Ghana National Education Assessment

2011 Findings Report

January 24, 2012

Ministry of Education
Ghana Education Service
Assessment Services Unit

Ghana National Education Assessment

2011 Findings Report

January 24, 2012

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 do not necessarily reflect the views of USAID or the United States Government.

Prepared under EdData II Technical and Managerial Assistance, Task Number 12, Contract Number AID-641-BC-11-00001, Strategic Objective 3. Submitted to Luis Tolley, USAID/Ghana; and Sandra Bertoli, Office of Education, Bureau for Economic Growth, Agriculture and Trade (EGAT/ED), USAID/Washington.

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

Table of Contents

Page

List of Figures

List of Tables

Page

Abbreviations

ASU	Assessment Services Unit
CONFEMEN	*Conférence des Ministres de l'Éducation des pays ayant le français en partage*
CRDD	Curriculum Research and Development Division
DHS	Demographic and Health Survey
EdData II	Education Data for Decision Making (project)
EFTS	Education Fund Tracking Survey
EGRA	Early Grade Reading Assessment
EMIS	education management information system
FAWE	Forum for African Women Educationalists
GAR	gross admission rate
GDP	gross domestic product
GES	Ghana Education Service
GETFund	Ghana Education Trust Fund
GHc	Ghana cedi (currency)
GSFP	Ghana School Feeding Program
MAF	Materials Allocation Form
MC35	minimum competency NEA test score of 35% or better
NALAP	National Literacy Acceleration Program
NEA	National Education Assessment
OR	odds ratio
P3	primary school, grade 3
P6	primary school, grade 6
PASEC	*Programme d'Analyse des Systèmes éducatifs des États et gouvernements membres de la CONFEMEN*
PCE	per child expenditures
PF55	proficiency NEA test score of 55% or better
PIRLS	Progress in International Reading Literacy Study
QC	quality control
RTI	RTI International (trade name of Research Triangle Institute)
SD	standard deviation
SMC	school management committee
SPIP	school performance improvement plan
SRIMPR	Statistics, Research, Information Management, and Public Relations (unit)
TIMSS	Trends in Mathematics and Science Study
UPE	universal primary enrollment
USAID	United States Agency for International Development

Acknowledgments

This work was funded by the United States Agency for International Development (USAID) through the Education Data for Decision Making (EdData II) project, Task Order 12, contract number AID-641-BC-11-00001. The authors are grateful to Luis Tolley and USAID/Ghana's Education Team for their guidance and support throughout this process. Thank you also to the Curriculum Research and Development Division's (CRDD's) Director, Sarah Agyeman-Duah; and Assessment Services Unit (ASU) staff members Isaac Asiegbor, Antwi Aning, Anthony Sarpong, and Kwabena Nyamekye for their partnership on this activity and for their very successful application of the 2011 National Education Assessment (NEA). Thanks to all the members of the RTI International Ghana team (Emmanuel Sam-Bossman, Peter Gyekye, and Janet Amaning), whose participation ensured a timely and smooth application of the NEA. We are also grateful for the Statistics, Research, Information Management, and Public Relations/Education Management Information System (SRIMPR/EMIS) department, and more specifically, Hebert Gorman, for providing essential EMIS data needed for the NEA sample selection and school characteristics analysis included in this report. We also acknowledge and extend our thanks to the test monitors and administrators who administered the test, and to all the Primary 3 and Primary 6 students who participated in this assessment. Finally, Emily Kochetkova and Erin Newton of RTI International assisted with revising and editing the report.

Executive Summary

NEA Report Background

This report is an analysis of findings from the 2011 administration of the Ghana National Education Assessment (NEA), carried out by the Assessment Services Unit (ASU) within the Ghana Education Service's (GES's) Curriculum Research and Development Division (CRDD).

The NEA is a biannual nationally and regionally representative measure of student competency in mathematics and English in primary grades 3 and 6 (P3 and P6). The 2011 NEA was the fourth application of the NEA and it covered all 10 regions of Ghana, sampling 580 schools and testing as many as 60,000 students over the course of three days. The NEA tests are based on national curricula and are made up of 40 to 60 multiple-choice questions, administrated collectively to students. Test items are distributed in the following domains:

English	**Mathematics**
Listening	Basic Operations
Reading Comprehension	Numbers and Numerals
Writing	Measurement
Usage (Grammatical Structure)	Shape and Space
	Collecting and Handling Data

Test Results

Students are defined as having achieved "minimum competency" if they answer at least 35% of the NEA test items correctly. Students who answer at least 55% of the items correctly are defined as having achieved "proficiency." The 2011 NEA summary results are presented in ***Table ES1.***

Table ES1 Percentages of students achieving minimum competency and proficiency levels, 2011 NEA

Test score cut point	Percent estimate[a] (95% confidence interval)[b] P3 maths	P3 English	Percent estimate[a] (95% confidence interval)[b] P6 maths	P6 English
Minimum competency (score ≥35%)	**52.6** (50.3, 54.9)	**66.3** (64.0, 68.5)	**56.9** (54.3, 59.5)	**78.9** (76.8, 80.8)
Proficiency (score ≥55%)	**18.2** (15.9, 20.8)	**24.2** (21.3, 27.4)	**16.1** (13.3, 19.3)	**35.3** (31.8, 38.9)

[a] Weighted estimates.
[b] 95% confidence intervals were estimated taking the final weights and cluster design into consideration.

English proficiency levels were greater than those for maths, with 24.2% of P3 students and 35.3 of P6 students achieving proficiency-level scores. However, ideally, proficiency levels would be higher than those observed.

Although the validity of comparing scores across years is limited—given that analyses confirming comparability of test difficulty by year are not possible—trend analysis still provides a sense of change in student performance over time. As can be seen in ***Table ES2,*** the percentage of P3 students achieving minimum competency and proficiency scores in English increased between 2009 and 2011. Similarly, the percentage of P6 students achieving minimum competency in English saw a slight increase, while those achieving proficiency remained stable. Conversely, the percentage of P3 students achieving minimum competency and proficiency in math decreased. Among P6 students, the percentage achieving minimum competency also decreased while, surprisingly, those achieving proficiency increased slightly.

Table ES2 Overall distribution of students reaching minimum competency and proficiency levels, all years of NEA administration (%)

Year	Primary 3				Primary 6			
	English		Mathematics		English		Mathematics	
	Minimum compe-tency	Profi-ciency	Minimum compe-tency	Profi-ciency	Minimum compe-tency	Proficiency	Minimum compe-tency	Proficiency
2005	50.6	16.4	47.2	18.6	63.9	23.6	47.2	9.8
2007	50.2	15.0	42.6	14.6	69.7	26.1	46.2	10.8
2009	57.6	20.0	61.2	25.2	76.9	35.6	61.9	13.8
2011	66.3	24.2	52.6	18.2	78.9	35.3	56.9	16.1

The mean scores for the 2011 administration were low (***Table ES3***), although English scores were decidedly better than math scores for both grades.

Table ES3 Mean scores by grade and subject

Grade	Mathematics	English
P3	37.8%	42.9%
P6	38.7%	48.6%

Examining student performance on the English test's domains (***Table ES4***) reveals that student performance on the listening comprehension questions was relatively strong for both P3 and P6, indicating that students' oral comprehension skills and understanding of English were good. However, when students' ability to read and write English was tested directly (Grammar, Reading, and Writing domains), the scores were considerably lower, with mean domain scores ranging from 33.9% to 44.1%. As for P3 student performance

on the six reading comprehension questions[1] within the Reading domain, 42% of the students had either 0 or 1 correct answer out of these six items, indicating that these students were having great difficulty with reading.

Table ES4 Average English test scores, by grade and domain

Domain	P3	P6
Listening	63.6%	72.8%
Grammar	35.1%	44.1%
Reading	40.6%	44.1%
Writing	33.9%	32.4%

Maths Results by Domain

Mean maths scores by domain indicate that the majority of students had not mastered the skills targeted in the assessment. The percentage of P3 students unable to answer a single domain question for the Numbers (7.8%), Collecting and Handling Data (16.1%), and Shapes and Space (17.8%) domains was substantial. Performance appeared strongest, although still weak, in the Operations domain ***(Table ES5).*** P3 performance appeared weakest in the Collecting and Handling Data domain, whereas P6 performance was strongest in this area. The difference in the percentage of students achieving proficiency-level scores among P3 and P6 students in Collecting and Handling Data is striking (9.3% versus 42%). However, given the small number of items in this domain, generalizing from this finding is inadvisable. Apart from this domain, the percentage of students achieving minimum competency and proficiency scores is similar.

Table ES5 Average maths test scores, by grade and domain

Domain	P3	P6
Numbers	39.3%	40.7%
Operations	42.9%	42.4%
Measurement	38.8%	Not applicable
Shapes and Space	41.2%	37.3%
Collecting and Handling Data	29.0%	44.0%

Results by Gender

In Ghana, while progress has been made in gender parity as it concerns access to school, differences remain between girls' and boys' learning outcomes in maths. Logistics regression that takes into consideration region, school location, and school type indicates that P3 girls were 16% less likely (odds ratio [OR] = 0.84) to achieve minimum

[1] Questions 31 through 36.

competency scores than boys in maths. Similarly, logistics regression indicates that P6 girls were 25% less likely (OR=.75) to achieve minimum competency in maths and 36% less likely (OR=.64) to achieve proficiency in maths than were boys. With regard to English, the difference in scores was not statistically significant.

Results by School Type

Private schools outperformed public schools and National Literacy Acceleration Program (NALAP)[2] pilot schools significantly in both grades and subjects. Although the share of students achieving minimum competency and proficiency scores at NALAP pilot schools does not, at first glance, appear to have been better than at public schools, when student gender, region, and school location are taken into consideration, P6 students at NALAP pilot schools ***were*** more likely to have achieved minimum competency and proficiency on both the maths and English tests ***(Table ES6).*** P6 NALAP pilot school students were 43% (OR=1.43) more likely to achieve minimum competency in maths; 78% (OR=1.78) more likely to achieve minimum competency in English; and 55% (OR=1.55) more likely to achieve proficiency in English. All of these differences were significant, with the exception of P6 students achieving proficiency in maths. The difference between NALAP pilot and other public schools was less pronounced among P3 students. P3 students in the NALAP pilot schools fared only slightly better than the public schools in terms of reaching the minimum-competency level in maths (OR=1.18), and significantly (statistically speaking) better in English in reaching minimum competency (OR=1.38).

Table ES6 Weighted odds ratios (adjusted by core demographic variables) of students achieving minimum competency and proficiency levels, by subject (excluding private schools)

Variable		Maths: Odds of achieving (95% confidence interval)[a]		English: Odds of achieving (95% confidence interval)[a]	
		Minimum competency	Proficiency	Minimum competency	Proficiency
P3	Public[b]	1	1	1	1
	NALAP pilot school	1.18 (0.95, 1.47)	1.03 (0.75, 1.42)	1.38 (1.04, 1.84)*	1.03 (0.68, 1.54)
P6	Public[b]	1	1	1	1
	NALAP pilot school	1.43 (1.11, 1.85)**	1.39 (0.98, 1.98)	1.78 (1.37, 2.33)***	1.55 (1.15, 2.08)**

[a]Weighted odds ratio estimates and 95% confidence intervals were calculated using Taylor linear series to account for the sample design's cluster effect.
[b]Reference class.
*$p < .05$. **$p < .01$. ***$p < .001$.
Note: Significance was calculated based on Pearson's Chi-squared test.

[2] This program—recently upgraded from pilot to national scale—teaches students to read in their mother tongue before being introduced to English in the third grade. It also publishes and distributes textbooks and leads teacher training, both on a large scale.

Results by School Location

Ghana's very significant investment in education has allowed the country to make tremendous progress toward achieving universal primary enrollment (UPE); 2010-2011 data from the Education Management Information System (EMIS) indicate a primary gross enrollment rate of 98.4%, a primary gross admission rate of 99.6%, and a net enrollment rate of 77.8%. However, despite increases in funding as well as efforts to improve equity in resource allocations, substantial differences remain, with urban areas and certain regions enjoying greater resource allocations (including a greater share of trained teachers, greater textbook availability, and greater per child funding levels). Although resources alone do not determine student performance, there are correlations.

Differences between rural and urban schools are considerable and significant for both grades and subjects ***(Table ES7)***. Student performance in urban schools was much higher than in rural schools. More than twice as many P3 students from urban schools achieved proficiency-level scores on both the maths test (22% vs. 9%) and English test (29% vs. 11%). The difference in performance between urban and rural P6 students is even more striking. Urban P6 students were more than three times more likely to achieve proficiency math scores than their rural counterparts (21% versus 6%).

Table ES7 Percentage of students achieving minimum competency and proficiency levels in maths and English, by grade and school location

Grade	Type of school	Maths: Percent achieved[a]		English: Percent achieved[a]	
		Minimum competency	Proficiency	Minimum competency	Proficiency
P3	Rural	42%	9%	56%	11%
	Urban	74%	29%	74%	29%
P6	Rural	44%	6%	69%	17%
	Urban	65%	21%	88%	46%

[a]Weighted estimates.
Note: Confidence intervals and odds ratios are available in an annex to the main report.

Similarly, regional differences are quite strong ***(Table ES8).*** P6 and P3 students in Greater Accra outperformed their peers in other regions and in both subject areas by a wide margin. For example, the ratio of Greater Accra students achieving proficiency compared to students in the lowest-performing regions ranged from 4.9 on the P3 English test to 9.5 on the P6 maths test. In other words, in comparing students from Greater Accra to those from the lowest-performing regions, between nearly 5 and 10 times more students from Greater Accra were able to achieve proficiency-level scores than students from lower-performing regions. Greater Accra also outperforms the other regions when

only public schools are taken into consideration[3]. In addition, the logistic regression model indicates that these findings cannot merely be explained by the fact that a higher proportion of Greater Accra schools are urban.

Table ES8 Percentage of students achieving minimum competency and proficiency levels in maths and English, by grade and region

	MC	Prof.	MC	Prof.	MC	Prof.	MC	Prof.
	P3				P6			
	Mathematics		English		Mathematics		English	
Greater Accra	**74%**	**40%**	**86%**	**54%**	**81%**	**38%**	**97%**	**74%**
Ashanti	55%	22%	70%	30%	62%	20%	85%	43%
Western	53%	17%	69%	22%	59%	16%	80%	32%
Eastern	52%	16%	66%	20%	59%	14%	77%	34%
Central	51%	16%	67%	21%	51%	11%	78%	28%
Brong Ahafo	51%	13%	66%	18%	52%	10%	72%	26%
Volta	44%	10%	53%	13%	51%	12%	76%	27%
Upper East	46%	13%	53%	13%	50%	9%	69%	20%
Northern	41%	11%	62%	19%	36%	4%	67%	17%
Upper West	41%	9%	46%	11%	48%	8%	70%	16%

It is worth noting that nearly all P6 students achieved minimum competency in English in the Greater Accra Region (97%). Even more important is the fact that nearly three quarters of students in the Greater Accra region achieved proficiency-level scores in English. Understanding what it is that allows the Greater Accra students to perform well will provide insights into what is needed to help the other regions improve their performance.

Factors Associated with Student Performance

In an effort to identify factors that may be hindering or helping performance, the researchers linked student results to school characteristic data available via the EMIS. Although being able to link individual student demographic data and individual teacher characteristics and classroom practices would have provided even more information, this analysis does provide some useful insights. In addition to the factors mentioned above, school factors found to be significantly associated with stronger student performance included:

[3] the differences between Greater Accra (the reference region) and the others region are all significant except Eastern and Western region in Maths grade 6 (minimum and proficiency level) as well as for Upper east (proficiency level only) and for Western Grade3 maths proficiency and Grade6 minimum level.

- availability of textbooks
- proportion of female teachers
- proportion of teachers with training
- visits from circuit supervisors
- schools keeping administrative registers
- student transfer rates

School factors negatively associated with student performance included:

- higher repetition and dropout rates
- multigrade classrooms
- higher percentage of orphans at the school

Recommendations Based on NEA Findings

Resource Allocation

Ghana's significant investment in primary education should be maintained, while per child expenditures should be made more equitable across regions and districts. Nevertheless, important geographical disparities remain. Rural remote schools have significantly lower results than others, especially in the Northern Region and in poorer learning conditions (large classes—i.e., with 70 or more students—and untrained teachers). A significant proportion of schools have no textbooks and most schools lack sufficient textbooks.

Potential measures for reducing these disparities include:

- **using capitation grants to reduce large disparities in the availability of school equipment and to secure minimum allocations of textbooks in each school;**
- **reviewing the funding allocation formula or developing school-level allocation formulas for use at the district level;**
- **publishing school funding and resource allocations so that parents can help monitor allocations (following a similar initiative in Uganda);**
- **initiating an Education Fund Tracking Survey.**

Schools that must maintain multigrade classrooms should receive **specific pedagogical supervision and in-service training** to help teachers overcome the challenges associated with teaching in multigrade-classroom settings.

Finally, as the proportion of students transferring to or from other schools was found to be associated with higher learning outcomes, this may mean that parents are seeking better education for their children. This social demand for education, a necessary

condition for achieving universal enrollment, should be supported by efforts toward more equitable distribution of funds, teachers, and materials, with a specific focus on deprived districts. Although teachers receive financial support to teach in certain zones, they should also be trained to teach to specific student populations that have little contact with the English language at home or in their local environment, as within NALAP.

Teacher Training

Results indicate that teachers are not teaching several subjects effectively.

NEA data collection should include classroom observations to evaluate teaching approaches. Data gathered through these observations could then be fed into the design of in-service and pre-service teacher training materials. Additionally, specific teacher training could be focused on teaching the subjects/domains in which students received the lowest scores.

As a related point, given that students taught by trained teachers scored significantly better than those taught by untrained teachers, especially if the trained teachers were female, **teacher-training programs should be continued regardless of the deployment issues discussed earlier.**

Also, although teachers' academic qualifications were not associated with better learning outcomes, **measures should be taken to enforce a strict selection process to ensure that the applicants are sufficiently qualified for courses designed to prepare candidates for teaching.**

In order to reduce the gender gap on maths tests, **teaching and learning materials and practices should be revised to be more girl-friendly, and teacher training should include consideration of gender issues during teaching of mathematics.**

As the data show that the proportion of students not reaching minimum reading competencies does not decrease between P3 and P6, with regard to a policy of automatic or collective grade promotion, **teachers' practices should be reformed such that all students achieve a minimum literacy level in the early grades.**

Greater emphasis should be placed on teaching foundational English and mathematics skills and ensuring that all students have mastered these skills. Mastery of these skills will, in turn, help to ensure that students are prepared to successfully acquire more advanced skills taught at higher grade levels.

School Management

Data show that basic school records such as teacher attendance reports are not kept systematically. While the existence of school management committees (SMCs) and school performance improvement programs (SPIPs) were not found to lead to better learning outcomes per se, the data are silent regarding the attributes of the SMC and the content and targets of the SPIP. However, given that the existence of an SMC or SPIP was found to contribute to better administrative record-keeping, it is recommended that:

- **SMCs and SPIPs should be expanded, with more focus on learning outcomes, teacher and student attendance, and effective time on task.**
- **SPIPs should set pedagogical objectives at the school level with regard to basic skills such as reading and counting.**
- **Teacher and student attendance should be monitored more frequently at the school level by greater community involvement and by visits from the circuit supervisor.**

Circuit supervisors can be an avenue for improving learning outcomes but seem to have little effect on students reaching minimum performance levels. Therefore it is recommended that circuit supervisor interventions should focus more on reforming teacher practices designed to reach the lower-performing students, and circuit supervisor interventions should target remote schools.

The issue of **teacher absenteeism should be addressed**, as this affects student learning time. This challenge could be approached by involving the teachers unions in discussions about ways to resolve teacher absenteeism.

Finally, this report provides several recommendations regarding specific ways to improve the design and future applications of the NEA test itself.

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

This report is an analysis of findings from the 2011 administration of the Ghana National Education Assessment (NEA), which was carried out by the Assessment Services Unit (ASU) within the Ghana Education Service (GES). For this effort, the United States Agency for International Development (USAID) sponsored technical assistance to the ASU through the Education Data for Decision Making Project (EdData II), led by RTI International.

This was the fourth round of the biennial NEA. The NEA is a nationally and regionally representative measure of student competency in mathematics and English in grades 3 and 6 (known as Primary 3 [P3] and Primary 6 [P6]). The 2011 NEA (conducted in July 2011) covered all 10 regions of Ghana, sampling 580 schools and testing as many as 60,000 students in the course of three days. As part of the analysis prepared for this report, the results of the test were stratified by gender, by region, and by type of school (public vs. private), with the findings presented below. The assessment also compares student performance between public schools and National Literacy Acceleration Program (NALAP) pilot schools.[4]

The report is organized as follows. The remainder of Section 1 presents some historical background from the Ghana education sector that is relevant to the administration and analysis of the NEA. Section 2 describes the methodology for the 2011 implementation of the NEA, from construction of the test forms through data entry. Section 3 presents the results of the research team's analysis of the NEA data; Section 4 follows up with a discussion of demographic, environmental, and school factors that may be affecting learning outcomes. Section 5 discusses the research team's overall recommendations, and Section 6 concludes with some limitations and challenges in the 2011 instruments that should be reviewed and addressed before the next round of implementation. Finally, several annexes offer more detailed technical discussions of material presented in the main text.

1.1 Education Spending

Ghana has made great strides in supporting its education sector and increasing access to education. Ghana spends more of its budget on education than the majority of African countries. Economic growth, specific funding mechanisms such as the Ghana Education

[4] As of 2010, all public primary schools in Ghana implement the National Literacy Acceleration Program, an approach to literacy, first introduced in 2004, which teaches students to read in their mother tongue before being introduced to English in the third grade. Students transition over to English entirely in the fourth grade. In addition to this mother-tongue instruction approach, NALAP includes the publication and distribution of millions of textbooks and the training of 80,000 teachers. When the results of the pilot were shown to be successful, the GES decided to scale up the program to the national level.

Trust Fund (GETFund), and donor support has resulted in 11% of gross domestic product (GDP) geared toward education annually (or 30% of government spending). This figure rose from 6% in 2003 to 11% in 2008.

The share of primary education budget (as a percentage of total current expenditure on education) is 43%. The unit cost as a share of GDP per capita is 19% in primary versus 11% average in sub-Saharan Africa.[5] Funds have been budgeted specifically to foster conditions favorable to universal primary education. For example, in 2003, school fees were eliminated, resulting in greater enrollment levels. Capitation grants, first introduced in districts classified as being deprived, and then universally to all districts, were created to compensate for the elimination of school fees.[6] A number of additional initiatives have been established or expanded, including the Ghana School Feeding Program (GSFP) started in 2005 and generalized to all districts in 2008 (resulting in increased enrollment levels).

1.2 Access, Enrollment, and Retention Rates

The impact of this investment is evidenced by the increases in student enrollment levels. Ghana has made tremendous progress toward achieving universal primary enrollment (UPE); 2010-2011 data from the Education Management Information System (EMIS) indicate a primary gross enrollment rate of 98.4%, a primary gross admission rate (GAR) of 99.6%, and a net enrollment rate of 77.8%. In 2009, the completion rate was 83%, while the net enrollment rate was 76%. With the support of the private sector, which enrolls 23.9% of students, Ghana is well on track to achieving UPE.[7] Enrollment for girls has also benefited from these new policies; the gender parity index is 0.96, one of the highest in sub-Saharan Africa.[8] The introduction of the automatic promotion policy has resulted in a very low average repetition rate (around 4%). Similarly, primary school completion rates have increased steadily, reaching 91.6% according to the 2010-2011 EMIS.

