., items with performance statistics) items available for 2016 operational selection for each of the standards measured. This information on the current status of the item bank made it possible to determine how many validated items will be available for the subsequent operational test assembly (in 2018), and consequently the number of items that needed to be piloted in 2016 in order to ensure the required number of items in the item bank for 2018 operational assembly. These non-validated items for 2018, which exist but were not piloted, augmented by a small number of specific items developed during this workshop, were inserted into the “new” item slots in the 2016 forms for piloting, as required for the embedded field test design applied to the 2016 operational test design. Information regarding the number of items needed per content standard to be able to assemble 2018 operational forms in turn enabled the development team to determine the number of forms required for 2016 operational testing. The results of the item bank audit and estimates of numbers of items and forms required for 2016 operational test assembly as well as 2018 test assembly are provided in Table 7 below. Note that in order to be able to assemble an appropriate test form, given that items must meet content standards and cognitive processing specifications, up to three times the number of items should be available in the item pool; thus in Table 7 column E the number of items required for 2018 assembly (column C) was multiplied by 3 minus the number of validated items already available (column D) to determine how many items needed to be piloted in 2016, information which enables calculating of the number of forms needed to be assembled for July 2016 operational administration (with a total of 5 new items per form).

Table 8: Number of Items Available in Item Bank and Needed for 2016 and 2018 Operational Test Form Assembly

P4 English Language	A	B	C	D	E
	# items required per form (2016) (operational + anchor)	**# items available per 2016 form**	**# items required per form (2018) (operational only)**	**# items available per 2018 form (B-A)**	**# items needed to pilot in all 2018 forms (C x 3 – D)**
Listening	8	11	6	3	6 x 3 – 3 = **15**
Grammar	11	24	9	12	9 x 3 – 12 = **15**
Reading	16	40	12	24	12 x 3 – 24 = **12**
TOTAL	**35**	**76**	**27**	**39**	**42 (= 9 forms)**
P6 English Language	**# items required per form (2016) (operational + anchor)**	**# items available per 2016 form**	**# items required per form (2018) (operational only)**	**# items available per 2018 form (B-A)**	**# items needed to pilot in all 2018 forms (C x 3 – D)**
Listening	10	21	7	11	7 x 3 – 11 = **10**
Grammar	14	23	11	9	11 x 3 – 9 = **24**
Reading	16	41	12	25	12 x 3 – 25 = **11**
TOTAL	**40**	**85**	**30**	**45**	**45 (= 9 forms)**
P4 Math	**# items required per form (2016) (operational + anchor)**	**# items available per 2016 form**	**# items required per form (2018) (operational only)**	**# items available per 2018 form (B-A)**	**# items needed to pilot in all 2018 forms (C x 3 – D)**
Numbers	5	13	4	8	4 x 3 – 8 = **4**
Basic operations	16	27	13	11	13 x 3 – 11 = **28**
Measurement	6	7	4	1	4 x 3 – 1 = **11**
Shape & space	3	3	2	0	2 x 3 – 0 = **6**
Data & chance	5	16	4	11	4 x 3 – 11 = **1**
TOTAL	**35**	**66**	**27**	**31**	**50 (= 10 forms)**
P6 Math	**# items required per form (2016) (operational + anchor)**	**# items available per 2016 form**	**# items required per form (2018) (operational only)**	**# items available per 2018 form (B-A)**	**# items needed to pilot in all 2018 forms (C x 3 – D)**
Numbers	4	5	3	1	3 x 3 – 1 = **8**
Basic operations	16	31	13	15	13 x 3 – 15 = **24**
Measurement	8	21	6	13	6 x 3 – 13 = **5**
Shape & space	6	13	5	7	5 x 3 – 7 = **8**
Data & chance	6	16	5	10	5 x 3 – 10 = **5**
TOTAL	**40**	**86**	**32**	**46**	**50 (= 10 forms)**

Once the item bank audit was completed, the test developers turned their attention to the selection of test items for each of the operational test forms required for P4 and P6 Math and English Language. Note that what was required for 2016 operational forms was ONE operational form for each subject/grade, and TEN pseudo-forms per operational form (i.e., 10 versions of the same operational form with only variation in the order of some of the operational items making the difference between each pseudo-form).