Despite these improvements, however, dropping out remains a challenge and efforts are needed to curb its occurrence, especially in the early grades. In addition, student and teacher attendance are both problematic, resulting in reduced time on task. A 2004 World Bank survey found a teacher absenteeism rate of 22.3%. More recent data from NALAP classroom observations show that in 25% of the classrooms, activities did not begin on time and in another 25%, time was wasted.[9]

---

[5] World Bank EdStats database, 2011, http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTEDUCATION/EXTDATASTATISTICS/EXTEDSTATS/0,,menuPK:3232818~pagePK:64168427~piPK:64168435~theSitePK:3232764,00.html

[6] Capitation grants are sent to schools to cover special school events and purchases of sporting equipment such as footballs. The capitation grant funding is quite small and amounts to just 4.5 GHc per student per year.

[7] EMIS 2010-2011.

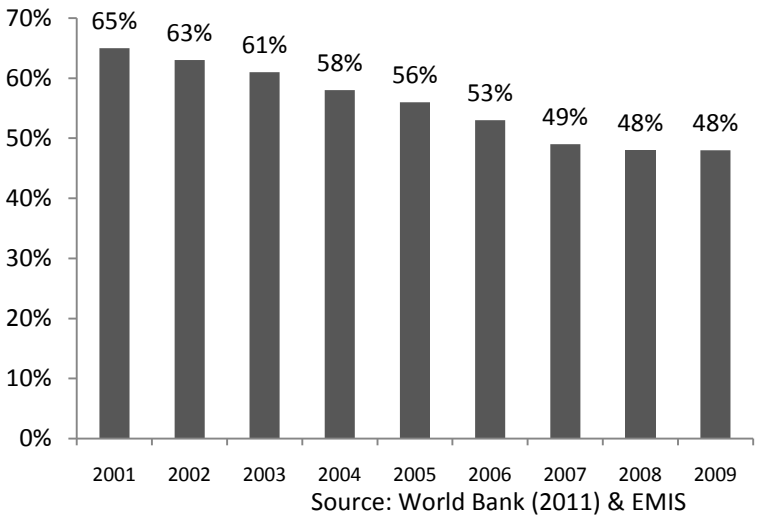
[8] World Bank. (2011). *Education in Ghana: Improving equity, efficiency, and accountability of education service delivery*.

[9] USAID. (2011). *NALAP formative evaluation.* PowerPoint presentation, March 2011.

1.3 Teachers

In recent years, the student/teacher ratio has been maintained at approximately 32 students per teacher, with a goal of 35. This is well below the African average. However, the share of *trained* teachers has decreased in recent years. Enrollment in teacher-training colleges only increased from 23,999 in 2003 to 26,100 in 2008, which was insufficient given the increases in student enrollment. In an effort to address this issue, a program of in-service training, which delivers diplomas through distance education, was put in place in 2005, enrolling 27,183 teachers in 2008. By 2008, the share of trained teachers seems to have stabilized at 48% (see ***Figure 1.1***), although this percentage is still well below the 65% rate that was experienced in 2001.[10] (Section 1.5, on equity issues, breaks down the proportion of trained teachers by region.)

Figure 1.1 Percentage of trained primary education teachers in Ghana's overall teacher corps, by year, 2001–2009

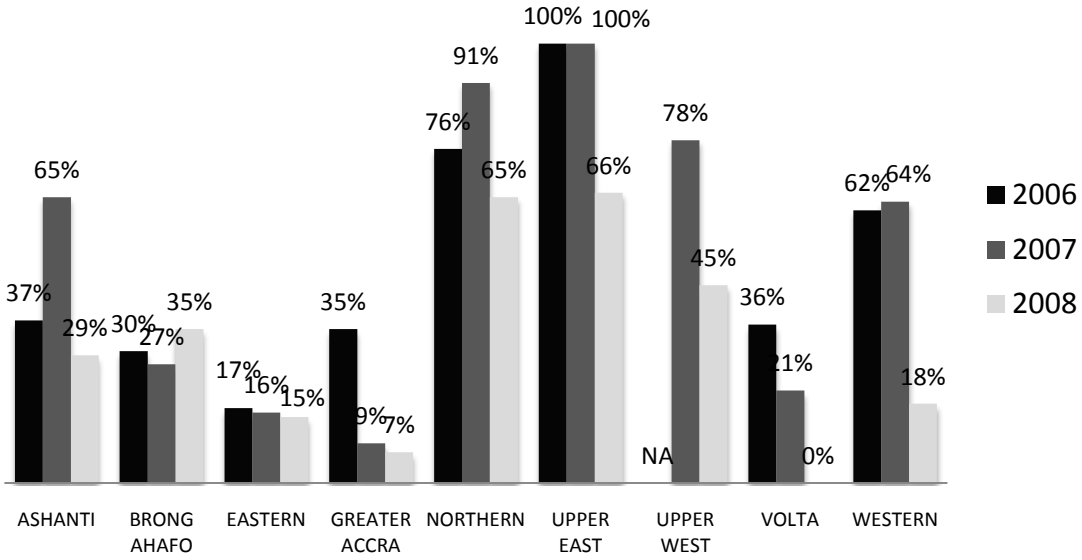


1.4 Equity Issues

Despite significant education spending on average and pro-poor policies (such as the identification of deprived districts to be used in budget allocation calculations), significant geographical disparities remain. Substandard per child expenditure (PCE) is defined as funding that falls within the bottom third of the per capita expenditures in the country. Although ***Figure 1.2*** shows a reduction in the funding disparity across the regions between 2006 and 2008, considerable disparities still existed. As of 2008, the percentage of children in districts with substandard PCE was about 66% in Northern and Upper East and 45% in Upper West, while in Greater Accra, only 7% of children were affected by substandard PCE.

[10] World Bank EdStats database, 2011.

Figure 1.2 Funding disparities: Percentages of children in districts with substandard education expenditures, by region and year



Source: World Bank (2011), p. 132

Only recently has the PCE in deprived districts increased in the primary and junior high school cycles, but even now (2011), expenditures remain lower than in non-deprived districts.

As could be expected, the disparities in funding are reflected in the regional distribution of trained teachers. Most trained and experienced teachers tend to teach in Accra, despite incentives to teach in deprived districts. The percentages of trained teachers range from 60.2% in the Eastern Region to just 32.9% in the Northern Region.[11] Similarly, the distribution of pedagogic materials is uneven. While the average number of textbooks per child has increased in recent years,[12] this ratio declined more in deprived districts than in non-deprived ones.[13]

[11] World Bank (2011).
[12] World Bank (2011).
[13] In addition to regional differences, students' socioeconomic status is clearly linked to issues such as school access, attendance, retention, and—possibly—learning outcomes. For example, a Demographic and Health Survey (DHS) found that while 10.2% of children from the top economic quintile were not in school, 37.5% of children from the poorest quintile were not in school. Similarly, a World Bank study (World Bank, 2011, p. ii) found that a rural girl from the poorest quintile is 13.9 times more likely not to have attended school than an urban boy from the richest quintile.

As long as the strong relationship between student achievement and independent factors such as teacher training, location, and other school indicators[14] exists, disparities in inputs will yield inequities in student performance.

1.5 Background of the NEA and EMIS Data Collection

As described earlier, the National Education Assessment involves tests administered every two years on a sample of P3 and P6 students, in maths and English. Tests are based on national curricula and are made up of 40 to 60 multiple-choice questions. The questions are presented in four forms, ordered differently on each. New questions are developed and added each year, but the tests retain a few common items over time. Following Bloom taxonomy, the tests target two levels of cognitive abilities (knowing and understanding, applying); they do not measure higher-order skills (analysis, synthesis, and evaluation/creation).

Less than a third of primary school children reached proficiency levels in English or in mathematics, according to the NEA results of 2005, 2007, and 2009. In an effort to better understand these results, this report evaluates student performance at the subdomain level[15] and also links the test results to available EMIS data to evaluate factors that could be associated with stronger or weaker student performance.[16]

EMIS variables considered in this analysis include student/teacher ratios, teacher training, availability of textbooks, keeping of attendance registers, and frequency of visits by the circuit supervisor.[17]

2. Methodology

2.1 Content of the NEA

The NEA is designed to evaluate students' facility with P3 and P6 maths and English skills. Further, the 2011 forms contained the following numbers of items:

- P3 English – 10 listening comprehension plus 30 read/write
- P3 mathematics – 40
- P6 English – 15 listening comprehension plus 45 read/write

[14] USAID (2009). *Basic education in Ghana: Progress and problems.*

[15] The NEA content is based on topics and subtopics—known as domains and subdomains—from within the Ghana national curriculum. Sample domains from the P3 mathematics curriculum include are "Numbers," "Operations," and "Shapes and Space."

[16] However, no questionnaires were administered to pupils, teachers, or directors; instead, the analysts relied on administrative data. Moreover, no measurements were available for some factors, such as pupils' socioeconomic status.

[17] J. Gillies (2008). *Opportunity to learn: A high impact strategy for improving educational outcomes in developing countries.* Report prepared for USAID/EQUIP2. More information on the tests can be found in Chapter 2, while Chapter 5 provides a list of policy issues addressed by EMIS data.

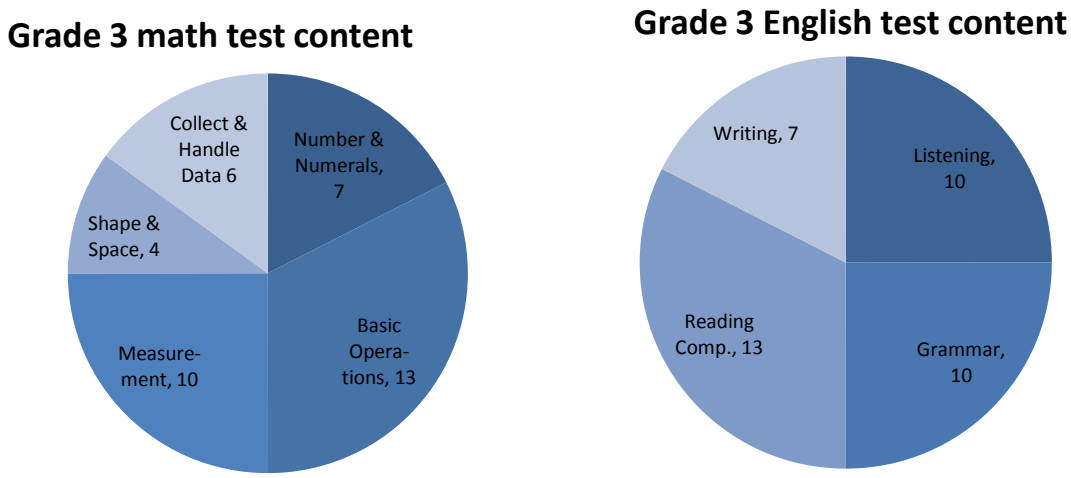
- P6 mathematics – 60

As noted earlier, there were four forms (versions) of the NEA for each subject in both P3 and P6. Equal numbers of the forms were distributed in the selected schools. Using multiple forms of the test reduces the probability that students will copy each others' work.

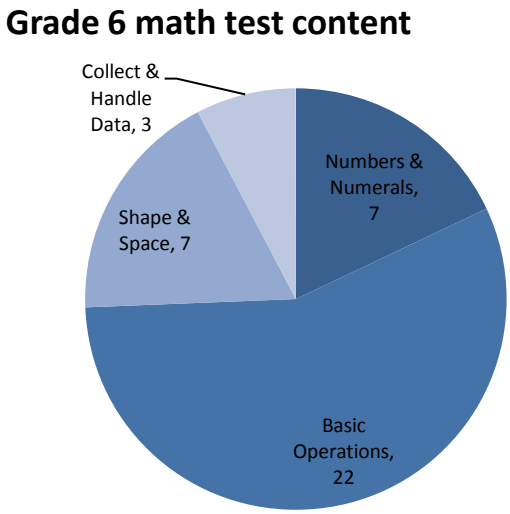
Item analysis (Annex G) showed great test reliability and internal consistency. In terms of difficulty, English items were well aligned with students' abilities, whereas the maths items seem to have been rather difficult. As is standard international practice, the test item difficulty progressed within the test so that the easier questions appeared first.

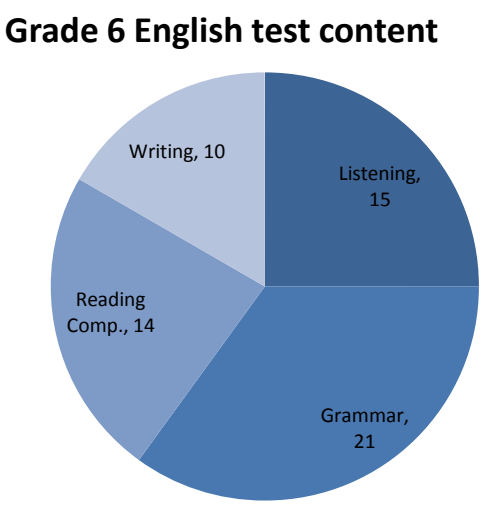
Each subject area covered multiple sub-topics or domains included in the Ghanaian curriculum for P3 and P6. As may be seen in the four graphs in ***Figure 1.3,*** the number of test items designed to measure each of these domain skills was uneven within each set of forms. This was particularly true for the P6 mathematics test forms. Some of the domains—for example, "Basic Operations" in the P6 mathematics test—included an unnecessarily large number of items. Conversely, some other domains—such as "Shapes and Space"—had too few items.[18]

Figure 2.1 Test content: Proportions of specific domain skills in overall examination



[18] Note that the Ghanaian domain definitions differ slightly from those found in international surveys, such as the Progress in International Reading Literacy Study (PIRLS) or the Trends in Mathematics and Science Study (TIMSS). In addition, some domains covered in most international surveys are not explicitly tested in the NEA, such as problem solving and written expression (producing messages).





2.2 Sample Design

This section presents technical information about the samples of schools and students that are represented in the analysis.

2.2.1 Population

The population of interest for the 2011 NEA was all Ghanaian primary schools that educated at least ten P3 and ten P6 students during the 2009-2010 academic year.

The population frame, or list of eligible primary schools, was drawn from the 2009-2010 Ghana EMIS data. This frame contained 14,137 eligible primary schools.

2.2.2 Sample Methodology

A sample of 580 schools was drawn from the population frame using a stratified systematic sampling methodology. Schools were stratified by the ten regions and a separate stratum was created for NALAP pilot schools, for a total of 11 strata. Before proceeding with systematic selection, the research team sorted each stratum by district, school type, and the EMIS 2009-2010 enrollment figures for P6 students.

To maximize the statistical power at the region level, an equal sample of 55 schools was systematically selected at random for each region. For the NALAP pilot school stratum, 30 schools were systematically selected (from the Northern and Volta regions). For each selected school, all P3 and P6 students present on the day of testing were automatically chosen to participate, unless there were more than 200 students per grade. In this case, 200 students were to be randomly selected. Only one school was found to have more than 200 students in P6, however, and none of the sampled schools had more than 200 students in P3. In this one instance where there were more than 200 P6 students, field

staff had enough forms available and did not want to exclude any children, so all 232 P6 students were given the exam.

Appropriate weights were calculated to account for equal sampling of schools by region and for schools with more than 200 students in P3 or P6. Further weighting adjustments were calculated and applied to the test score data for nonresponses, such as student absenteeism on the day of test, refusal to participate, or incomplete test forms. For more information on the weighting, please refer to ***Annex A.***

2.2.3 Enrollment Updates and Replacement Schools

Each sampled school had two replacement schools that were to be used as a reserve only if the original schools were deemed ineligible or did not consent to participating in the study. The field staff was given only the original list of sampled schools; if a school had to be replaced, the main office in Accra informed the field staff of the name and location of the first replacement school. This ensured that the main office and not the field staff would make final decisions regarding school replacements.

Once a school was sampled, the two schools on the list next to the sampled school were automatically selected as its two replacement schools. Because the list of schools was sorted by district, school type, and P6 enrollment, prior to selection, schools with similar characteristics were listed next to each other. Automatically selecting the two schools next to the originally selected school ensured the replacement schools were similar to the sampled school at the district, school type, and enrollment levels.

Although 2009-2010 EMIS data were used to draw the sample, EMIS enrollment data from the *following* year—2010-2011—were used to update the expected number of students, to verify sampled schools were open for 2010-2011 academic year, and to verify that enrollment figures for P3 and P6 were greater than ten students. Before the ASU sent the field staff out to administer the NEA to the sampled schools, it contacted the schools to obtain consent to participate. A total of 11 schools were replaced with reserve schools.[19]

2.3 Sample Description: Expected vs. Obtained Schools and Test Forms

Of the 580 sampled schools, 573 (97.8% completion rate after replacement) schools successfully completed the 2011 NEA exam. Schools were defined as "completed" if completed testing materials from all four test subjects (P3 maths, P3 English, P6 maths, and P6 English) were successfully returned to Accra headquarters, and marked answer sheets could be successfully scanned.[20]

[19] Five schools were replaced because the current P3 and/or P6 enrollments were fewer than 10 students. One school was replaced because it was a special school for the hearing impaired. Staff were not able to contact three schools prior to collecting data. While staff were in the field, two schools were replaced because they were deemed not accessible during the rainy periods of data collection.

[20] Of the seven sampled schools deemed not complete, the reasons for non-completion included: answer sheets from three schools were not returned to headquarters, answer sheets from two schools were returned but were blank, and

The left side of ***Table 2.1*** shows the breakdown of schools sampled and the expected number of P3 and P6 students (based on EMIS data). The right side of Table 2.1 shows the number of completed schools as well as the actual number of completed P3 and P6 tests.

There were discrepancies between the total number of completed maths and English tests within each grade, as seen in Table 2.1. For instance, in the Ashanti Region, 2214 students in P3 completed the maths test but only 2151 completed the P3 English test (a difference of 63 completed tests). These differences are explained primarily because of improper or incomplete answer sheet markings. An answer sheet was defined as "complete" if the following requirements were met: grade, subject, and test form number were properly marked, and at least five questions were attempted (regardless of whether the answers were correct or not). The far majority of the incomplete answer sheets were missing the test form number. Because of this omission, it was not possible to link the answer sheets to the proper answer key, and thus the scorers were not able to evaluate them.

On average, 40 P3 students per school completed the maths and English tests (standard deviation [SD] = 25.6) and an average of 35 P6 students per school completed the maths and English tests (SD = 25.3).

Table 2.1 Expected school and student sample sizes plus numbers of "completed" schools and tests, 2011 NEA, by region

Region	Expected sample sizes: *n* (% private school[a])			Completed schools[b] and tests[c]: *n* (% private school[a])				
	Sampled schools	Expected P3 students	Expected P6 students	Completed schools	Completed P3 maths tests	Completed P3 English tests	Completed P6 maths tests	Completed P6 English tests
Ashanti	55 (25.5)	2402 (23.6)	2188 (21.0)	55 (25.5)	2214 (26.4)	2151 (26.9)	1945 (23.7)	1928 (23.9)
Brong Ahafo	55 (12.7)	2549 (17.4)	2151 (17.7)	55 (12.7)	2191 (18.0)	2108 (18.3)	1974 (19.9)	1954 (20.1)
Central	55 (21.8)	2391 (16.2)	2196 (19.1)	55 (21.8)	2094 (26.7)	1948 (15.6)	1804 (19.7)	1764 (19.7)
Eastern	55 (18.2)	2104 (17.6)	1855 (22.3)	55 (18.2)	1872 (21.7)	1789 (21.7)	1592 (25.2)	1565 (25.3)
Greater Accra	55 (47.3)	2718 (34.6)	2675 (31.5)	54 (48.1)	2242 (36.3)	2176 (37.7)	2240 (32.2)	2208 (32.3)
Northern	55 (1.5)	3252 (1.2)	2722 (1.3)	53 (1.8)	1884 (.48)	1772 (0.5)	1678 (0.8)	1664 (0.8)
Northern (NALAP pilot schools)	12 (N/A)	790 (N/A)	762 (N/A)	11 (N/A)	633 (N/A)	591 (N/A)	675 (N/A)	665 (N/A)
Upper East	55 (3.6)	3351 (1.1)	2603 (1.6)	55 (3.6)	2732 (1.9)	2569 (1.9)	2394 (1.7)	2385 (1.7)
Upper West	55 (1.8)	2889 (3.4)	2153 (4.5)	55 (1.8)	2405 (5.3)	2232 (5.6)	2126 (4.4)	2108 (4.5)

two schools did not have testing materials at all because they were replacement schools and the test materials were mistakenly sent to the schools they were to replace.

Region	Expected sample sizes: *n* (% private school[a])			Completed schools[b] and tests[c]: *n* (% private school[a])				
	Sampled schools	Expected P3 students	Expected P6 students	Completed schools	Completed P3 maths tests	Completed P3 English tests	Completed P6 maths tests	Completed P6 English tests
Volta	55 (12.0)	3052 (10.5)	2686 (14.1)	52 (13.7)	1730 (16.6)	1665 (16.8)	1538 (17.4)	1567 (17.1)
Volta (NALAP pilot schools)	18 (N/A)	768 (N/A)	716 (N/A)	18 (N/A)	693 (N/A)	689 (N/A)	1740 (N/A)	595 (N/A)
Western	55 (16.7)	2444 (13.0)	1982 (15.8)	55 (16.7)	2090 (13.7)	2000 (13.8)	605 (17.5)	1714 (17.6)
Total	**580 (15.5)**	**28,710 (12.3)**	**24,689 (13.7)**	**573 (15.7)**	**22,780 (14.5)**	**21,690 (14.6)**	**20,311 (15.0)**	**20,117 (15.0)**

[a]Unweighted percentages of private schools.
[b]Schools were defined as "completed" if completed testing materials from all four test subjects (P3 maths, P3 English, P6 maths, and P6 English) were successfully returned to Accra headquarters, and test forms were able to be successfully scanned.
[c]An answer sheet was defined as "complete" if the grade, subject, and test form number were properly marked, and at least five questions were attempted (regardless if the answers were correct).

2.4 Training and Preparation of Test Materials

2.4.1 Training the Trainers

RTI worked with ASU to develop and implement a train-the-trainers event, which was held in May over three days in Accra. The purpose of this activity was to prepare ASU and Curriculum Research and Development Division (CRDD) staff to conduct the training of test administrators, who would then administer the 2011 NEA. To this end, RTI revised the existing trainer's guide, test administrator's manual, test monitor's manual and form, and practice mocks, as well as introducing a certification exam, all of which trainers would use to train the administrators and monitors.

The trainer's guide was designed to teach trainers how to present a standardized training program to trainees. The test monitor's manual was designed to be used as a primary reference for monitoring activities. The accompanying test monitor form was designed to collect, systematically, the observations of test administration procedures in a standard manner, including such factors as testing time, preparation, test procedures, materials, and student behavior.

During the training of trainers, the participants became thoroughly familiar with the NEA and learned about the fundamentals of delivering an effective training program. As part of the learning experience, they were given an opportunity to practice their delivery of test instructions to the students during peer group practice sessions. These sessions also provided an opportunity for the trainers-in-training to receive feedback about the proper delivery of instructions and reading of questions. They also practiced filling out the training monitor form during the practice sessions.

2.4.2 Training the Test Administrators and Monitors

Test administrator and monitor training was held at five regional centers. At each training center, test administrators attended 2½ days of training, while test monitors attended 1½ days of training. Overall, approximately 570 test administrators and 200 test monitors attended training at these centers. Test administrators and monitors were Circuit Supervisors or School Inspectors and Assistant Directors of Supervision (Head of Circuit Supervisors), respectively, in a school district.

Trainers used PowerPoint presentations during the sessions, and, as described above, trainees were given either an administrator or a monitor manual and forms, participated in practice mocks, and finished with a certification exam. The assignment of schools to test administrators occurred on the final day of the training. Each data collection team was made up of at least two test administrators and one test monitor. The precise composition of data collection teams depended largely on the number and size of schools to be administered in a district. (Trainee numbers by home region and gender are provided in ***Annex B.***)

2.4.3 Preparing and Packaging Test Materials

To ensure that all the sample schools received the appropriate test materials on time, RTI and ASU staff worked together to assemble, organize, and pack materials. The goals of the packing activity were (1) to ensure that schools received adequate test materials in a timely manner, (2) to ensure that test materials were organized in a way that was convenient for the administrators to administer the test and secure test materials while at sampled schools, and (3) to ensure that test materials were repacked such that they could be easily retrieved and stored after they arrived at the ASU office at the end of the test period. To achieve these goals, the packing of test materials was organized around several factors.

First, all test materials were packed into security bags. The test materials from only one school could be packed into a single security bag; in other words, no two schools were permitted to share a single security bag. The goal was to eliminate any misallocation of test materials between schools.

A second consideration was the number of students enrolled in each sampled school. RTI developed a spreadsheet to calculate material requirements based on enrollment data contained in Ghana's EMIS. If the number of students in either grade exceeded 69, then the recommendation was to pack test materials into two separate security bags by grade. For schools that had higher enrollments, the goal was to avoid overloading the security bags with materials to ease the test administrators' burden of transporting and distributing the materials.

Third, an overage amount was factored into the number of all test materials provided to schools. Overage was calculated at 20% for all materials. The goal was to provide more than enough materials in case there was a need for extras.

Fourth, most test materials were identified using preprinted labels and/or envelopes that contained the region, district, school name, and EMIS code. Preprinting (as opposed to hand-writing) all the labels reduced the amount of time required for labeling, helped to reduce errors in labeling, and ultimately helped to ensure that the right materials reached the right schools.

To expedite packing in future rounds, counting and assembly of test materials could be completed well in advance of actual packing of the bags. However, RTI recognizes that the possibility of doing so depends on both early purchase and delivery of test materials and the availability of workspace.

At the start of the packing session, ASU had seven temporary workers on board to assist with counting, assembly and packing. As packing progressed, it was determined that more staff were needed to assist with the work. Therefore, three more workers were hired. In future rounds, RTI would recommend hiring additional staff before packing begins. In addition, more space to count and assemble materials might have increased the speed with which packing was completed.

The security bags for each school were laid out on the assembly hall floor by region and district. Staff used Materials Allocation Forms (MAFs) to count and assemble the needed test materials, which were then placed on top of the security bags. As a quality control measure, the test materials were recounted before they were packed inside the bags. Then the bags were locked using padlocks and the keys were placed inside an envelope addressed to the receiving district. This envelope, which also contained a copy of the MAF, was handed to the truck driver who would deliver it to district personnel. Security bags were then loaded into trucks to be transported to the districts.

2.5 Data Collection

2.5.1 Test Administration

The 2011 NEA test administration was conducted July 11–13, 2011. As noted above, in every participating school, the two test subjects—English and mathematics—were administered to students in Primary 3 and Primary 6. Each test lasted for 1 hour and 30 minutes, with a 30-minute break in between.

The mathematics test was administered first for both P3 and P6. During this test, each student was supplied with a pencil, an eraser, a sharpener, a question booklet, a pink answer sheet (for shading correct answers), and rough paper to work out mathematical problems. In each class, a test administrator read only the instructions and the example test items to students, and the students were allowed to ask questions about the instructions.

After the maths test, students were tested in English listening comprehension first. Subsequently, they were tested using a written English test. The materials used during the English test included pencils, erasers, sharpeners, question booklets, a listening-

comprehension answer booklet (to identify multiple-answer options from which to choose a correct answer) and green answer sheets (for shading correct answers). Similar to the maths testing process, test administrators read students the instructions and the example test items, plus the English listening comprehension items. As students worked, test administrators and monitors walked around the room periodically and quietly offered different forms of assistance to students. Before the end of each test, test administrators alerted students of the remaining time.

As part of the test administration exercise, completed answer sheets were packed and secured with padlocks (using a specific packing protocol) separately from the question booklets and returned immediately to the District Education Offices. The secured bags were then transported to each of the 10 regional education offices. By July 20, 2011, national monitors had delivered to the ASU office completed answer sheets in test security bags, under lock. At ASU, the answer sheets were removed from the bags, inventoried, and stored by region and subject on wooden shelves ahead of the editing and scanning process.

2.5.2 Editing, Scanning, and Electronic Data Sets

Editing, scanning and e-mailing of scanned data from the answer sheets received from the schools began on August 22, 2011. The editing process began a day after CSX South Africa had serviced the five scanners, programmed them, and trial-tested them with completed answer sheets. Seven non-GES workers served as the editing staff. They primarily ensured that all the fields on the answer sheet were correctly filled in and all unwanted marks that might interfere and interrupt the scanning process were erased completely. They also addressed cases in which students had shaded two or more answers for the same item. The editing process was carefully documented.

While the editors worked, three ASU staff members and one RTI staff person scanned the edited answer sheets to convert them into electronic data. Scanning was done per school and compiled by region. On August 24, the first of the scanned data sets was sent to RTI's home office. By the end of September 2011, data from all 10 regions had been scanned and e-mailed to RTI.

2.6 Data Processing

2.6.1 NEA Test Achievement Data

Once at RTI, data were checked for accuracy and completion, and accidentally double-scanned test forms were omitted. Sample weights[21] were adjusted at the school level (to account for the seven incomplete schools) and applied to the test achievement data set. Weights were further adjusted to account for incomplete (nonresponse) test forms at the student level and were then scaled to the student population by class using the 2010-2011 EMIS enrollment figures. Completed tests were corrected based on subject, class, and

[21] For sampling information, see Section 2.3. For more information on weighting, see Annex A.

test-form number. Note that because the maths and English forms were not linked to the students, it was not possible to merge the maths and English results at the student level; thus, test achievement data are at the class and subject levels.

2.6.2 EMIS Data

EMIS data not only were used as the sample frame to randomly select schools, but also were linked to the test achievement data (by EMIS school IDs) to produce statistical inferences about student learning achievement. The EMIS provided three different data sets—teacher information, school information, and enrollment figures—for the current school year (2010-2011) and past academic year (2009-2010). Some data sets did not contain all the sampled schools. In these cases, data from 2009-2010 were used as proxies. If schools were found in neither the 2010 nor the 2011 EMIS data set, then the researchers omitted those schools when analyzing student achievement with variables taken from the EMIS data sets.

Please see ***Annex C*** for a description of the EMIS data used during the NEA analysis.

2.6.3 Application of Analysis Weights

As was noted in the sampling section, in order to have representative data for each of the ten regions in Ghana, the research team had to sample a sufficiently large and equal number of schools from each region. This is why 55 schools were sampled from each region. However, given that the different regions are not all the same size with equal population, school, and student numbers, final results must be weighted to provide unbiased national estimates. In other words, regions with greater numbers of schools and students should receive a larger weight in the analysis than those with smaller numbers of schools and students. For example, near equal numbers of students from Ashanti (2214 students) and Brong Ahafo (2191 students) completed the P3 maths exam (see Table 2.1 above). However, as Ashanti has almost twice as many schools as Brong Ahafo, Ashanti received a greater weight in the analysis. Looking at ***Table 2.2,*** which provides the weighted percentages of exams by urban/rural, gender region and, school type, one can see that completed P3 maths exams from Ashanti account 19.2% of the weighted total number of P3 maths exams while Brong Ahafo made up 10.7% of the weighted proportion. In other words, Ashanti representation was nearly double that of Brong Ahafo because there were nearly twice as many schools, P3 students, and P6 students in Ashanti than in Brong Ahafo. Upper Western Region represents the smallest proportion of P3 maths (3.6%) because it had the fewest primary schools and students.

Table 2.2 Descriptive variables: Weighted proportion estimates of core demographic, by subject and grade

Variable		Percent estimate[a] by subject and grade			
		P3 maths	P3 English	P6 maths	P6 English
Urban/rural status	Urban	42.1	56.0	44.3	68.6
	Rural	58.8	74	65.4	87.6
Gender	Female	49.1	49.7	48.7	49.1

Variable		Percent estimate[a] by subject and grade			
		P3 maths	P3 English	P6 maths	P6 English
	Male	50.9	50.3	51.3	50.9
Region	Ashanti	19.2	19.2	19.3	19.2
	Brong Ahafo	10.7	10.7	10.5	10.5
	Central	10.4	10.3	10.4	10.4
	Eastern	10.5	10.5	10.6	10.5
	Greater Accra	11.0	11.0	12.0	11.9
	Northern	9.7	9.7	9.3	9.4
	Upper East	5.5	5.5	5.3	5.4
	Upper West	3.6	3.5	3.4	3.4
	Volta	8.2	8.3	8.2	8.4
	Western	11.3	11.3	11.1	11.0
School type	Public	82.6	82.5	82.2	82.2
	Private	17.0	17.2	17.4	17.4
	NALAP pilot	0.3	0.3	0.4	0.4

[a]Weighted estimates.
[b]95% confidence intervals may be found in Annex A, Table A1.

Slightly more males than females completed the examinations; the difference was about 2%. Public schools made up about 83% of the total number of completed exams, with private schools making up the remaining 17%, and NALAP pilot schools making up almost 0.5% of the completed tests.[22]

3. NEA Results

3.1 Achievement Cut Points: Minimum Competency and Proficiency

As indicated in Section 1.5, all four subject tests (P3 maths, P3 English, P6 maths, and P6 English) used the same test score cut points to indicate a student achieved the *minimum-competency level* (MC35) and the *proficiency level* (PF55). That is, students who scored 35% or better are defined as having reached minimum competency and students scoring 55% or better are defined as having reached proficiency. Note that this is the definition of minimum competency and proficiency that has been applied in all previous NEA applications. Thirty-five percent originally was set as the minimum competency as it would indicate that students had achieved a score higher than what they could have achieved if they had guessed all the questions. Analysis of the probability of achieving the minimum competency and proficiency scores is provided in ***Annex E.*** Note, however, that international standards generally classify students as proficient if they have achieved a minimum score of 70%. For comparability reasons, this report presents NEA's traditional classifications of proficiency (≥55%), but readers should be aware that this

[22] The four test forms for each class/subject were included as demographic characteristics because each exam asked different questions that could have an impact on pupils' final scores. This means that each test form should make up 25% of the total subject examinations. The test forms were well balanced across schools and regions. P3 math showed the largest imbalance of completed tests, but the imbalance still was within 2.5 percentage points (P3 math Form 1 = 26.3%, P3 math Form 4 = 23.8%).

definition does not effectively identify students who truly are proficient in the subject area.[23]

3.2 Scores and Achievements

Percentages of students achieving the minimum competency and proficiency thresholds are provided in ***Table 3.1.*** P3 and P6 students clearly performed better on the English than on the maths tests, as larger numbers of students were able to achieve competency and proficiency in the English tests. Slightly over half of the P3 students achieved minimum competency in maths (52.6%) while 66.3% achieved minimum competency in English in 2011 (see Table 3.1). A much smaller fraction of students achieved proficiency, but again, the number achieving proficiency in English was greater than in maths. Close to one-fourth of the P3 students achieved proficiency in English (24.2%) while almost one-fifth of the P3 students achieved proficiency in maths (18.2%). A similar but more striking pattern is seen among P6 students: Twice as many students achieved proficiency in English than in maths (35.3% versus 16.1%). Again, however, although the proficiency and minimum competency definitions used by the GES are applied for reasons of comparability, students who are able to correctly answer only slightly more than half of the questions correctly have not mastered the subject matter.

Proficiency levels increased between P3 and P6, indicating that a larger share of the older students had become comfortable with English. The converse is true for maths, where a slightly smaller proportion of the P6 than P3 students achieved proficiency scores.

Table 3.1 Percentages of students achieving minimum competency and proficiency levels

Test score cut point	Percent estimate[a] (95% confidence interval)[b]		Percent estimate[a] (95% confidence interval)[b]	
	P3 maths	P3 English	P6 maths	P6 English
Minimum competency (score ≥35%)	**52.6** (50.3, 54.9)	**66.3** (64.0, 68.5)	**56.9** (54.3, 59.5)	**78.9** (76.8, 80.8)
Proficiency (score ≥55%)	**18.2** (15.9, 20.8)	**24.2** (21.3, 27.4)	**16.1** (13.3, 19.3)	**35.3** (31.8, 38.9)

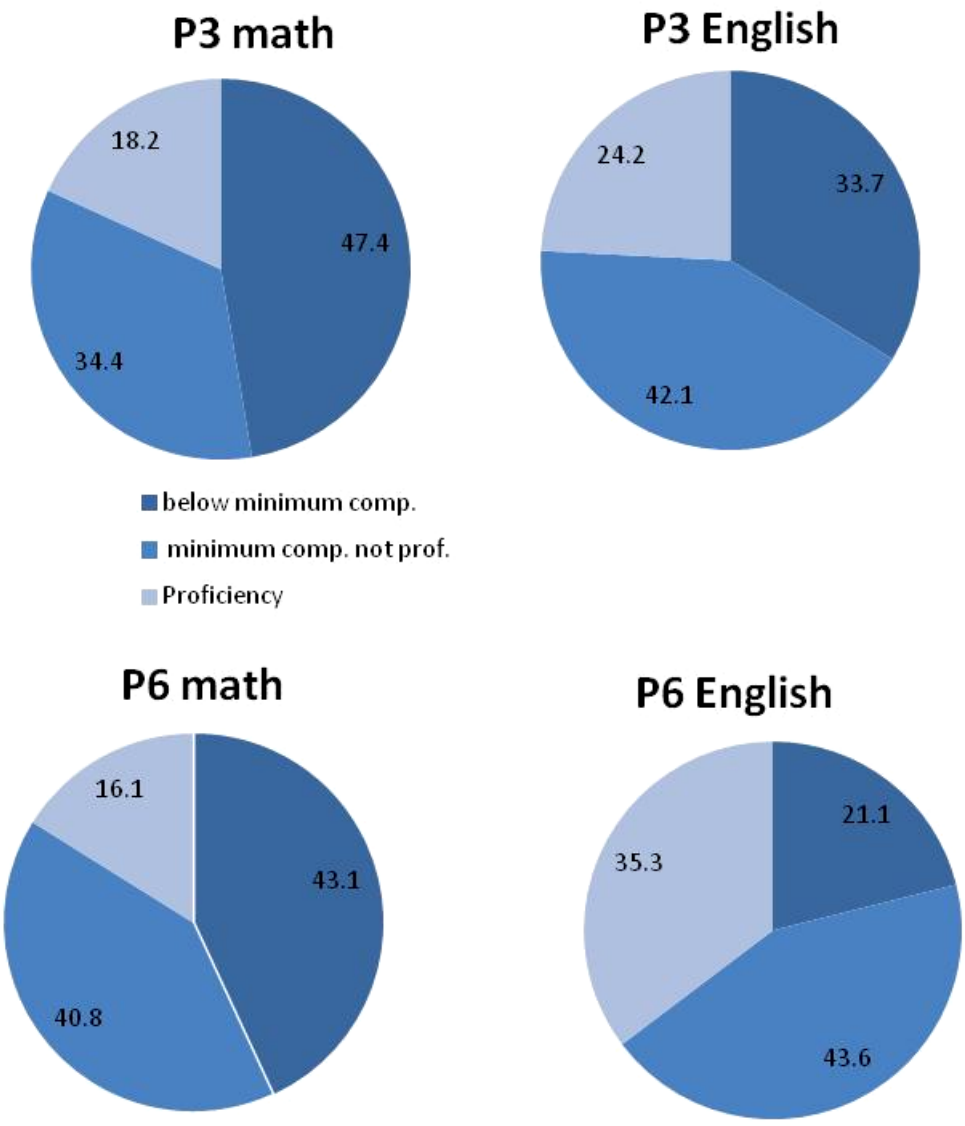
[a] Weighted estimates.
[b] 95% confidence intervals were estimated taking the final weights and cluster design into consideration.

As important as it is to look at the percentage of students achieving the minimum competency and proficiency scores, it is equally important to look at the proportion of students not achieving the minimum scores. The graphs in ***Figure 3.1*** show the proportions of students who achieved proficiency, who achieved only competency, and

[23] As a comparison, the international PASEC assessment, which has a similar number of test items, sets the minimum competency threshold at 40% and the proficiency threshold at 70%. (PASEC is the *Programme d'Analyse des Systèmes éducatifs des États et gouvernements membres de la CONFEMEN [Conférence des Ministres de l'Éducation des pays ayant le français en partage]*, an assessment implemented by a consortium of francophone countries.)

who did not achieve proficiency or competency. As can be easily seen, close to half of the students in P3 and P6 did not reach a score of at least 35%. Difficulty understanding the English instructions may have affected these scores, particularly at the P3 level, as English was still relatively new to these students. On a more positive note, the number of P6 students unable to achieve even minimum competency was lower in P6 than in P3. This, again, was particularly true in English.

Figure 3.1 Percentages of students not reaching minimum levels, reaching minimum but not proficiency levels, and reaching proficiency levels



3.3 Time Series Analysis

As described in Section 1.5, the NEA was administered every other year between 2005 and 2011. From a psychometric context, there are several data requirements for comparing test-takers' performance on examinations such as this one over time. However, some of these critical pieces of information, such as question-by-question responses (item-level data) and copies of test booklets (item content information), are missing from most of the historical NEA data. Without these pieces, only a superficial and qualitative glance at trend data is possible.[24]

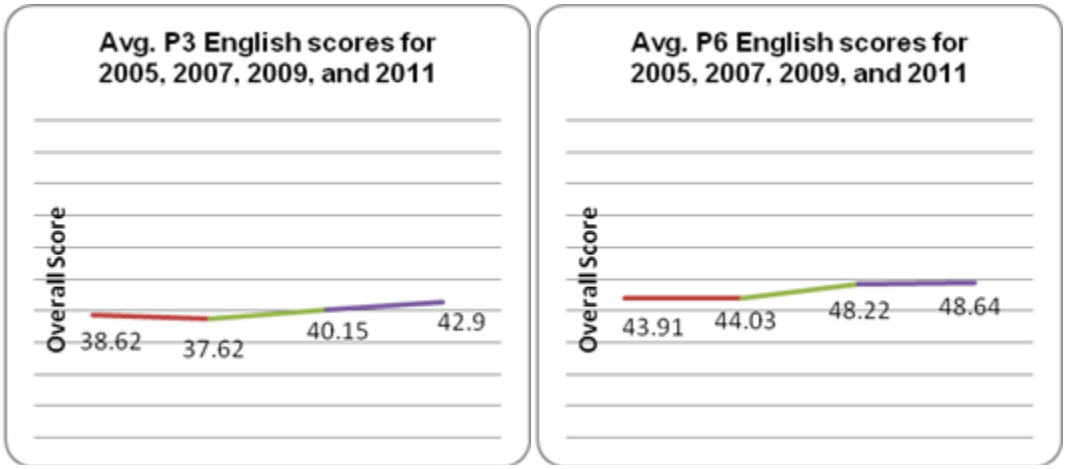
Table 3.2 illustrates which portions of the necessary information are present for each year of the NEA data collection. If the ASU can maintain the item-level data and the item content information for every administration going forward from 2011, then they will be able to make valid comparisons back to the 2011 testing period.

Table 3.2 Status of required data for each year of NEA administration

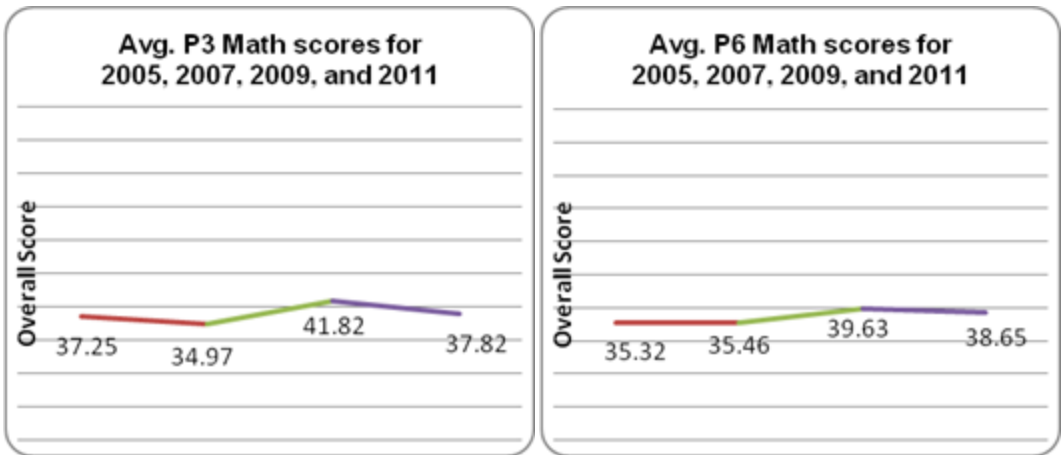
Requirement	Year of data collection			
	2005	2007	2009	2011
Item-level data	NO	**YES**	NO	**YES**
Item content	NO	NO	**YES**	**YES**

The average achievement data presented in ***Figure 3.2*** provide year-specific information.

Figure 3.2 Average English and maths scores for Primary 3 and Primary 6 in 2005, 2007, 2009, and 2011



[24] An explanation of why this information is needed to evaluate trend data is presented in ***Annex F.***



Although valid direct year-to-year quantitative comparisons are not possible, the graphs by year seem to indicate, at the very least, stability in overall assessment. This stability is likely due to similar questions over time, although it is unknown whether the stability truly means that students' ability is being measured accurately. It is possible that students are being consistently tested using unreliable materials. Only item-level assessments at each year can answer the question of whether ability in English and mathematics is being reliably assessed. (See ***Annex G*** for an item-level evaluation of the 2011 instruments.)

In looking at individual subject information, differences appear between English and mathematics. Working under the assumption that the 2009-2011 English tests were qualitatively more or less equivalent, with the only major difference being that the reading comprehension questions at the P6 level were harder in 2011, there are broad trend-like generalizations that can be noted. Based on a qualitative evaluation of the 2009 and 2011 tests, in the both the Primary 3 and Primary 6 tests, there appear to have been improvements in student performance in the 2011 English test.[25] As noted previously, item analysis would really be required to definitely measure whether this increase in the English test is statistically significant; however, based on the qualitative review of the test items, this does indicate a positive trend in performance. Conversely, student performance on the mathematics tests showed a slight drop between 2009 and 2011.

3.4 Results by Subject Domain

As mentioned in Section 2.1, the questions in the maths and English tests covered multiple domains. Examining student performance by domain or sub-topic can provide additional insight regarding students' strengths and areas that may need additional focus. The following sections present the average scores for students by grade and domain.

[25] Moreover, confidence intervals are to be calculated for each year.

3.4.1 English Results by Domain

On the English test, student performance on the listening comprehension questions was relatively strong for both P3 and P6 ***(Table 3.3)***, indicating that students' oral comprehension skills and understanding of English were good. However, when students' ability to read and write English was tested directly (grammar, reading, writing domains), the scores were considerably lower, with mean domain scores ranging from 33.9% to 44.1%. Taking as a hypothesis that students have mastered a domain subject when they have achieved a score of at least 70%, with the exception of listening comprehension, students had not mastered the subject matter tested.