Item selection, using the item cards, was a sequential process:

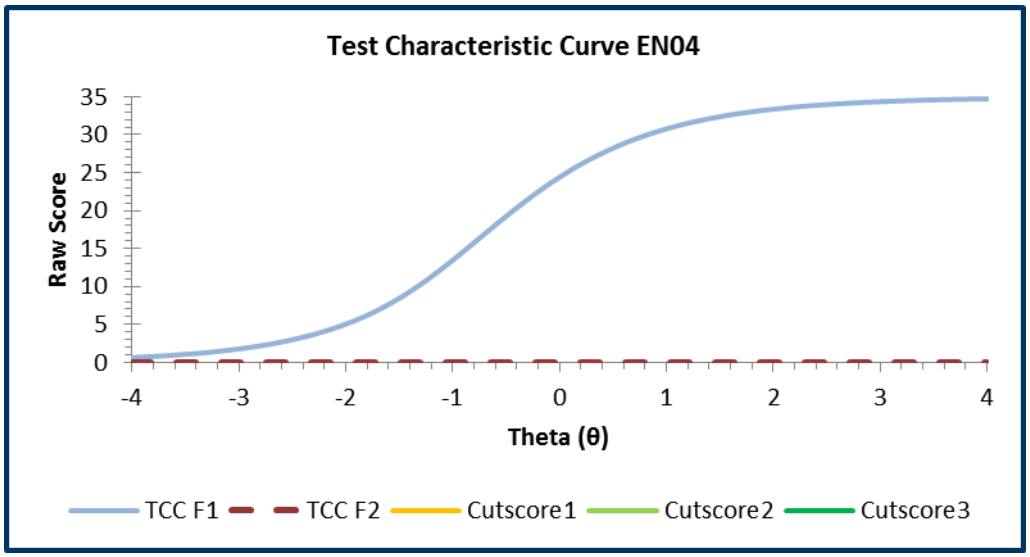
1. Identified all anchor items (from the 2013 or 2011 forms to the extent possible supplemented by new anchor items where necessary) and noted the item IDs in the relevant slots on the test maps; these anchor items occupied the same slots on all 10 2016 forms;
2. Identified all operational items as per test map specifications and noted the item IDs on the test maps; operational items were common across all forms within the same

testing event (i.e., July 2016) although their position varied slightly (by 2-3 positions on the test form) across the 10 forms in order to increase test administration security;

3. Identifed all "new" items for piloting to occupy the "new" slots on the 10 test forms; new items were different across the 10 forms.

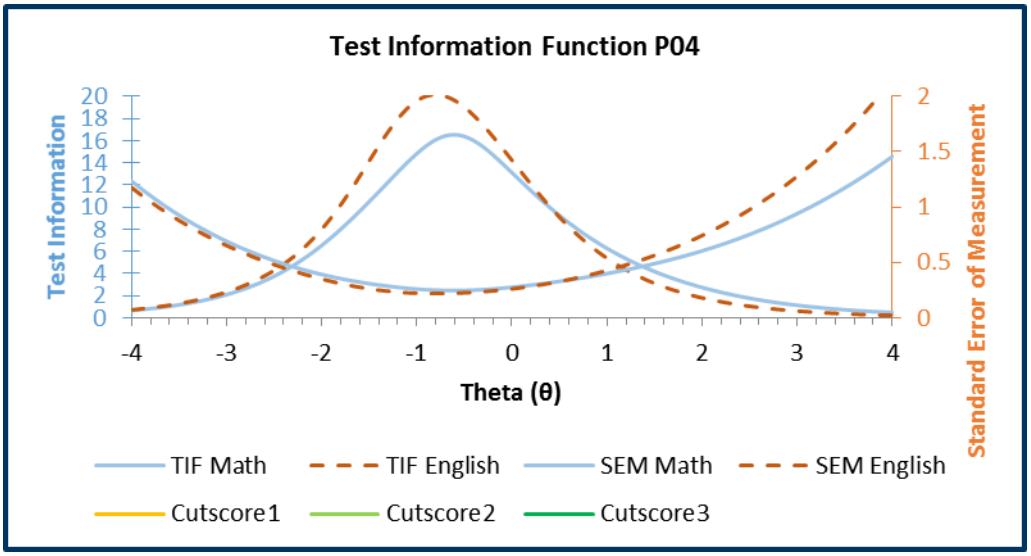
Once the test forms were assembled, the form information (i.e., sequence of item IDs) was entered into the electronic test assembler. The test assembler automatically generated test form point-biserial correlations as well as test characteristic curves (see Figure 8 below). This information can be used to make changes to the form to improve its correlation – this can be achieved by identifying then exchanging an item with weaker item correlation for another with stronger correlation, thus having an effect on the overall test correlation. This process of test form refinement allows for the construction of the optimum quality test form. Note that in Figure 8 the curve for P4 English Language is almost perfectly inclined indicating that no changes needed to be made to the form structure.