Delving a little deeper to examine P3 student performance on the six reading comprehension questions[26] within the reading domain, 42% of the students had either 0 or 1 correct answer out of these six items. As these six questions required a very basic skill—retrieving explicit information from a simple text—it is fair to say that 42% of the students were having great difficulty with reading. This conclusion is supported by Early Grade Reading Assessment (EGRA) data from many countries, where students who were not able to respond to comprehension questions tended to have low reading fluency scores and at times were not even able to read a single word. As the NEA does not evaluate foundational reading skills, it is not possible to identify where the difficulty in students' reading originates. However, one could safely assume that additional teacher focus on foundational English reading skills would be beneficial to many students.[27]

Table 3.3 Average English test scores, by grade and domain

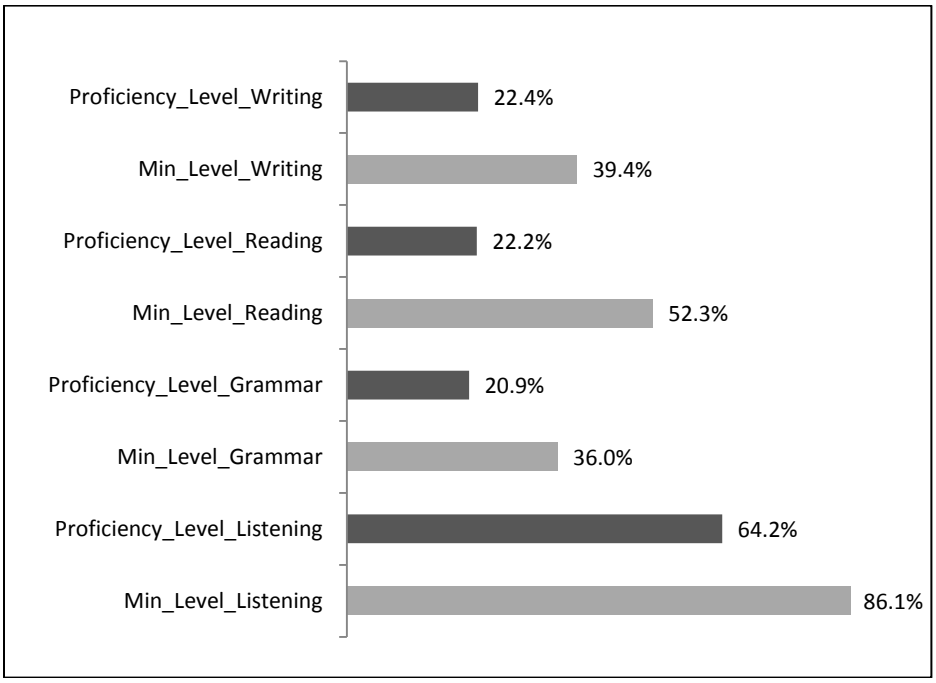
Domain	P3	P6
Listening	63.6%	72.8%
Grammar	35.1%	44.1%
Reading	40.6%	44.1%
Writing	33.9%	32.4%

Note: Standard errors and confidence intervals may be found in Annex D, Tables D7 and D9.

The percentages of students achieving minimum competency and proficiency levels show an understandably similar pattern. For example, 86.1% and 92.6% of P3 and P6 students achieved the minimum-competency level in listening comprehension (see ***Figures 3.3 and 3.4***). A full 78.1% of P6 students achieved proficiency in this domain. Far fewer students achieved proficiency levels in the other English domains. For example, among P3 students, only 22.4% reached proficiency in Writing and 22.2% in Reading. With the exception of the Writing domain, a larger share of P6 students achieved proficiency-level scores.

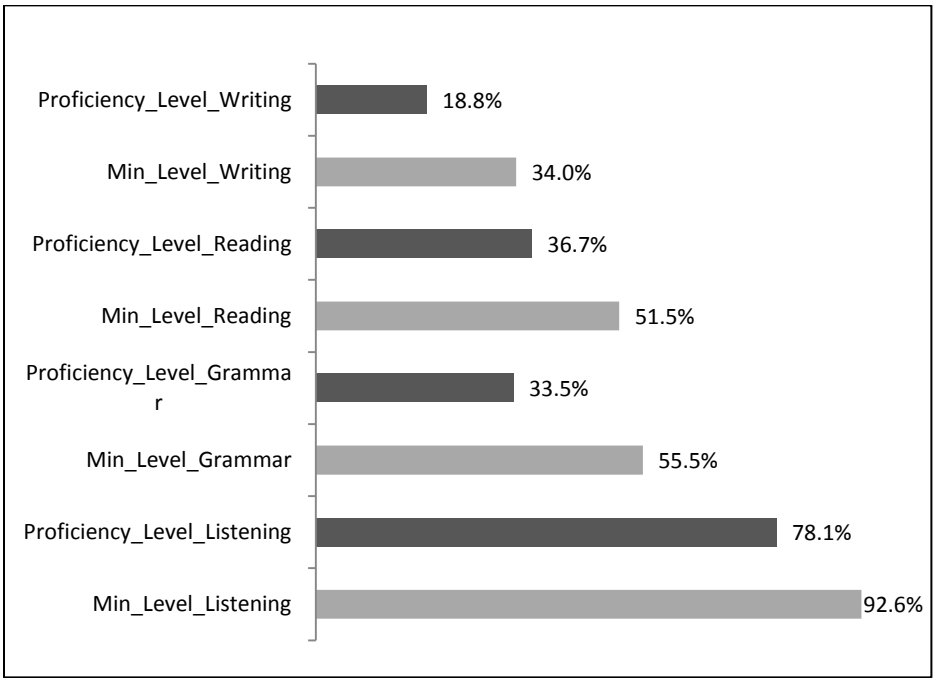
[26] Questions 31 through 36.
[27] Foundational reading skills include, among others, letter-sound recognition and simple decoding.

Figure 3.3 P3 students achieving minimum competency and proficiency scores on English test, by domain



Note: Confidence intervals may be found in Table D3 in Annex D.

Figure 3.4 P6 students achieving minimum competency and proficiency scores on English test, by domain



Note: Confidence intervals may be found in Table D5 in Annex D.

3.4.2 Maths Results by Domain

Mean domain maths scores indicate that the majority of students had not mastered the skills targeted in the assessment. The percentage of P3 students unable to answer a single domain question for the "Numbers" (7.8%), "Collecting and Handling Data" (16.1%), and "Shapes and Space" (17.8%) domains was substantial. Performance appeared strongest, although still weak, in the "Operations" domain, with P3 and P6 students achieving an average score of 42.9% and 42.4% respectively ***(Table 3.4).*** As discussed in Section 2.1, the Shapes and Space and Collecting and Handling Data domains contained only a small number of test items. The findings from these domains nonetheless provide some useful insights. P3 performance appeared weakest in the Collecting and Handling Data domain. As with the English assessment, foundational skills in maths (such as single-digit counting, number identification, or quantity discrimination) were not included in the test and, therefore, it is not possible to pinpoint the origin of students' difficulties. Many test items involved a combination of skills, making it difficult to understand which specific skill or sets of skills were more challenging to the students. In addition, as mentioned earlier, students' apparent difficulty reading English may have hindered their ability to read, understand, and respond to the mathematics questions. As with the English test, poor performance on all the maths domains indicates that additional teacher focus on fundamentals could be beneficial.

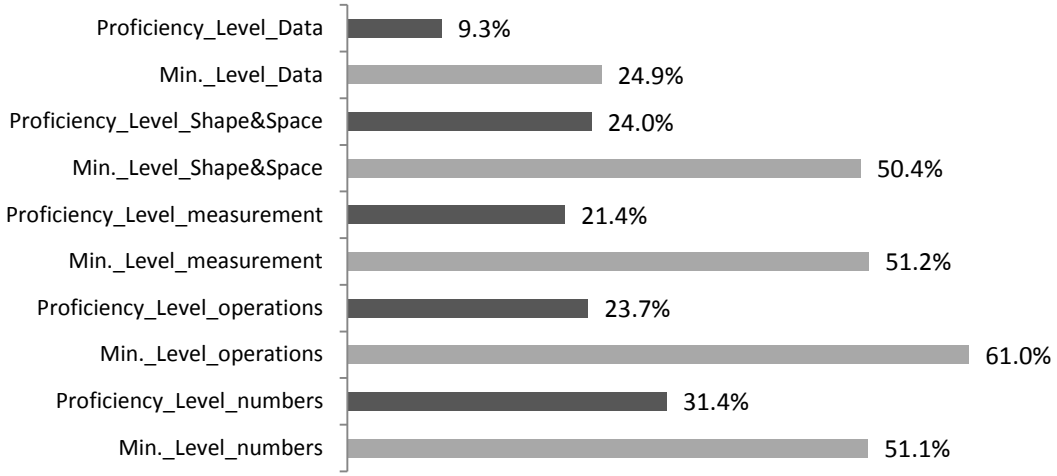
Table 3.4 Average maths test scores, by grade and domain

Domain	P3	P6
Numbers	39.3%	40.7%
Operations	42.9%	42.4%
Measurement	38.8%	Not applicable
Shapes and Space	41.2%	37.3%
Collecting and Handling Data	29.0%	44.0%

Note: Standard errors and confidence intervals may be found in Annex D, Tables D8 and D10.

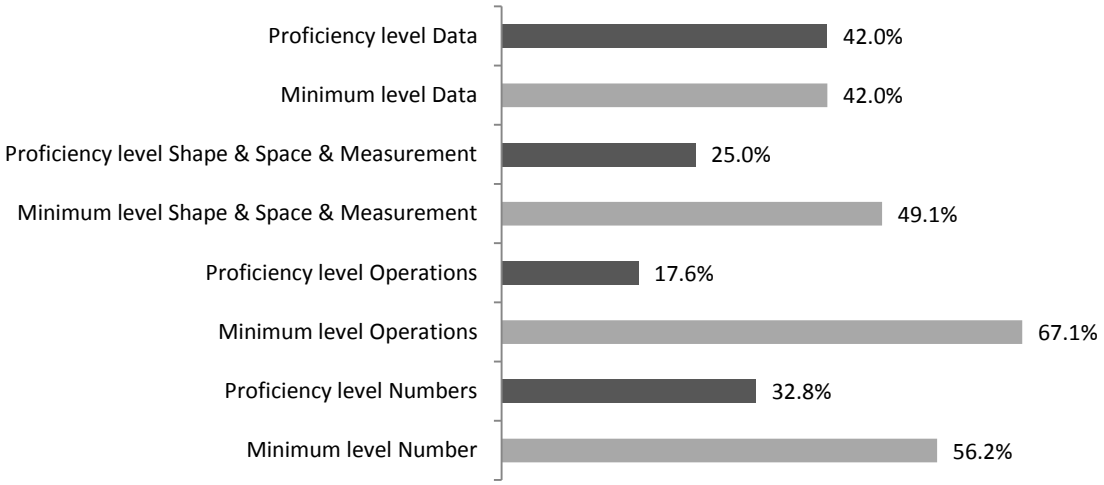
The percentages of students achieving minimum competency and proficiency levels are presented in ***Figures 3.5*** and ***3.6*** below. The difference in the percentage of students achieving proficiency-level scores among P3 and P6 students in Collecting and Handling Data is striking (9.3% versus 42%). However, given the small number of items in this domain, generalizing from this finding is inadvisable. Apart from this domain, the percentage of students achieving minimum competency and proficiency scores is similar. The numbers domain saw the greatest number of P3 and P6 students achieving proficiency levels. But even here, only just under a third of students were able to achieve the proficiency-level scores.

Figure 3.5 P3 maths results, by domain



Note: Standard errors and confidence intervals may be found in Annex D, Table D4.

Figure 3.6 P6 maths results, by domain

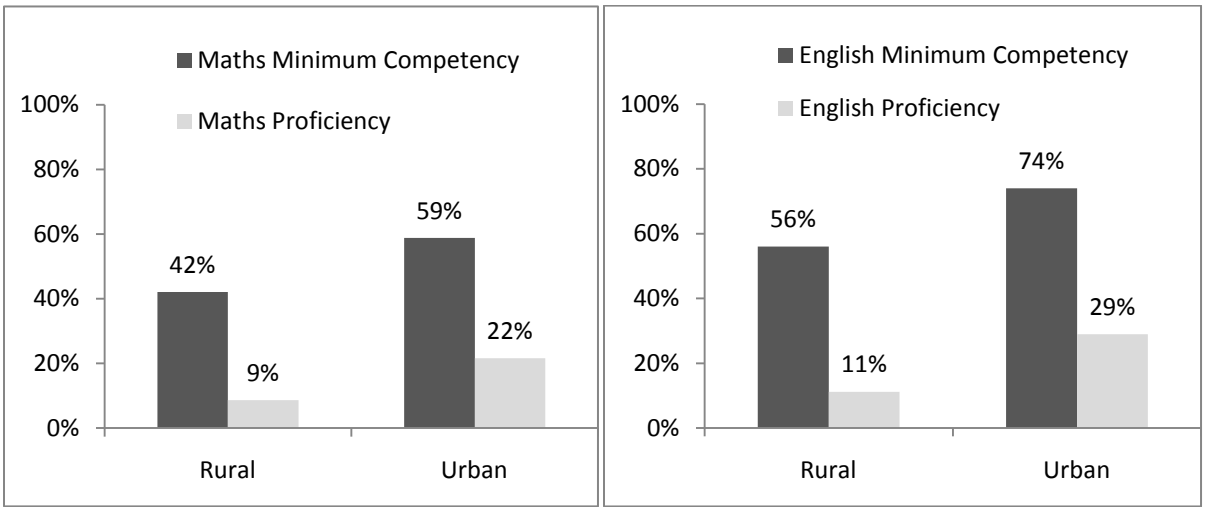


Note: Standard errors and confidence intervals may be found in Annex D, Table D6.

3.5 Achievements by Urban/Rural Location, School Type, and Region

Figure 3.7 shows the weighted percentages of students achieving minimum competency and proficiency levels for P3 maths and English tests, disaggregated by urban/rural location and school type. As can be seen, student performance in urban schools was much higher than in rural schools. More than twice as many students from the urban schools achieved proficiency-level scores on both the maths and English tests.

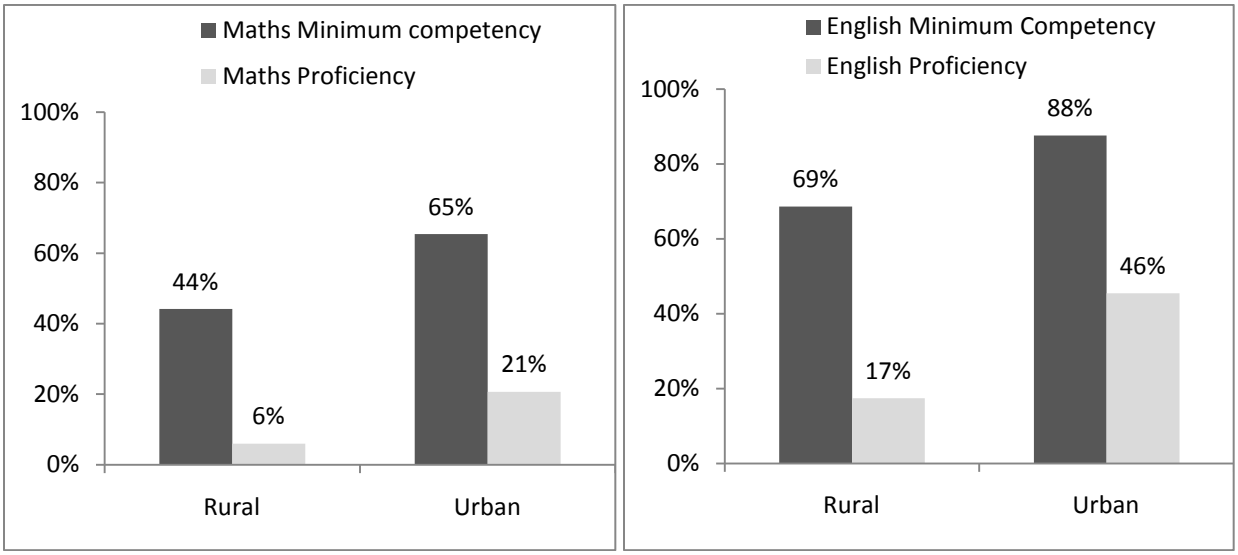
Figure 3.7 Percentage of P3 students achieving minimum competency and proficiency levels in maths and English, by urban/rural schools



Note: Confidence intervals may be found in Annex D, Table D1. Odds ratios appear in Annex H, Table H1.

The differences in performance between urban and rural P6 students is even more striking *(Figure 3.8).* Urban P6 students were more than three times more likely to achieve proficiency maths scores than their rural counterparts (21% versus 6%).

Figure 3.8 Percentage of P6 students achieving minimum competency and proficiency levels in maths and English, by urban/rural schools

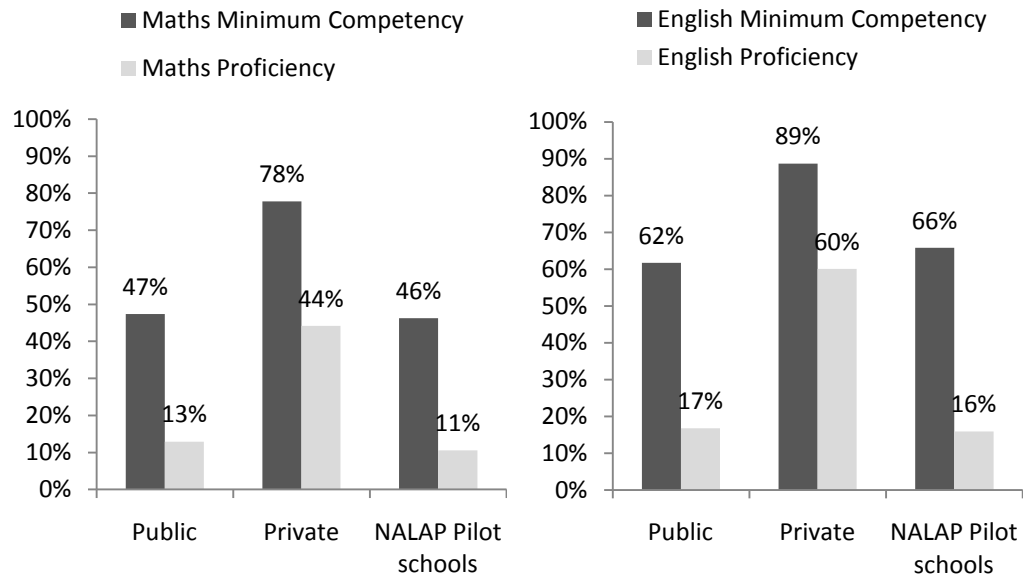


Confidence intervals may be found in Annex D, Table D2. Odds ratios appear in Annex H, Table H2.

In addition to looking at student performance by school setting, the research team looked for differences between public and private schools, and also differentiated between public schools that had been part of NALAP since its pilot phase in 2004 and those that had implemented the program only since 2010. The schools selected to participate in the pilot phase of NALAP seemingly would have had an advantage over the rest of Ghanaian public schools, in that they had been implementing the program for a longer period of time. Once NALAP went national, however, these pilot schools were no longer treated any differently than other public schools.

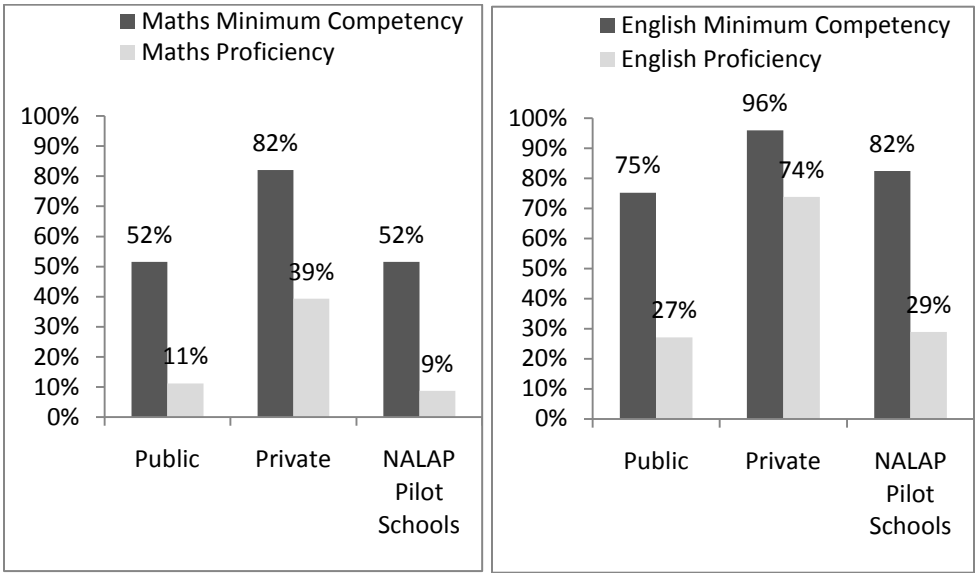
The 2009 NEA results found that the NALAP pilot schools had much higher English "proficiency" scores than other public schools that had not yet begun the program. For the 2011 NEA, although the share of students achieving minimum competency and proficiency scores at NALAP pilot schools does not, at first glance, appear to have been better than at public schools ***(Figures 3.9 and 3.10),*** when student gender, region, and school location are taken into consideration, students at NALAP pilot schools ***were*** more likely to have achieved minimum competency and proficiency on both the maths and English tests. This difference is more striking at the P6 level, where NALAP pilot students were 43% (odds ratio [OR] =1.43) more likely to achieve minimum competency in maths; 78% more (OR=1.78) more likely to achieve minimum competency in English; and 55% (OR=1.55) more likely to achieve proficiency in English. All of these differences were significant, with the exception of P6 students achieving proficiency in maths. The difference between NALAP pilot and other public schools was less pronounced among P3 students. P3 students in the NALAP schools fared only slightly better than the public schools in maths MC35 (OR=1.18) and significantly (statistically speaking) better in English MC35 (OR=1.38). Private schools outperformed public and NALAP pilot schools in both grades and on both subject tests.

Figure 3.9 Percentage of P3 students achieving minimum competency and proficiency levels in maths and English, by school type



Confidence intervals may be found in Annex D, Table D1. Odds ratios appear in Annex H, Table H1.

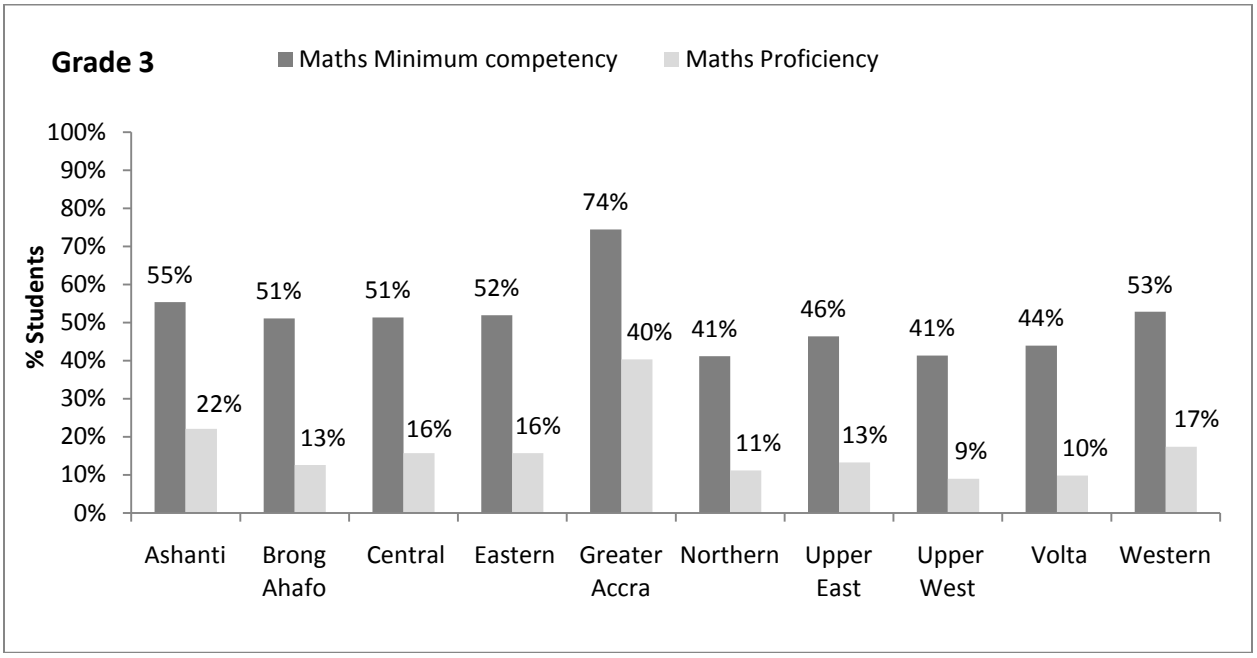
Figure 3.10 Percentage of P6 students achieving minimum competency and proficiency levels in maths and English, by school type



Confidence intervals may be found in Annex D, Table D2. Odds ratios appear in Annex H, Table H2.

Given differences in resource allocation by region, the research team was curious to see if student performance results varied by region. As shown in ***Figure 3.11,*** Greater Accra Region had the largest proportion of P3 students obtaining MC35 for maths (74% for MC35 and 40.% for PF55).

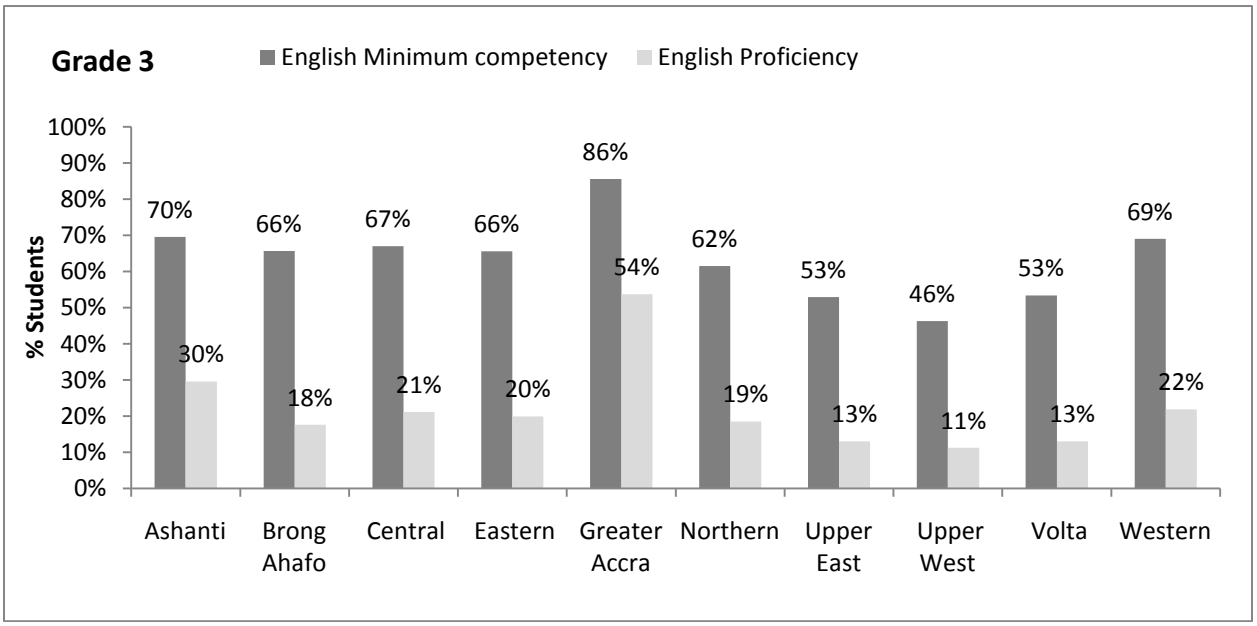
Figure 3.11 Percentages of P3 students achieving minimum competency or proficiency in maths, by region



Note: Confidence intervals are presented in Annex D, Table D1. Odds ratios appear in Annex H, Table H1.

Similarly, ***Figure 3.12*** indicates that proficiency and minimum competency levels for English were dramatically higher in Greater Accra Region's schools (86% for MC35 and 54% for PF55).

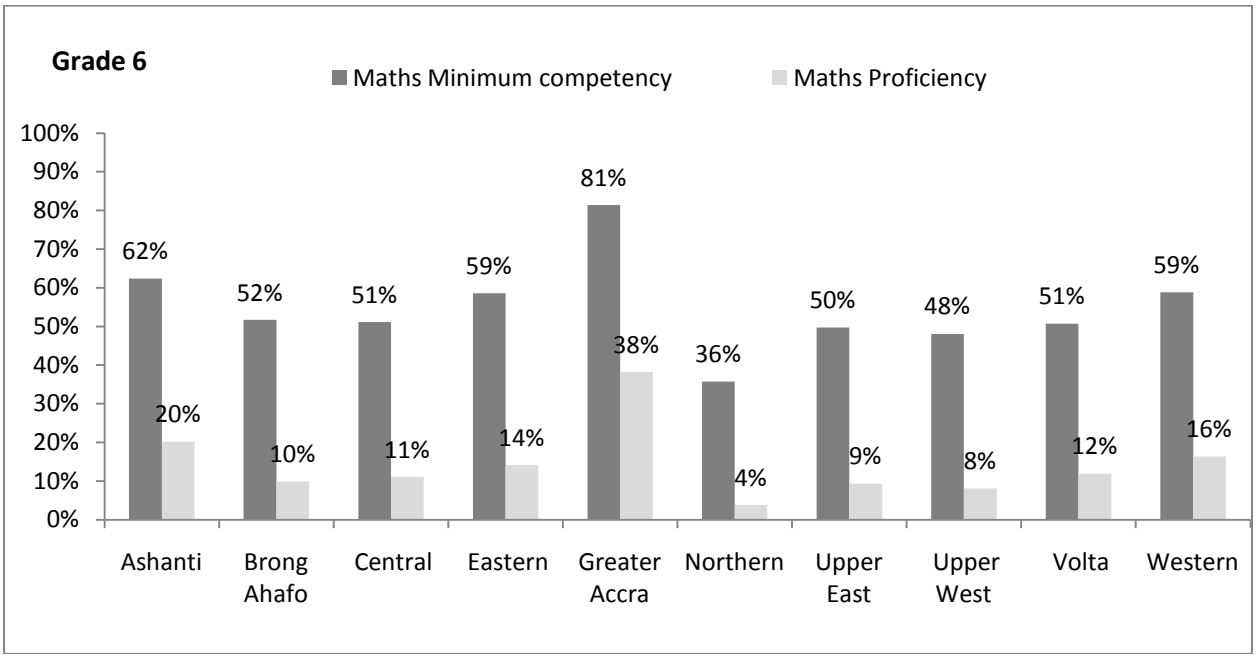
Figure 3.12 Percentages of P3 students reaching minimum competency or proficiency in English, by region



Note: Confidence intervals are presented in Annex D, Table D1. Odds ratios are presented in Annex H, Table H1.

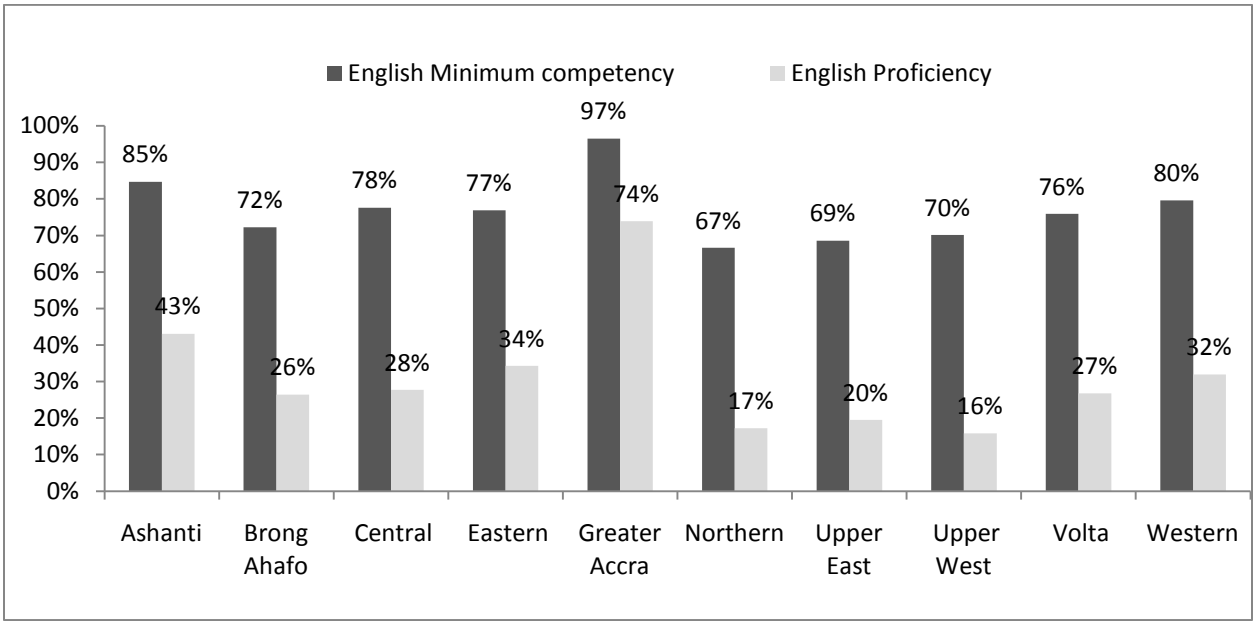
Not surprisingly, the P6 students from Greater Accra region also outperformed all other regions in maths and English; see ***Figures 3.13 and 3.14,*** which present P6 minimum-competency and proficiency data for English and maths by region.

Figure 3.13 Percentages of P6 students reaching minimum competency and proficiency in maths, by region



Note: Confidence intervals are presented in Annex D, Table D2. Odds ratios appear in Annex H, Table H2.

Figure 3.14 Percentages of P6 students reaching minimum competency and proficiency in English, by region



Note: Confidence intervals are presented in Annex D, Table D2. Odds ratios are presented in Annex H, Table H2.

It is worth noting that nearly all grade 6 students achieved minimum competency in English in the Greater Accra Region (97%). Even more important is the fact that nearly three quarters of students in the Greater Accra region achieved proficiency-level scores in English.

4. Factors Related to Learning Outcomes

4.1 Brief Literature Review of Factors Affecting Learning Outcomes in the African Context

From the existing literature, a framework can be produced to categorize factors influencing specific learning outcomes and to identify areas of action.[28]

Variables are generally split in two categories:

- Contextual or demographic variables that go beyond the scope of education interventions, such as socioeconomic status, students' age, household living conditions, and rural/urban status.
- Policy variables in which the education system can intervene: teacher training, class size, time on task, availability of pedagogic materials.

Being able to link school, community, and student characteristics to student performance allows researchers and policy makers to identify factors that are most closely associated with strong or poor performance. This information, in turn, allows policy makers to develop policies tailored to improve student performance.

Although the effect of specific factors varies from country to country, there are some fairly commonly occurring patterns. For instance, a meta-review of *Programme d'Analyse des Systèmes éducatifs des États et gouvernements membres de la CONFEMEN* (PASEC) assessments in 14 sub-Saharan (mostly West African) countries identifies:

- Factors generally associated with lower learning outcomes: repetition, students' age, female gender, over age, poverty, teacher absenteeism, overcrowded classrooms, child labor, and rural location.
- Factors generally associated with higher learning outcomes: the use of mother tongue as language of instruction, possession of textbooks that students can take home, initial and in-service training of teachers, availability of teacher's guidebooks, student homework, parents' literacy, better child nutrition, teachers' academic level of attainment, and school inspection.[29]

[28] Gillies (2008).
[29] PASEC (2010), *Diagnostic et préconisations pour une scolarisation universelle de qualité.* Dakar: Rapport Union des Comores, CONFEMEN, pp. 98–99.

Recent Southern and Eastern Africa Consortium for Monitoring Education Quality (SACMEQ) data yield similar results for southern Africa.[30]

4.2 Demographics and Contextual Variables

Table 4.1 presents the research questions that can be addressed with EMIS data, the indicators that can be calculated, and the exact source. The final column is marked YES when the indicator has a significant effect on the students' competencies, and NO otherwise.

Table 4.1 Effects of demographic and contextual variables

Research question	EMIS reference	Indicator	Significant effect?
Do girls perform less well than boys?	Included in the NEA	Female vs. male student	YES, especially in mathematics
Do urban schools perform better than rural schools?	1.4; 1.8	Urban/rural classification	YES
Do remote schools perform less well than others?	2.11	A school is remote if it is at least 10 km from district office, head house, or next primary school	YES
Do students having health problems have lower results?	13.1;13.2; 13.3	% students deceased last year; % teachers deceased; % orphans	YES, at least when measured as % orphans
Are higher dropout rates associated with lower learning outcomes?	10.4	% dropout: 0%, less than 2%, more than 2%	YES
Are higher repetition rates associated with lower learning outcomes?	10.5	% repeaters: Less than 10%, more than 10%	YES
Is there a relationship between the proportion of students transferring FROM another school and learning outcomes?	10.1	% students transferred FROM another school: 0%, less than 5%, more than 5%	YES, positive relationship
Is there a relationship between the proportion of students transferring TO another school and learning outcomes?	10.2	% students transferred TO another school: 0%, less than 5%, more than 5%	YES, positive relationship

Note: For all of these analyses, gender, region, school location, and test form are taken into consideration. Thus, the effect reported was appropriately adjusted for any effect these factors may have had on student performance.

Variables reported in Table 4.1 as having an effect on students' competencies were those for which the logistic regression models produced a significant coefficient for at least one

[30] See the SACMEQ website, http://www.sacmeq.org/PostersSACMEQIII.htm, retrieved Dec. 14, 2011.

grade or subject. (Annex H contains details regarding the logistic regressions.) The graphs presented in the following subsections, which further examine various probabilities, show the percentages of students reaching minimum competency and proficiency levels according to the different variables.

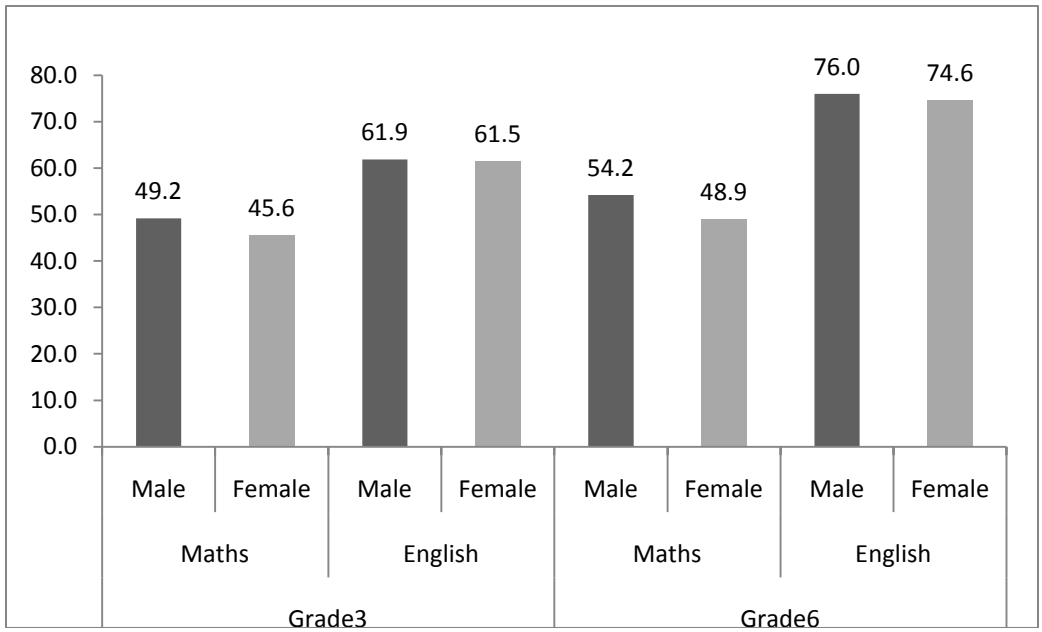
Note: EMIS information was missing for 88 private schools of the sample. The following results are for public schools only.

4.2.1 Gender Effects

Recent PASEC data from francophone Africa show small differences between female and male students' learning outcomes in most countries,[31] and SACMEQ data show that little narrowing of the gender gap has been achieved over time in anglophone Africa.

In Ghana, while progress has been made in gender parity as it concerns access to school, differences between girls' and boys' learning outcomes remain and are significant in mathematics. These differences in maths performance between genders are larger than in many African countries. Moreover, in maths, the gender gap does not diminish between P3 and P6. There is a 5.3 percentage-point gap between P6 girls and boys who reach minimum competency in maths, and a 2.9-point gap in those reaching proficiency ***(Figures 4.1 and 4.2).***

Figure 4.1 Percentage of students reaching minimum competency, by grade, subject, and gender

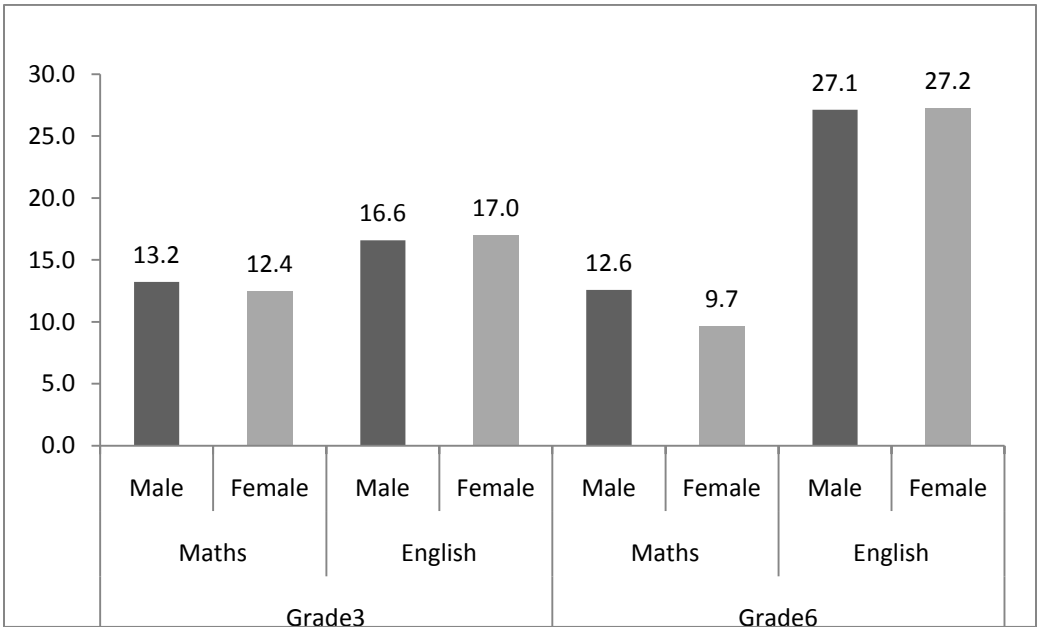


Note: Odds ratios are presented in Annex H, Tables H1 and H2.

[31] World Bank EdStats database, retrieved Dec. 14, 2011.

Logistics regression that takes into consideration region, school location, and school type indicates that P3 girls were 16% less likely (OR=0.84) to achieve minimum competency scores than boys in maths, but were only 3% less likely to achieve minimum competency in English (OR=0.97). Note that the difference in English scores was not statistically significant. Similarly, logistics regression indicates that P6 girls were 25% less likely (OR=.75) to achieve minimum competency in maths and 36% less likely (OR=.64) to achieve proficiency in maths than were boys.

Figure 4.2 Percentage of students reaching proficiency, by grade, subject, and gender



To reduce the gender gap in mathematics, programs and manuals could be reviewed to include more girl-friendly content, and maths teaching to girls should be monitored through lesson observations. A recent CONFEMEN/Forum for African Women Educationalists (FAWE)[32] report contains a survey framework that could serve as a good basis for designing a specific gender component in a classroom observation protocol.

Teacher training could also include gender issues and could aim to modify teachers' representation and persistent stereotypes of girls beings "less talented" in maths.

4.2.2 School Type and Location

As noted in Section 3, differences between rural and urban schools were considerable and significant for both grades and subjects, taking into account students' ability to reach

[32] CONFEMEN/FAWE. (2010). *Genre et acquisitions scolaires en Afrique, Etude sur les performances des élèves au cycle primaire*, draft report.

minimum competency or proficiency levels (see odds ratios in Annex H). The percentage of students dropping out also was higher in rural schools than in urban ones. The data do not yield information on languages spoken at home, but urban children are likely more exposed to English at home and in their living environment than rural children, explaining part of the difference in terms of English competencies.

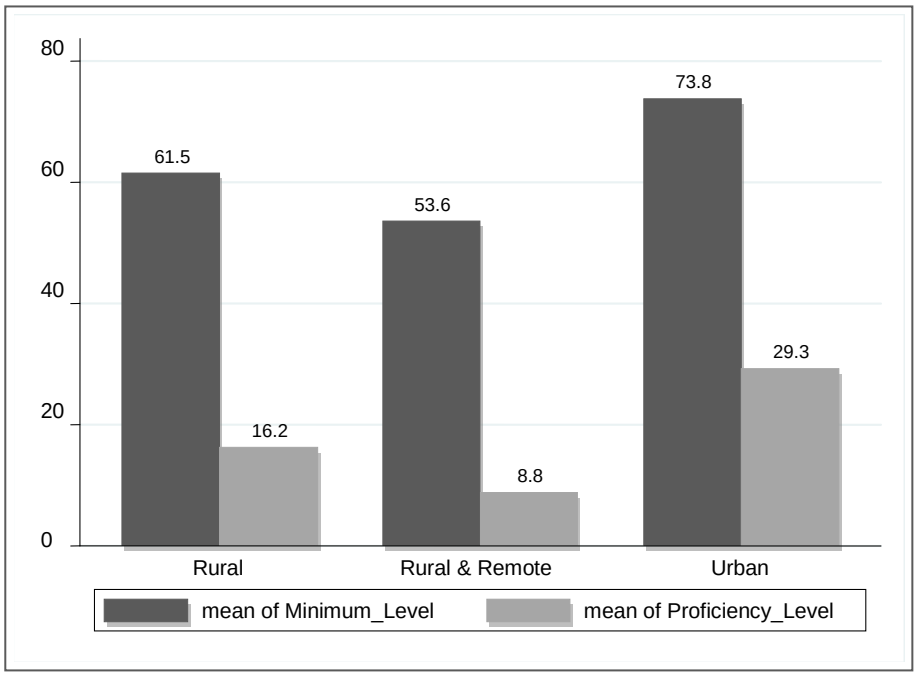
Going further into analyzing the effect of school location, "Remote schools" could have been defined by variables such as *"not accessible by a vehicle"* or *"the road is not tarred,"* but the accessibility of the school by a vehicle does not seem to have had any significant effect on students' competencies (see odds ratios in Annex H). Instead, a school has been defined as "remote" when:

- Head teacher's house is more than 5 km (approx. 1-hour walk) from school, or
- District Education Office is more than 10 km from school, or
- Next primary school is more than 10 km away.