Figure 8: Test Characteristic Curve for P4 English Language



Once the team was satisfied with the final form of the operational test, they could compare the item test curves for P4 Language and P4 Math, likewise the 2 P6 forms from both subject areas, to determine if the tests were of equal or similar difficulty levels and correlations. Further adjustments could be made to bring the 2 tests at the same grade level closer to each other if so required. In Figure 9 below is a comparison between the P6 English Language Test Information Function and that of P6 Math. The functions show that P4 English Language had greater discrimination (provided greater information) than P4 Math and was slightly easier than the P4 Math test (was further to the left).

Figure 9: Test Information Function for P4 English Language and P4 Math



Assembly of Electronic Version of Final Test Forms

During the third week of the workshop, final forms of all tests were assembled in Word using the completed test maps to provide information on item IDs and sequencing. A test template, used to assemble pilot test forms, was used for the purposes of assembling operational test forms.

Once in final digital format, the tests were printed and reviewed and all editing and formatting modifications were made.

Annex B: NEA 2016 Sample Methodology

The population of interest for the 2016 NEA was all P4 and P6 pupils who attended Ghanaian primary schools (private and public) during the 2015/2016 school year.

The 2016 NEA sample used the 2013/2014 EMIS census data as the sample frame. After exclusion of schools which contained a P4 or P6 pupil enrolment of less than 10 pupils[10] (*n* = 4,725 schools), 15,754 schools remained in the sample frame. The 15,754 schools were thought to contain approximately 530,956 P4 pupils and 481,644 P6 pupils.[11]

Schools were stratified by region and sorted by district, locality, school type (public or private) and enrolment. For each region, 55 schools were randomly sampled with equal probability. This provided for a total sample size of 550 schools. All P4 and P6 pupils attending the selected schools on the day the NEA was administered (11–13 July 2016) were automatically selected to complete the assessment.

In all, 18,915 P4 pupils and 17,081 P6 pupils were assessed from 546 schools. ***Table B1*** summarizes the number of schools and the number of P4 and P6 pupils who completed the NEA 2016 assessment, by region and class.

Table B1: Final counts of schools, P4 and P6 pupils who completed the NEA 2016, by region

Region	Schools	P4 pupils	P6 pupils	Total
Ashanti	55	1,874	1,728	3,602
Brong Ahafo	55	1,842	1,532	3,374
Central	55	1,798	1,677	3,475
Eastern	54	1,520	1,383	2,903
Greater Accra	53	2,115	1,988	4,103
Northern	55	1,804	1,601	3,405
Upper East	55	2,303	2,204	4,507
Upper West	55	2,189	1,805	3,994
Volta	55	1,779	1,630	3,409
Western	54	1,691	1,533	3,224
Total	**546**	**18,915**	**17,081**	**35,996**

Although at least one school from all 216 districts was randomly sampled, the sample size was insufficient to make appropriate 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.

[10] Exclusions based on enrollment of less than ten P3 or P6 pupils were done in the previous four NEAs (2007, 2009, 2011, 2013).

[11] Estimates are based on the weighted counts of the NEA 2016 pupil-level data (the estimated number of pupils present on the day of assessment).

Sample weights were generated at the school level as the total number of schools divided by the sampled number of schools in each region.