When the analysts controlled for school location (urban/rural) and region, the remoteness of a school did have a significant effect on students' competencies, as the odds ratios were significant for both grade and subject and for minimum and proficiency levels (Annex H). Larger differences between remote and non-remote schools (in terms of both statistical significance and magnitude) were found in P3 in English ***(Figure 4.3).***

The probability of P3 students from remote schools reaching proficiency in English was found to be only half that of students in non-remote schools. Remote schools might be suffering from lower student or even teacher attendance, and they also likely benefit less from circuit supervision or in-service training than other schools.

Figure 4.3 Percentage of students reaching minimum competency and proficiency levels, by school location, P3, English

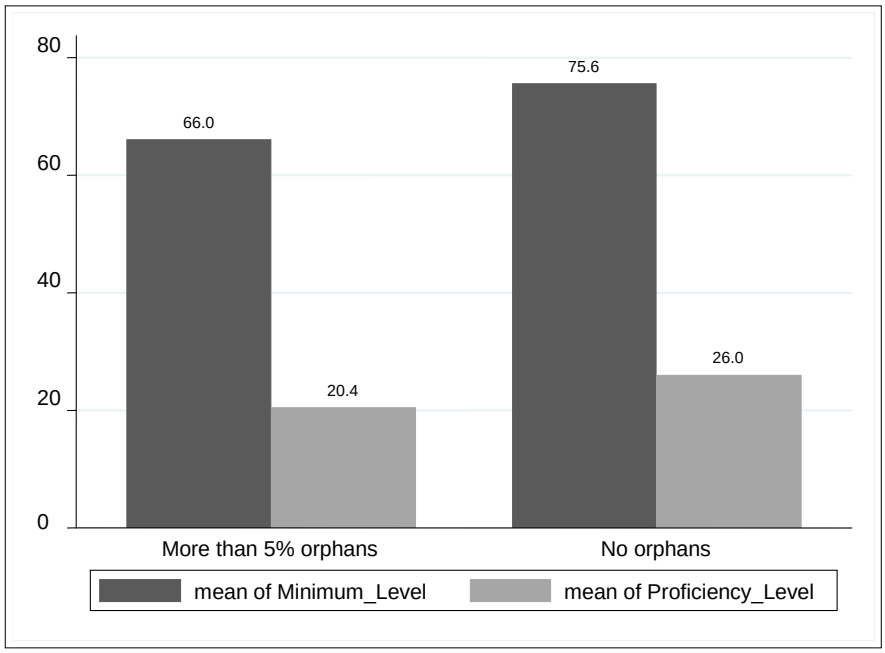


4.2.3 Health Conditions and Mortality

In order to capture other contextual factors, such as living conditions and mortality, EMIS data contain information on the number of students who are orphans (i.e., have lost one or both parents), as well as the number of students and teachers that have died. Considered as proxy indicators for health conditions, these figures are compared to total enrollment to calculate (1) percentage of students who are orphans and (2) percentage of students or teachers deceased.

The "% orphans" factor was found to have an effect on reaching proficiency or minimum competency, particularly in P6 English ***(Figure 4.4).*** Schools that had no orphans (or did not report any) had better results.

Figure 4.4 Percentage of students reaching minimum competency and proficiency levels, by % orphans, P6, English



4.2.4 Other Demographic Variables

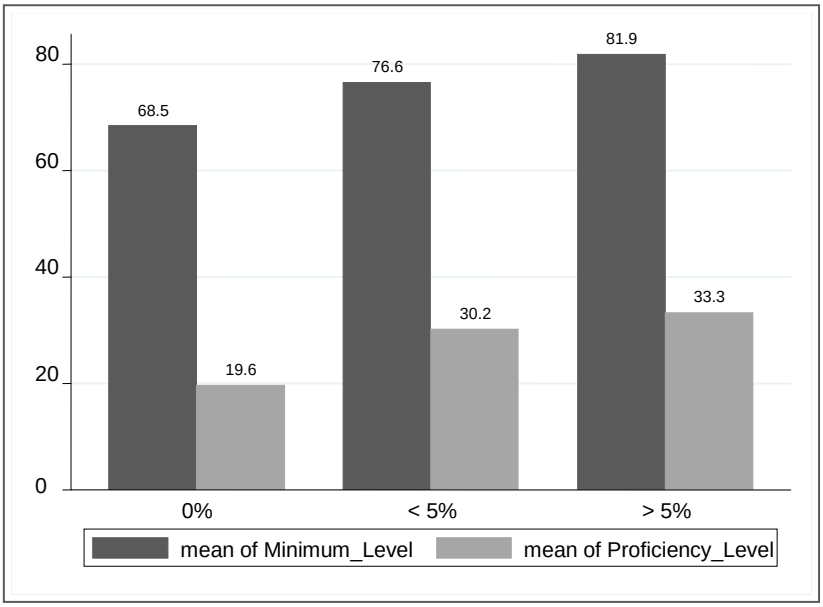
The EMIS data contain relevant information on students' in-school and between-school flows. This is measured through:

- Repetition rates
- Dropout rates
- Students transferred to another school
- Students transferred from another school

A higher proportion of students dropping out was found to be associated with poorer learning outcomes only in P6 ***(Figure 4.5).*** Similarly, a higher repetition rate (especially as compared to Ghana's average of 4%) at the school level was found to be associated with lower learning outcomes in P6. Of interest within the overall picture regarding repetition is the fact that while 57% of the schools reported 0 repeaters, 5% of the schools reported more than 15% of students were repeaters.

Looking at between-school transfers (flows), the percentage of students who transferred TO or FROM another school was found to be associated with higher learning outcomes. If a transfer to another school resulted from the parents' decision, this indicator of school choice might reflect parents' commitment to a better education for their children.

Figure 4.5 Percentage of students reaching minimum competency and proficiency, by percentage of students transferring from another school, P6



4.3 Teaching and Learning Environment

This section describes findings related to the teaching and learning environment. Similar to Table 4.1 above, ***Table 4.2*** presents research questions, exact source of data, indicators, and significance.

Table 4.2 Effects of variables on teaching and learning environment

Research question	EMIS reference	Indicator	Significant effect?
Do students in large classes have lower results?	1.11	Student/teacher ratio less than 35; between 35 and 70; above 70	Some effect but not significant and also highly correlated to rural location
Do multigrade classes experience lower learning?	2.1	School has at least one multigrade classroom	YES
Do better-resourced schools (electricity, water, and toilets) tend to have higher learning levels?	3.3; 3.6; 3.1.0	Number of items (water, functioning electricity, toilets)	YES
Does school feeding program improve attendance, retention, and learning?	4.10	School has free meals program	No improvement in retention or learning

Note: For all of these analyses, gender, region, school location, and test form are taken into consideration. Thus, the effect reported was appropriately adjusted for any effect these factors may have had on student performance.

Schools were grouped into three categories according to their student/teacher ratio:

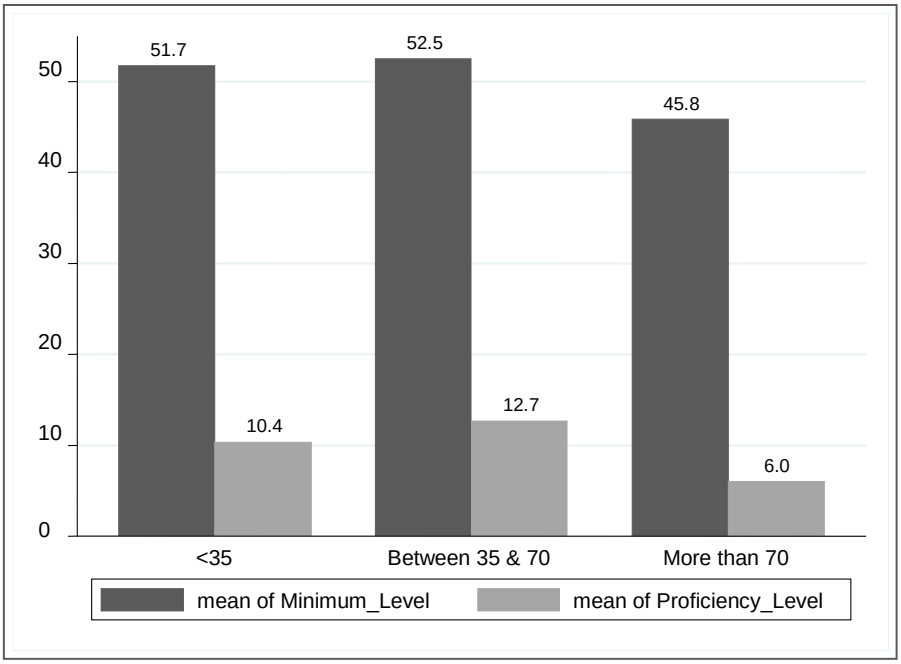
- Fewer than 35 students per teacher
- Between 35 and 70 students per teacher

- More than 70 students per teacher

This indicator served as a proxy for class size, as the research team did not have access to information about how many students were in each sampled class or how many teachers served each grade.

In the data that *were* available, very few urban schools had more than 70 students per class (only 3.8% of urban schools vs. 10.2% in rural schools). Overcrowded schools (over 70 students per teacher on average) were found to have lower performance than others ***(Figure 4.6)*** but effects were not significant for any grade or subject, as the effect might have been captured by the variable for urban/rural location**.**

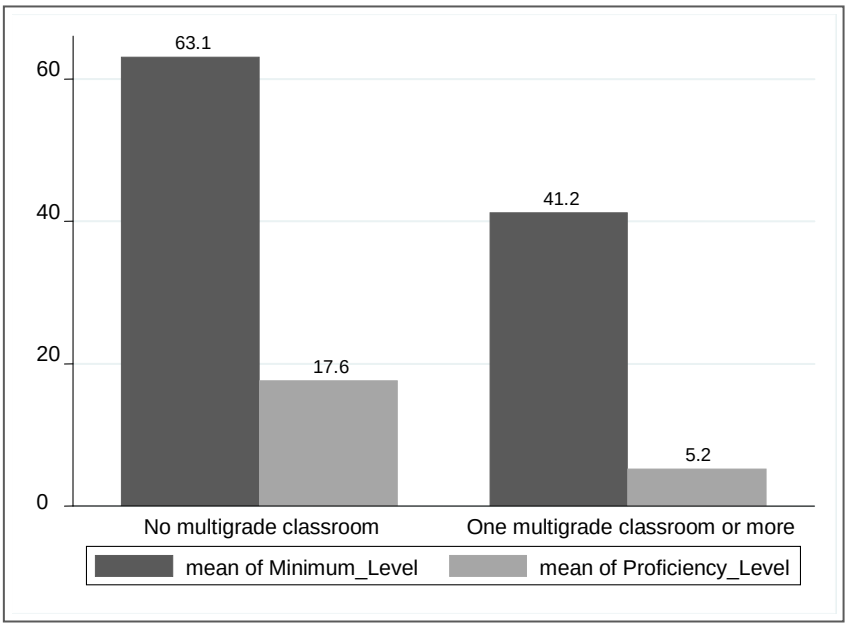
Figure 4.6 Percentage of students reaching minimum competency and proficiency, by class size, P6 maths



Multigrade classrooms were found to be associated with poor learning outcomes in nearly all grades and subjects taught ***(Figure 4.7).*** Only 3.5% of P6 students in schools with multigrade classrooms were found to reach proficiency in maths, and roughly two-thirds of these students did not even reach minimum competency.

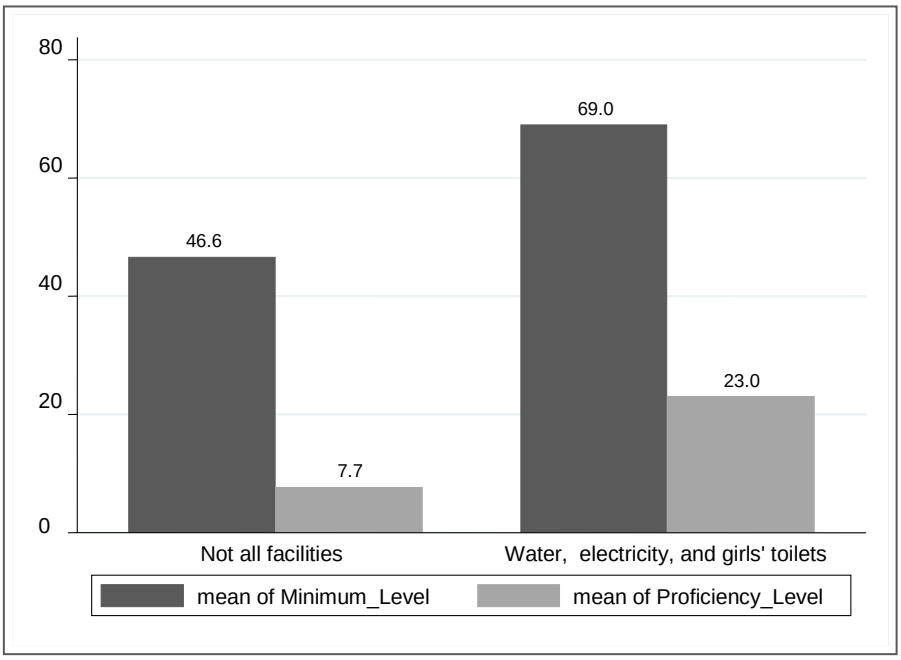
Teaching in multigrade classrooms presents particular challenges as the teachers must prepare multiple lessons to be able to teach students learning at different grade level. Teachers must also be able to organize the day in such a way that each of the grades are given sufficient instructional time.Teachers do not seem to be well-prepared for multigrade teaching. Twenty-three percent of schools with at least one multigrade classroom had no trained teachers, versus 6% in the schools with no multigrade classes. Moreover, the data showed that multigrade schools were not benefiting from more circuit supervor visits or in-service training than non-multigrade schools.

Figure 4.7 Percentage of students reaching minimum competency and proficiency, by multigrade classrooms, P3 English



Electricity, water, and toilets for girls are assembled into one indicator called "school facilities" ***(Figure 4.8).*** In the majority of schools that did not have all of these facilities, the proportion of students reaching minimum competency was 46.6%. By contrast, 69% of students in schools having all of these facilities reached minimum competency. These facilities are also individually associated with higher skills.

Figure 4.8 Percentage of students reaching minimum competency and proficiency, by school facilities, P6 maths



A final indicator in this category, the existence of a school feeding program, was found to have no effect on learning or dropout rates but can positively impact both students' enrollment and attendance rates.

4.4 Teaching and Learning Materials and Resources

This section discusses findings related to the availability and use of teaching and learning materials. As before, ***Table 4.3*** indicates the relevant research questions, EMIS source, indicators, and significance.

Table 4.3 Effects of variables on teaching and learning materials

Research question	EMIS reference	Indicator	Significant effect?
Do textbooks/teacher guidebooks improve students' learning?	7.1; 7.2; 9.5	Textbook/student ratio; at least one teacher guidebook in school	YES
Does having library books at the school improve students' learning?	5.4	Ratio: books per student	NO
Does the use of information and communication technology (ICT) improve students' learning?	proxy	Indicator of equipment available in the schools rather than use of ICT in the classroom	YES, but this is an indicator for the presence of equipment, not for its use

Note: For all of these analyses, gender, region, school location, and test form are taken into consideration. Thus, the effect reported was appropriately adjusted for any effect these factors may have had on student performance.

The textbook-to-student ratio was categorized into three groups:

- No textbooks in the school
- Fewer than 0.5 textbook per student on average
- More than 0.5 textbook per student on average.

Despite better materials provision in the last year compared to prior years, there remained a significant proportion of students in schools without any textbooks. Consequently, these students tended to underperform ***(Table 4.4a).***

Table 4.4a Percentage of students in schools without textbooks, by subject and grade

Subject	P3	P6
English	7.7%	8.1
Maths	12.4%	13.1%

Pedagogic materials are essential for both students and teachers. Teachers need textbooks and reference materials to help them properly follow the ministry curriculum. Students that do not have access to textbooks tend to spend more valuable class time copying down materials from the chalkboard and less time actively learning.[1] Though EMIS data indicate that relatively few students attend schools with no text books (***Table 4.4a***), the majority of schools had English textbooks for fewer than half of their students and, on average, classrooms had English textbooks for slightly under a quarter of their students (average book to student ratio=.22, ***Table 4.4b***). The scarcity of English books was strongest in remote areas where the text book per student ratio was .19.

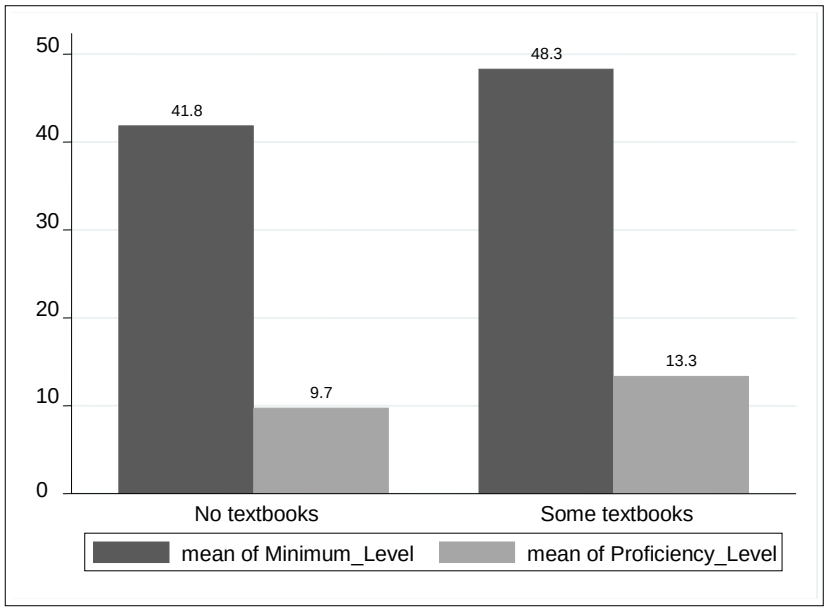
EMIS data indicate that between 12.4 and 13.1% of P3 and P6 students attend schools have no mathematics textbooks. However, the overall availability of maths text books is greater as the textbook to student ratio is .51 (***Table 4.4b***) indicating that, on average there is one text book for every two children. This seemingly contradictory information indicates that there must be a fairly wide disparity in the allocation of maths textbooks. Surprisingly, rural schools have a slightly higher textbook to student ratio than do urban schools (.57 vs. .54) while remote schools once again are less well resourced with a textbook to student ratio of .45.

[1] Helen Abadzi. (2007, October). *Absenteeism and beyond: Instructional time loss and consequences*. World Bank Policy Research Working Paper Series, p. v.

Table 4.4b Average textbook-to-student ratio, by subject and school location

Subject/school location	Textbook-to-student ratio
English	0.22
Maths	0.51
English	
Rural and remote	0.19
Rural	0.34
Urban	0.25
Maths	
Rural and remote	0.45
Rural	0.57
Urban	0.54

Figure 4.9 Percentage of students reaching minimum competency and proficiency, with or without maths textbooks, P3 maths



4.5 School Management and Community Involvement

The following section discusses findings related to school management practices and community involvement. ***Table 4.5*** indicates the relevant research questions, EMIS source, indicators, and significance.

Table 4.5 Effects of variables on school management and community involvement

Research question	EMIS reference	Indicator	Significant effect?
Do schools where communities are involved (financing, monitoring, material support…) have better performance?	4.1; 4.2	Existence of a school management committee (SMC)	NO
Do schools with performance improvement plan have better performance?	4.3	Existence of a school performance improvement plan (SPIP)	NO
Do schools that keep registers have better performance?	6.5	Number and types of registers	YES

Note: For all of these analyses, gender, region, school location, and test form are taken into consideration. Thus, the effect reported was appropriately adjusted for any effect these factors may have had on student performance.

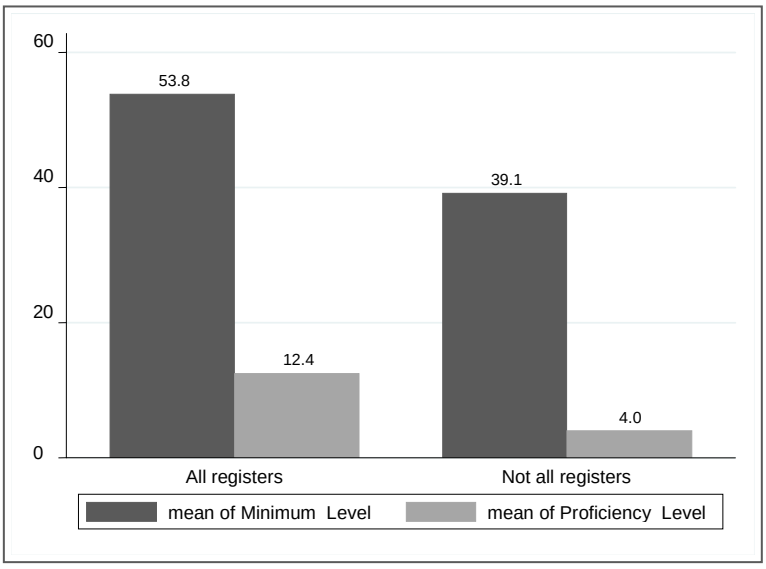
The existence of a school management committee (SMC) or a school performance improvement plan (SPIP) did not have a significant effect on learning outcomes. The data do not reveal, however, the attributes of the school management committee, how this committee works to monitor student and teacher attendance or even learning outcomes, or anything about the management of capitation grants. (Information on capitation grants was not reported by schools in the EMIS data.)

The data also do not indicate (for schools that had them) if the SPIP focused on learning outcomes, or specific objectives in terms of reading skills; or if it included a component on time-on-task and students' attendance, for instance.

However, having an SMC or SPIP reduced by half the proportion of schools not keeping all administrative registers such as admission or classroom registers, inventories, teachers' attendance registers, logbooks, or visitors' books.

Keeping registers is an indicator of better school administration/management. As the keeping of registers in the school is associated with a better chance for students to reach proficiency, SMCs and SPIPs might indirectly improve learning outcomes. In schools that were not keeping all types of registers, only 4% of P6 students had reached proficiency in English, versus 12.4% of students in schools that were keeping all registers ***(Figure 4.10).***

Figure 4.10 Percentage of students reaching minimum competency and proficiency, by registers kept in school, P6 maths

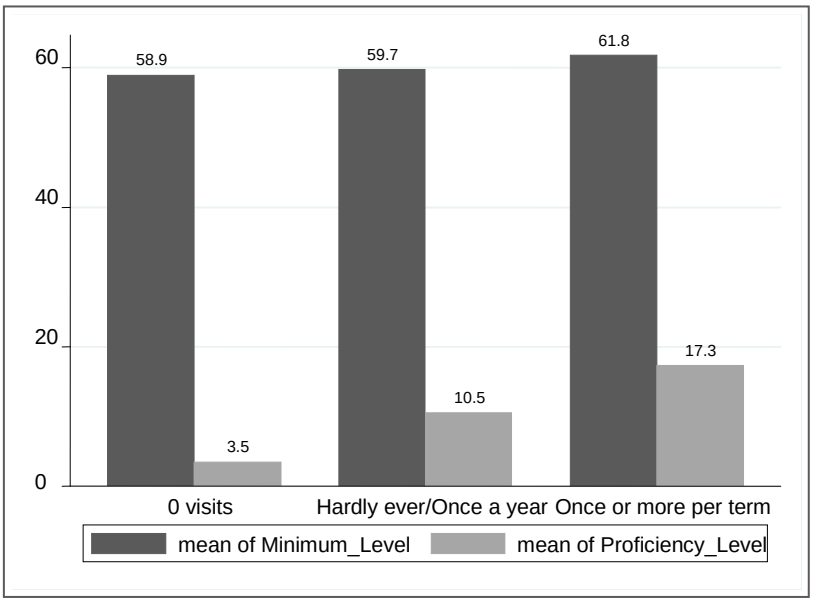


Visits from a circuit supervisor were found to improve record-keeping practices, especially inventories and visitors' books. These visits also were found to have a net positive effect on test scores ***(Figure 4.11).***

Students in schools that were more frequently visited by circuit supervisors were found to be far more likely to reach proficiency than those in schools that were not visited often (17.3% versus 3.5% in English in P3). Visits from circuit supervisors was also found to be among the strongest determinants of more time spent on task.[33]

[33] World Bank (2004). *Books, buildings and learning outcomes: An impact evaluation of World Bank support to basic education in Ghana.* Washington, DC: World Bank.

Figure 4.11 Percentage of students reaching minimum competency and proficiency, by visit from circuit supervisor, P3 English



4.6 Teacher Characteristics and Practices

This section discusses associations between teacher characteristics and practices and learning outcomes. It is important to keep in mind that the data are at the school level only. Therefore, for example, it is not possible to associate one student's performance with the characteristics of his or her teacher. Indicators (see ***Table 4.6***) are calculated as a share of teachers in the school.

Table 4.6 Effect of variables on teacher characteristics and practices

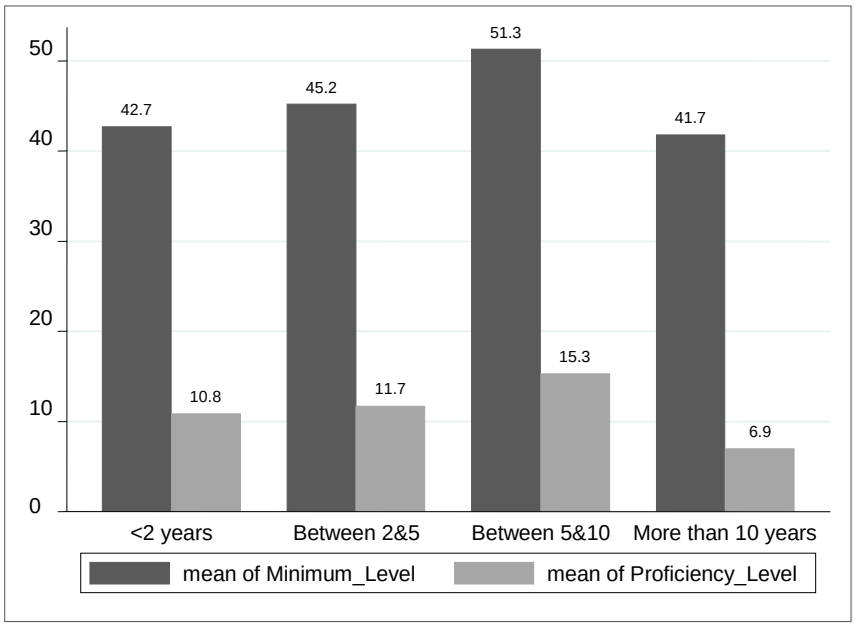
Research question	EMIS reference	Indicator	Significant effect?
Is the proportion of female teachers in the school associated with student performance?	11	% female teachers in the school	YES
Do more experienced teachers have better results?	11	Average years in service at the duty station	NO
Do schools with a larger proportion of trained teachers have better student performance than others?	11	% trained (qualified) teachers among classroom teachers	YES
Is in-service training associated with better performance?	4,12;4,14	Frequency of school-based in-service training; frequency of visits of in-service trainers	NO

Note: For all of these analyses, gender, region, school location, and test form are taken into consideration. Thus, the effect reported was appropriately adjusted for any effect these factors may have had on student performance.

The average years of teachers' experience in a particular school was found to be associated with better learning outcomes only in P3 ***(Figure 4.12).*** The results show steadily greater performance among students when the average of teachers' time at the

school increases from 2 years up to 10 years. After 10 years at the same school, performance levels drop. Again, note that accurately evaluating the impact of an individual teacher's years of experience or education on student performance requires linking a specific teacher's characteristics to his or her students.

Figure 4.12 Percentage of students reaching minimum competency and proficiency, by average years of teacher experience in the school, P3 maths



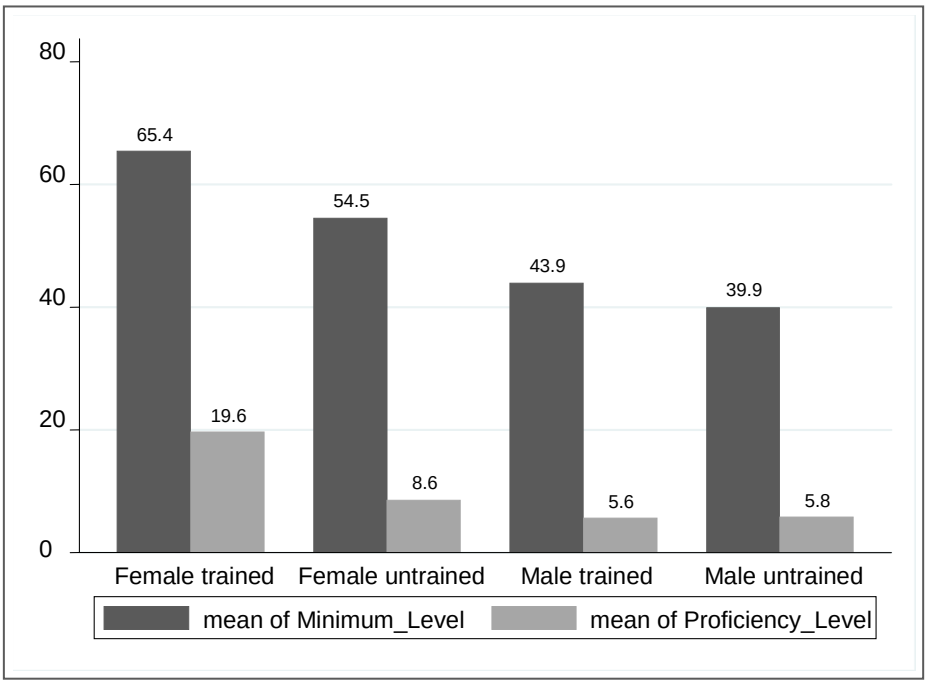
Teachers' training can have an effect on learning outcomes only when the training modifies teachers' practice. In this study, the frequency of visits from in-service trainers or school-based in-service training was not found to have any effect on learning outcomes ***(Figure 4.13).***

In Ghana, teacher training—measured here as the proportion of trained teachers in a school—is highlighted in several donor or government documents as a key factor for improving learning outcomes.[34] However, studies have shown conflicting information on the effect of teacher training. According to the NEA data, teacher training measured as the percentage of qualified teachers in the school is associated with better learning outcomes for nearly all grades and subjects. It is important to emphasize that the EdData II research team was not able to link specific teachers and their qualifications to specific students, and therefore instead examined only overall training levels among the teachers at the schools.

Note also that if trained teachers are being allocated to better schools, it is possible that these findings are the result of teacher allocation and not teacher training.

[34] See, for example, World Bank (2011).

Figure 4.13 Percentage of students reaching minimum competency and proficiency, by teachers' gender and training, P6 English



5. Recommendations Based on NEA Findings

This section summarizes the research team's reactions and recommendations based on the analysis results reported above. The concerns are grouped into four key areas: allocation of resources, teacher training needs, school management, and logistics of the test administration itself.

5.1 Resource Allocation

Ghana's significant investment in primary education should be maintained, while per child expenditures should be made more equitable across regions and districts. As noted earlier, enrollment figures are increasing and schooling conditions are improving with programs such as school feeding, among other initiatives.

Nevertheless, important geographical disparities remain. Rural remote schools have significantly lower results than others, especially in the Northern Region and in poorer learning conditions (large classes—i.e., with 70 or more students—and untrained teachers). A significant proportion of schools have no textbooks.

Potential measures for reducing these disparities include:

- **using capitation grants to reduce large disparities in the availability of school equipment and to secure minimum allocations of textbooks in each school;**

- **reviewing the funding allocation formula or developing school-level allocation formulas for use at the district level;**
- **publishing school funding and resource allocations so that parents can help monitor allocations (following a similar initiative in Uganda);**
- **initiating an Education Fund Tracking Survey (EFTS).**

The student/teacher ratio has remained stable over time (32) and is well below the African average. Although this indicator clearly demonstrates Ghana's commitment to maintaining reasonable class sizes and acceptable schooling conditions, it has implications for the budget, as most education spending is geared toward recruiting and training teachers.

Better teacher allocation would allow more funding for nonsalary expenses, such as textbooks and school facilities, which have positive effects on learning outcomes. Additionally, better teacher allocation could also serve to reduce the number of multigrade classrooms, which have significantly lower results than others. **Schools that must maintain multigrade classrooms should receive specific pedagogical supervision and in-service training.**

Finally, as the proportion of students transferring to or from other schools was found to be associated with higher learning outcomes, this may mean that parents are seeking better education for their children. This social demand for education, a necessary condition for achieving universal enrollment, should be supported by efforts toward more equitable distribution of funds, teachers, and materials, with a specific focus on deprived districts. Although teachers receive financial support to teach in certain zones, they should also be trained to teach to specific student populations that have little contact with the English language at home or in their local environment, as within NALAP.

5.2 Teacher Training

Results indicate that teachers are not teaching several subjects effectively, given that:

- half of the students studied did not reach minimum competency level in the core maths domains (Numbers and Operations) at P3 and P6;
- the results are especially low in domains such as Writing, Shapes and Space, and Collecting and Handling Data, with only roughly 20% of the students reaching proficiency; and
- data show similar patterns in P3 and P6.

NEA data collection should include classroom observations to evaluate teaching approaches. Data gathered through these observations could then be fed into the design of in-service and pre-service teacher training materials. Additionally, specific teacher training could be focused on teaching the subjects/domains in which students received the lowest scores (shapes and space, handling and collecting data as well as writing.

As a related point, given that students taught by trained teachers scored significantly better than those taught by untrained teachers, especially if the trained teachers were female, **teacher-training programs should be continued regardless of the deployment issues discussed earlier.**

Also, although teachers' academic qualifications were not associated with better learning outcomes, **measures should be taken to enforce a strict selection process** to ensure that the applicants are sufficiently qualified for courses designed to prepare candidates for teaching.

The gender gap in maths test scores is larger than in other African countries. The research team found a 5.3-percentage-point gap between P6 girls and boys in terms of those who reached minimum competency in maths, and the gap did not narrow between P3 and P6. Thus, it is recommended that:

- **maths teaching and learning materials and practices be revised to be more girl-friendly; and**
- **teacher training include consideration of gender issues during teaching of mathematics.**

As the data show that the proportion of students not reaching minimum reading competencies does not decrease between P3 and P6, with regard to a policy of automatic or collective grade promotion, **teachers' practices should be reformed such that all students achieve a minimum literacy level in the early grades.**

Teacher practices and behaviors can be improved by ensuring that teacher training emphasizes effective time spent on task and more individual attention being given to students. Again, these approaches are made possible by a reasonable student/teacher ratio.

Greater emphasis should be placed on teaching foundational English and mathematics skills and ensuring that all students have mastered these skills. Mastery of these skills will, in turn, help to ensure that students are prepared to successfully acquire more advanced skills taught at higher grade levels.

5.3 School Management

Data show that basic school records such as teacher attendance reports are not kept systematically. While the existence of school management committees and school performance improvement programs were not found to lead to better learning outcomes per se, the data are silent regarding the attributes of the SMC and the content and targets of the SPIP. We should note that we do not have information regarding the implementation of the SPIP nor the SMCs at schools. The effectiveness of these plans and committees can vary greatly from school to school and our data do not allow us to measure the activity of the SMC or the SPIP application. Therefore the fact that we are

not able to link SMC or SPIP to student performance is not surprising and the reader should not conclude that these committee's and plans are not important. However, given that the existence of an SMC or SPIP was found to contribute to better administrative record-keeping, it is recommended that:

- **SMCs and SPIPs should be expanded with more focus on learning outcomes, teacher and student attendance, and effective time on task.**
- **SPIPs should set pedagogical objectives at the school level with regard to basic skills such as reading and counting.**
- **Teacher and student attendance should be monitored more frequently at the school level by greater community involvement and by visits from the circuit supervisor.**

Circuit supervisors can be an avenue for improving learning outcomes but seem to have little effect on students reaching minimum performance levels. Therefore it is recommended that:

- **Circuit supervisor interventions should be more focused on reforming teacher practices to reach the lower-performing students; and**
- **Circuit supervisor interventions should target remote schools.**

Finally, the issue of **teacher absenteeism should be addressed**, as this affects student learning time. This challenge could be approached by involving the teachers unions in discussions about ways to resolve teacher absenteeism.

5.4 Logistics of Test Administration

Earlier sections of this report indicated several specific logistical changes, involving packaging and assembly of materials, which could facilitate future administrations of the NEA.

- More space to count and assemble materials might increase the speed with which packing is completed.
- At the start of the packing session, ASU already had some temporary workers on board who could assist with counting, assembly and packing. However, it became obvious that more staff were needed to assist with the work, and several more people were hired. In future rounds, ASU could hire additional staff for this task before packing begins.
- If space and personnel were available, and if test materials (labels, bags, pencils, etc.) could be purchased earlier, some of the counting and assembly could be completed in advance of actual packing of the bags, and well before the "rush period" for the test administration.

6. Limitations and Recommendations for Future NEA Tests

6.1 Test Forms and Instrumentation

In addition to the limitations summarized above, several other limitations emerged in the 2011 instrumentation.

- Typographical errors were not uncommon in the 2011 assessment in all forms. It was possible to fix some of these errors before data collection, but not all. It cannot be stressed enough how important **quality control** (QC) protocols are when instrumentation is designed for widespread ability assessment. In addition, errors in images presented in reference to specific questions within the assessments may have confused students. Typographical errors lead to misleading questions, multiple correct answers, and situations in which no correct response is present.
- In any ability-measurement situation, test designers must follow stringent criteria to ensure that the **items are constructed appropriately** to target the ability construct of interest, also known as *face validity.* For example, to assess students on their ability to listen to and comprehend stories and instructions, it is critical that the instrument used in the ability-measurement process contain vocabulary, sentence construction, and logic that are grade-appropriate and coherent. Questions about students' knowledge of etiquette and ability to read an analog clock, such as were included in the 2011 NEA, have no bearing on listening comprehension and, indeed, detract from its measurement.
- Passages intended to mea**sure reading comprehension need to** be **clearly written and aligned with the questions that follow.** There was an expectation of inference in the questions without enough context cues in the stories and poems to substantiate the inference. In some cases, the question led the students to infer too much, and in fact all answers could be correct, given the lead in the question. In other passages, errors resulted in no correct answers, even with student inference. This issue was compounded by reading passages that were unclear and did not provide the information the students needed to answer the questions. The expectation of inference was new since the 2009 assessment, where there were clear and specific story-related questions. Without item-level data, the impact of this change cannot be determined. However, even the EdData II education experts had in determining the correct answers to these questions.
- Among test forms, instructions, examples, and questions, **formats need to be consistent.** Varying the instructions for similar items, or omitting an example, may give some students an advantage and put others students at a disadvantage. For example, on one 2011 test form the instructions read "Choose the correct word or group of words that completes the sentence," but on another form,

instructions for the same task were "Choose the correct word that completes the sentence." If the correct answer to the question had multiple words, students receiving the second instruction were at a disadvantage, thinking that it was a "trick question" because the instructions specified a "word." **Instructions** given to students as they begin taking an ability test are critical and must be clear, concise, and followed by several examples to ensure that the students understand what is asked of them. Standardized testing (as in many other national and international assessments) is often artificial and initially difficult to adjust to. The presence of simple, yet thorough, instructions accompanied by examples can greatly reduce initial testing errors that likely have nothing to do with student ability.

- In order to better assess the skill of interest, be it addition or grammar, it is important to **include the appropriate number of items** to measure the skill. As may be seen in Section 2.1, the number of test items designed to measure each of these domain skills was uneven within each set of forms. This was particularly true for the P6 mathematics test forms. Some of the domains—for example, "Basic Operations" in the P6 mathematics test—included an unnecessarily large number of items (22). Conversely, some other domains—such as "Data Collecting and Handling"—had too few items (3). The EdData II research team recommends that each subject test not exceed 45 minutes.
- There should be enough questions (without fatiguing the student) to assess a range of abilities so that students with low ability are distinguished from those with high ability. For example, with an addition construct, a test should have simple single-digit addition (e.g., 2 + 3) and more complex items adding multiple numbers together of two digits or more (e.g., 742 + 92 + 1276), as well as items that are between the two difficulty extremes. In the 2011 P3 maths test forms, addition problems were all fairly complex from the beginning.
- Multiple-choice tests usually contain a **distinct correct answer, surrounded by distracters** that are less correct, or in some cases, obviously incorrect. Distracters are designed to encourage critical thinking, not guessing. In addition, there should not be multiple correct answers. Students should be encouraged to select the best answer, and it should be clear that there is a "best" response. By extension, a correct answer should be present in all questions with multiple-choice response options. Having multiple correct answers causes frustration and fatigue among the students taking the test. Questions in the 2011 NEA test included multiple correct response options and response options that were almost correct, which could confuse students.
- For test items that ask multiple questions in reference to a single passage or prompt, it is a common error to construct one question that is so closely related to another that it "gives away" the answer. This is particularly challenging in reading-comprehension sections, where several questions follow a reading

passage. It is critical that these questions are as independent from one another as possible, given that they all relate to a single reading passage. In other words, it is important to **make sure that students cannot find answers within other questions.**

- Test designers must be careful to make sure that there are **no clues in the response options** that make the correct answer stand out. For example, in a maths question…

  143 + 2 = ?

  a) 1124
  b) 7832
  c) 145
  d) 1101

  Option "c" stands out because it has only three digits, while the rest of the response options contain four digits. In this case, students could select the correct answer without reading the question if they simply chose the one that was different.

- Finally, as has been discussed with the ASU teams, although the P3 and P6 syllabi (curricula) are quite different, the distributions of domain and cognitive abilities measured in the P3 and P6 tests are roughly the same. The research team recommends that the P3 test focus on really basic skills, while the P6 test covers higher cognitive abilities.

6.2 Suggested Revisions to NEA Test Content

Reliable learning outcome trends information could be obtained by including common (anchor) items in the NEA tests. A minimum of 20% common anchor items would provide the information needed for trends analysis. In addition, before finalizing NEA tests, test developers should conduct psychometric analyses of pilot test results to quantitatively ensure the comparability of test content across years. Including Progress in International Reading Literacy Study (PIRLS) and Trends in Mathematics and Science Study (TIMSS) grade 4 items or SACMEQ grade 6 items in the NEA test would also allow researchers to:

- produce international comparisons,
- include more domains such as problem solving or written expression (production of meaningful messages), and
- include open ended questions.

Other recommendations for the NEA instrument are:

- Include open-ended questions in order to assess higher cognitive abilities and to prevent random guessing of answers, especially in P6.

- Design questions that evaluate a single skill so that the results are more easily interpretable.
- Include questions designed to assess student facility with foundational reading and maths skills in both P3 and P6 in order to be able to pinpoint the origins of difficulties students may be facing.
- Develop reading passages for reading comprehension questions that follow a clear sequence of events, with questions that are linked directly to information in the text.
- Consider assessing students' maths skills using mother-tongue questions and instructions rather than English.
- For grade 3, include a simplified oral reading fluency test.
- Include student identification numbers in order to allow linkage of English and maths results at the individual level.
- Include a student questionnaire that collects information such as gender, age, repetition, and kindergarten attendance. Include a head-teacher questionnaire to measure pedagogical supervision and community involvement in school management.
- Include a teacher questionnaire to assess teacher characteristics and practices. Being able to link specific teachers to their students would provide useful information regarding the impact that different teacher characteristics have on student learning.

6.3 Suggested Supplements to EMIS Information: Surveys and Classroom Observations

Teachers' practice in the classroom and effective time on task have been given considerable attention in recent research.[35] Unfortunately, this type of information is not collected via the EMIS. A survey could be conducted in a subsample of NEA schools to obtain additional information to help identify factors that enhance and those that hinder student performance. Data such as time on task, teacher pedagogic practice, school management, and community involvement would be invaluable in guiding future education policy. Survey data could be collected via:

- Direct interviews
- Review of student exercise books
- Collection of exercises
- Classroom observations.

[35] For example, see A. Gove & P. Cvelich. (2010). *Early reading: Igniting education for all. A report by the Early Grade Learning Community of Practice.* Research Triangle Park, NC: Research Triangle Institute.

Annex A: Population Size and Sample Weighting

This annex supplements the information about sample weighting presented in Sections 2.2.2 and 2.6.3. First, it explains how the analysis weights were created for the 2011 NEA. Second, ***Table A1*** at the end of this annex presents weighted proportion estimates and confidence intervals for the actual demographic variables examined in the analysis.

The final two weights used for the analysis of the 2011 NEA were: [$Wt_{Final_{subject/class}}$] and [$Wt_{final_{school}}$]. The first was created at the test level (student-subject-class level) and should be used for analyzing student achievement data. The second was calculated at the school level and should be used for analyzing school level data.

Sample Weights

Because the sample of schools was stratified by region and each school within a region had an equal probability of being sampled, the sample weights were created by simply dividing the total number of primary schools (which had at least ten P6 and ten P3 students) by the number of schools sampled in the region. NALAP pilot schools (found in Volta and Northern region) comprised a separate stratum. Therefore these schools had a sample weight of the total number of NALAP pilot schools divided by the sampled number of NALAP pilot schools.

$$Wt_{Sample} = \frac{Total\ Schools(by\ Region)}{Sampled\ Schools\ (by\ Region)}$$

Rate Propensity at the School Level

To account for nonresponse at the school level, adjustment weights were calculated at the regional level. The adjustment weights were the total number of sampled schools divided by the number of completed schools by region.

$$Wt_{resp\_schools} = \frac{\#Sampled\ Schools\ by\ Region}{\#\ Completed\ Schools\ by\ Region}$$

Rate Propensity Weights at the Student Level

To account for nonresponse at the test form level, adjustment weights were calculated at the P3 and P6 student (class-subject) level. The adjustment weights for P3 maths tests were calculated using the total number of attempted P3 maths tests divided by the number of completed P3 maths test for each school. Weight adjustments for the P3 English tests were calculated by the total number of attempted P3 English tests divided by the number of completed P3 English tests for each school. The response-adjusted weights were calculated for P6 maths and P6 English the same way as for P3 maths and P3 English.

$$Wt_{resp\_test} = \frac{\#Tests\ Attempted\ (by\ School/Class/Subject)}{\#Tests\ Completed\ (by\ School/Class/Subject)}$$

Absentee Weights

To account for absent students during the day of the exam, a second nonresponse weight was calculated at the test form level. These adjustment weights for P3 maths tests were calculated using the total number of P3 students enrolled in the school divided by the total number of attempted P3 maths tests for each school. Weight adjustments for the P3 English tests were calculated by the total number of P3 students enrolled divided by the total number of attempted English tests. The response-adjusted weights were calculated for P6 maths and P6 English the same way as for P3 maths and P3 English.

$$Wt_{abs} = \frac{\#\ Pupils\ Enrolled\ (by\ School/Class)}{\#Tests\ Attempted\ (by\ School/Class)}$$

Non-Scaled Weights at the Test (Class-Subject) Level

The non-scaled weight was calculated at the class-subject level by multiplying the sample weight, school nonresponse adjusted weight, class-subject nonresponse weights, and absent student weights.

$$Wt_{nonscale} = Wt_{Sample} * Wt_{resp\_schools} * Wt_{resp\_test} * Wt_{abs}$$

Final scaled Weight at the Student (Subject-Class) Level

The final subject-class weight was scaled to the regional population of P3 students by dividing the total number of P3 students by the sum of the non-scaled weights at the regional level. This weight should be used for analyzing student achievement data.

$$Wt_{Final_{subject/class}} = \frac{Total\ Pupils\ Enrolled\ (by\ Region/SchoolType/Class)}{\sum Wt_{nonscaled}\ (by\ Region/SchoolType/Class)}$$

Final Weight at the School Level

A final school level weight was applied to the EMIS school-level data. These weights were calculated by multiplying the sample weights by the school response propensity weights. This weight should be used for analyzing data at the school level.

$$Wt_{final_{school}} = Wt_{Sample} * Wt_{resp\_schools}$$

Table A1. Descriptive variables: Weighted proportion estimates of core demographic, by subject

Variable		Subject: Percent estimate[a] (95% confidence interval)[b]			
		P3 math	P3 English	P6 math	P6 English
Urban/rural status	Urban	42.1 (39.9, 44.5)	56.0 (53.5, 58.6)	44.3 (41.8,46.9)	68.6 (65.7,71.3)
	Rural	58.8 (54.3, 63.1)	74 (69.6,77.9)	65.4 (59.9,70.5)	87.6 (84.6,90.2)
Gender	Female	49.1 (48.1, 50.1)	49.7 (48.7, 50.8)	48.7 (47.5, 49.8)	49.1 (47.9, 50.2)
	Male	50.9 (49.9, 51.9)	50.3 (49.2, 51.3)	51.3 (50.2, 52.5)	50.9 (49.8, 52.1)
Region	Ashanti	19.2 (16.5, 22.1)	19.2 (16.5, 22.1)	19.3 (16.1, 22.9)	19.2 (16.1, 22.8)
	Brong Ahafo	10.7 (9.3, 12.3)	10.7 (9.3, 12.2)	10.5 (9.0, 12.2)	10.5 (9.0, 12.2)
	Central	10.4 (8.9, 12.0)	10.3 (8.9, 11.9)	10.4 (8.8, 12.2)	10.4 (8.8, 12.2)
	Eastern	10.5 (8.9, 12.3)	10.5 (8.9, 12.4)	10.6 (8.9, 12.5)	10.5 (8.9, 12.4)
	Greater Accra	11.0 (8.8, 13.6)	11.0 (8.8, 13.7)	12.0 (9.3, 15.3)	11.9 (9.2, 15.2)
	Northern	9.7 (8.0, 11.8)	9.7 (8.0, 11.7)	9.3 (7.5, 11.5)	9.4 (7.6, 11.5)
	Upper East	5.5 (4.8, 6.3)	5.5 (4.8, 6.3)	5.3 (4.6, 6.2)	5.4 (4.6, 6.2)
	Upper West	3.6 (3.1, 4.2)	3.5 (3.0, 4.1)	3.4 (2.8, 4.0)	3.4 (2.8, 4.0)
	Volta	8.2 (6.9, 9.8)	8.3 (7.0, 9.8)	8.2 (6.9, 9.7)	8.4 (7.1, 9.9)
	Western	11.3 (9.7, 13.0)	11.3 (9.7, 13.0)	11.1 (9.4, 13.0)	11.0 (9.4, 12.9)
School type	Public	82.6 (78.4, 86.2)	82.5 (78.2, 86.1)	82.2 (77.6, 86.1)	82.2 (77.6, 86.1)
	Private	17.0 (13.5, 21.3)	17.2 (13.6, 21.5)	17.4 (13.6, 22.1)	17.4 (13.6, 22.1)
	USAID	0.3 (0.3, 0.4)	0.3 (0.3, 0.4)	0.4 (0.3, 0.4)	0.4 (0.3, 0.4)
Test form number	1	26.3 (26.1, 26.5)	25.8 (25.6, 26.1)	26.0 (25.8, 26.3)	26.1 (25.9, 26.3)
	2	25.1 (24.9, 25.3)	25.5 (25.2, 25.8)	25.5 (25.4, 25.7)	25.2 (25.0, 25.4)
	3	24.8 (24.6, 25.0)	24.9 (24.7, 25.1)	24.5 (24.3, 24.8)	24.6 (24.4, 24.8)
	4	23.8 (23.6, 24.0)	23.8 (23.5, 24.0)	23.9 (23.7, 24.1)	24.1 (23.9, 24.3)

[a]Weighted estimates.
[b]95% confidence Intervals were estimated taking the final weights and cluster design into consideration.

Annex B: Numbers of Test Administrator and Monitor Trainees

Table B1. Number of trainees, by home region and gender

Region	Number of trainees		
	Male	Female	Total
Northern	87	4	91
Upper West	59	5	64
Upper East	48	18	66
Volta	85	9	94
Brong Ahafo and Ashanti	149	17	166
Eastern	68	4	72
Central	69	10	79
Greater Accra	41	26	67
Western	63	8	71
Totals	**669**	**101**	**770**

Annex C: Use of EMIS Data During NEA Data Analysis

EMIS Teacher Data

Because students were not linked to a specific teacher, teacher data were aggregated to the school level to help characterize the schools. Before the research team aggregated teacher data to the school level, all variables were checked for consistency and missingness; key variables[36] were imputed using the weight hot deck procedure.[37] Teacher variables were then aggregated to the school level by percent (e.g., percent of qualified teachers), by means (e.g., average age of teachers), or by a ratio (e.g., ratio of teacher books to teachers). These variables were then categorized to increase the cell counts for the logistic regression analysis.

EMIS School Information Data

School information data inherited the nonresponse-adjusted sample weights at the school level. As mentioned previously, school information for only public and USAID schools was obtained from the EMIS; therefore, only USAID and public schools were analyzed using logistical analysis. All school information variables were checked for consistency and for large portions of missing data. School information data were merged with the aggregated teacher data at the school level.

EMIS School Enrollment Data

EMIS enrollment data for P3 and P6 classes were compared to the total number of attempted test forms (i.e., completed tests + noncompleted tests) for consistency. For the 26 schools that were not found in the EMIS 2011 data, EMIS 2010 enrollment data were used, P2 figures were used as a proxy for the current P3 class, and P5 figures were used for the current P6 class. Discrepancies between the EMIS enrollment and NEA test numbers that exceeded ±15% were further scrutinized. Once all inconsistencies were resolved, the enrollment data, school information data, and teacher information (aggregated to the school level) were merged together to produce a final EMIS data set at the school level. These final EMIS school data were then merged with the NEA test information to produce a second final NEA-EMIS data set which was at the test level (i.e., by subject). Both data sets were used in the analysis.

[36] Teacher variables imputed: Teacher Rank, Teacher Type, and Teacher's Academic Achievements.
[37] *SUDAAN Language Manual, Release 10.0*. First Edition. Research Triangle Institute (2008), pp. 813–839.

Table C1. Number of "completed" schools found in EMIS data sets plus number of completed tests found within in those schools

Data type	Completed[a] schools: Number found in EMIS	Completed[b] tests: Number linked to the EMIS data[c]			
		P3 maths	P3 English	P6 maths	P6 English
NEA answer sheets[d]	573	22,780	21,690	20,311	20,117
Enrollment data	573	22,780	21,690	20,311	20,117
Teacher data	565	22,445	21,351	19,947	19,751
School information data	487	19,645	18,632	17,394	17,219
Combined data	481	19,483	18,463	17,187	17,013

[a]Schools were defined as "completed" if completed testing materials from all four test subjects (P3 maths, P3 English, P6 maths, and P6 English) were successfully returned to Accra headquarters, and answer sheets could be successfully scanned.
[b]An answer sheet was defined as "complete" if the grade, subject, and test form were properly marked, and at least five questions were attempted (regardless of whether the answers were correct or not).
[c]Unweighted counts.
[d]The NEA answer sheet data serve as the master file, which contains all 2011 data.

Annex D: Additional Scores and Mean Scores with Confidence Intervals for Student Performance in Maths and English

This annex presents confidence intervals along with the scores previously analyzed in Sections 3.4 and 3.5.

Table D1. Weighted percentages of P3 students achieving minimum competency and proficiency levels, by subject and core demographic variables

Variable		Maths: Percent achieved[a] (95% confidence interval)[b]		English: Percent achieved[a] (95% confidence interval)[b]	
		Minimum competency	Proficiency	Minimum competency	Proficiency
Urban/rural status	Rural	42.1 (39.9, 44.5)	8.7 (7.6, 10)	56.0 (53.5, 58.6)	11.2 (9.4, 13.2)
	Urban	58.8 (54.3, 63.1)	21.6 (16.7, 27.4)	74.0 (69.6, 77.9)	28.9 (22.5, 36.2)
Gender	Male	54.4 (52.1, 56.7)	19.0 (16.5, 21.6)	66.5 (64.2, 68.7)	24.1 (21.1, 27.3)
	Female	50.8 (48.2, 53.4)	17.4 (15, 20.1)	66.2 (63.7, 68.6)	24.4 (21.3, 27.7)
Region	Ashanti	55.4 (48.2, 62.4)	22.1 (15, 31.4)	69.6 (62.9, 75.6)	29.6 (21.2, 39.8)
	Brong Ahafo	51.1 (45.8, 56.4)	12.6 (9.1, 17.3)	65.7 (59.5, 71.4)	17.6 (12.8, 23.6)
	Central	51.3 (45.4, 57.2)	15.7 (11, 21.9)	67 (60.6, 72.9)	21.2 (15, 29.1)
	Eastern	51.9 (44.4, 59.3)	15.7 (10, 23.7)	65.6 (58.2, 72.2)	19.9 (13.4, 28.6)
	Greater Accra	74.4 (68.7, 79.4)	40.3 (32.2, 48.9)	85.6 (80.4, 89.7)	53.7 (43.4, 63.8)
	Northern	41.2 (35.8, 46.8)	11.2 (8, 15.4)	61.5 (55.5, 67.1)	18.5 (12.7, 26.1)
	Upper East	46.4 (40.4, 52.6)	13.3 (9.4, 18.7)	52.9 (45.6, 60)	13.0 (8.2, 20.1)
	Upper West	41.3 (35.9, 46.9)	9.0 (6.6, 12.3)	46.3 (39.6, 53.1)	11.2 (6.9, 17.8)
	Volta	43.9 (38.3, 49.7)	9.8 (6.8, 13.8)	53.4 (46.8, 59.9)	13.0 (8.4, 19.6)
	Western	52.8 (46.2, 59.4)	17.4 (12.7, 23.2)	69 (62.7, 74.6)	21.9 (15.6, 29.7)
School type	Public	47.4 (45.1, 49.8)	12.9 (10.8, 15.2)	61.7 (59.2, 64.1)	16.8 (14, 19.9)
	Private	77.8 (73.2, 81.8)	44.2 (37.1, 51.5)	88.7 (85.3, 91.3)	60.1 (53.4, 66.5)
	NALAP pilot schools	46.3 (42.4, 50.2)	10.6 (9, 12.4)	65.8 (61.3, 70)	15.9 (13.9, 18.2)
Test form number	1	48.3 (45.7, 50.9)	15.8 (13.5, 18.4)	68.7 (66.3, 71.1)	27.2 (24.2, 30.4)
	2	54.2 (51.7, 56.7)	18.4 (16, 21)	61.8 (59.2, 64.2)	19.8 (16.9, 23.1)
	3	53.2 (50.4, 56)	18.5 (15.9, 21.3)	66.4 (63.7, 68.9)	24.8 (21.5, 28.5)
	4	55.1 (52.4, 57.7)	20.4 (17.7, 23.3)	68.5 (66, 71)	25.1 (22, 28.5)

[a]Weighted proportions.
[b]95% confidence Intervals were estimated using Taylor linear series to account for the cluster effect of the sample design.

Table D2. Weighted percentages of P6 students achieving minimum competency and proficiency levels, by subject and core demographic variables

Variable		Maths: Percent achieved[a] (95% confidence interval)[b]		English: Percent achieved[a] (95% confidence interval)[b]	
		Minimum competency	Proficiency	Minimum competency	Proficiency
Urban/rural status	Rural	44.3 (41.8, 46.9)	6.0 (5.1, 7.1)	68.6 (65.7, 71.3)	17.4 (15.1, 20)
	Urban	65.4 (59.9, 70.5)	20.7 (14.2, 29.2)	87.6 (84.6, 90.2)	45.5 (37.9, 53.2)
Gender	Male	59.4 (56.7, 62)	17.7 (15, 20.9)	79.4 (77.3, 81.4)	35.0 (31.6, 38.6)
	Female	54.5 (51.5, 57.4)	14.3 (11.4, 17.9)	78.4 (76.1, 80.5)	35.6 (31.8, 39.6)
Region	Ashanti	62.4 (55.9, 68.4)	20.2 (13.3, 29.5)	84.7 (79.2, 88.9)	43.1 (33.8, 52.9)
	Brong Ahafo	51.7 (45.3, 58)	9.9 (6.3, 15.3)	72.2 (65.8, 77.8)	26.4 (19.1, 35.3)
	Central	51.1 (45.1, 57)	11.1 (6.5, 18.3)	77.6 (71.7, 82.5)	27.7 (20.8, 35.9)
	Eastern	58.6 (50.4, 66.4)	14.1 (10, 19.5)	76.9 (70, 82.7)	34.3 (25.5, 44.4)
	Greater Accra	81.4 (74.1, 87)	38.2 (26.5, 51.4)	96.5 (94.5, 97.7)	73.9 (64.6, 81.4)
	Northern	35.7 (30.4, 41.3)	3.8 (2.3, 6.1)	66.6 (59.4, 73.2)	17.2 (13, 22.5)
	Upper East	49.7 (43.7, 55.6)	9.3 (6, 14.2)	68.6 (62.3, 74.3)	19.5 (13.9, 26.7)
	Upper West	48.1 (43.3, 52.9)	8.1 (6.1, 10.6)	70.1 (64.3, 75.4)	15.8 (12.3, 20.2)
	Volta	50.7 (44.1, 57.2)	11.9 (7.5, 18.2)	75.9 (70.6, 80.5)	26.8 (20.4, 34.4)
	Western	58.8 (50.6, 66.5)	16.4 (9.6, 26.8)	79.6 (72.7, 85.1)	31.9 (22.1, 43.6)
School type	Public	51.6 (48.7, 54.6)	11.2 (8.4, 14.7)	75.2 (72.8, 77.5)	27.1 (23.5, 31.1)
	Private	82.1 (78.1, 85.6)	39.3 (32.1, 47)	96 (93.2, 97.7)	73.8 (67.5, 79.2)
	NALAP pilot schools	51.6 (48, 55.1)	8.7 (7.3, 10.2)	82.4 (79.8, 84.7)	28.9 (25.8, 32.2)

Table D3. P3 English results, by domain: Means, standard errors, and confidence intervals

Domain	Mean	Standard error	(95% confidence interval)	
Proficiency_Level_Writing	22.4%	1.2%	20.0%	24.9%
Min_Level_Writing	39.4%	1.2%	37.1%	41.6%
Proficiency_Level_Reading	22.2%	1.4%	19.4%	25.0%
Min_Level_Reading	52.3%	1.3%	18.3%	23.5%
Proficiency_Level_Grammar	20.9%	1.2%	33.6%	38.4%
Min_Level_Grammar	36.0%	0.9%	55.2%	58.8%
Proficiency_Level_Listening	64.2%	1.3%	61.7%	66.7%
Min_Level_Listening	86.1%	0.6%	84.9%	87.4%

Table D4. P3 maths results, by domain: Means, standard errors, and confidence intervals

Domain	Mean	Standard error	(95% confidence interval)	
Proficiency_Level_Data	9.3%	0.8%	7.8%	10.8%
Min._Level_Data	24.9%	1.0%	23.0%	26.9%
Proficiency_Level_Shape&Space	24.0%	1.1%	21.8%	26.2%
Min._Level_Shape&Space	50.4%	1.0%	48.4%	52.4%
Proficiency_Level_measurement	21.4%	1.2%	19.0%	23.8%
Min._Level_measurement	51.2%	1.1%	49.1%	53.3%
Proficiency_Level_operations	23.7%	1.1%	21.5%	25.8%
Min._Level_operations	61.0%	0.9%	59.2%	62.9%
Proficiency_Level_numbers	31.4%	1.1%	29.3%	33.5%
Min._Level_numbers	51.1%	1.0%	49.2%	53.1%

Table D5. P6 English results, by domain: Means, standard errors, and confidence intervals

Domain	Mean	Standard error	(95% confidence interval)	
Proficiency_Level_Writing	18.8%	1.3%	16.4%	21.3%
Min_Level_Writing	34.0%	1.4%	31.2%	36.7%
Proficiency_Level_Reading	36.7%	1.6%	33.6%	39.8%
Min_Level_Reading	51.5%	1.3%	48.9%	54.1%
Proficiency_Level_Grammar	33.5%	1.8%	30.0%	37.0%
Min_Level_Grammar	55.5%	1.5%	52.6%	58.4%
Proficiency_Level_Listening	78.1%	1.0%	76.1%	80.2%
Min_Level_Listening	92.6%	0.5%	91.6%	93.5%

Table D6. P6 maths results, by domain: Means, standard errors, and confidence intervals

Domain	Mean	Standard error	(95% confidence interval)	
Proficiency level_Data	42.0%	1.1%	39.9%	44.1%
Minimum level_Data	42.0%	1.1%	39.9%	44.1%
Proficiency level_Shapes and Space & Measurement	25.0%	1.3%	22.4%	27.6%
Minimum level_Shapes and Space & Measurement	49.1%	1.1%	46.8%	51.3%
Proficiency level_Operations	17.6%	1.4%	14.9%	20.4%
Minimum level_Operations	67.1%	1.0%	65.3%	69.0%
Proficiency level_Numbers	32.8%	1.3%	30.2%	35.4%
Minimum level_Numbers	56.2%	1.1%	54.0%	58.3%

Table D7. Average test scores, P3 English, by domain: Means, standard errors, and confidence intervals

Domain	Mean	Standard error	(95% confidence interval)	
Listening	63.6%	0.8%	62.2%	65.1%
Grammar	35.1%	0.8%	33.6%	36.6%
Reading	40.6%	0.9%	38.9%	42.2%
Writing	33.9%	0.8%	32.3%	35.4%

Table D8. Average test scores, P3 maths, by domain: Means, standard errors, and confidence intervals

Domain	Mean	Standard error	(95% confidence interval)	
Numbers	39.3%	0.7%	38.0%	40.6%
Operations	42.9%	0.6%	41.6%	44.1%
Measurement	38.8%	0.7%	37.4%	40.1%
Shapes and Space	41.2%	0.8%	39.6%	42.9%
Collecting and Handing Data	29.0%	0.5%	28.0%	30.1%

Table D9. Average test scores, P6 English, by domain: Means, standard errors, and confidence intervals

Domain	Mean	Standard error	(95% confidence interval)	
Listening	72.8%	0.7%	71.4%	74.2%
Grammar	44.1%	0.9%	42.2%	45.9%
Reading	44.1%	1.0%	42.2%	46.0%
Writing	32.4%	0.7%	31.1%	33.7%

Table D10. Average test scores, P6 maths, by domain: Means, standard errors, and confidence intervals

Domain	Mean	Standard error	95% confidence interval	
Numbers	40.7%	0.7%	39.4%	42.0%
Operations	42.4%	0.7%	41.1%	43.7%
Shapes and Space & Measurement	37.3%	0.7%	35.8%	38.7%
Collecting and Handling Data	44.0%	0.7%	42.6%	45.4%

Annex E: Students' Probability of Achieving Minimum Competency and Proficiency Scores

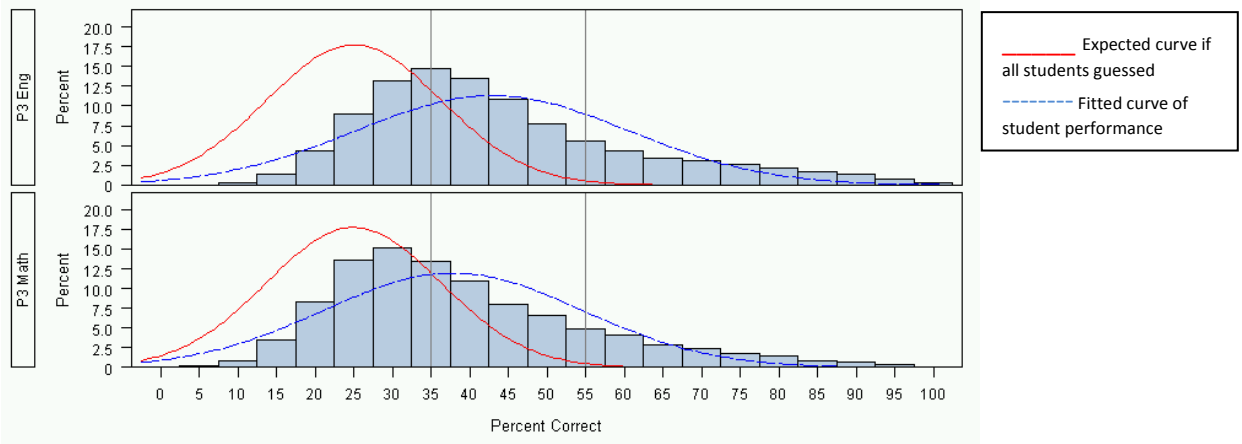
This annex further elaborates some probability issues raised in Section 3.1.

P3 Scores and Achievements

As noted earlier, the P3 maths and English exams each consisted of 40 questions. Based on probability statistics, if all P3 students were to guess every question by randomly selecting one of the four possible responses, one would expect a mean score of 25% or 10 correct answers. It also would be expected that 41.5% of the students would correctly answer at least 10 questions. P3 students had to correctly answer at least 14 questions to obtain minimum competency (35%) and 22 questions (55%) to reach proficiency. If P3 students were to guess all 40 questions at random, 5.4% should obtain minimum competency and 0.001% should reach proficiency.

Figure E1 depicts the weighted histogram of the P3 English and maths scores along with a normal distribution curve that best fits the histogram (dashed line) as well as the normal distribution if the students had randomly guessed all questions (solid line). The dashed curved line, which stands to the right of the curved line, indicates that students scored much better than had they randomly guessed.

Figure E1. Score distribution of P3 students



Note: Scores are weighted.

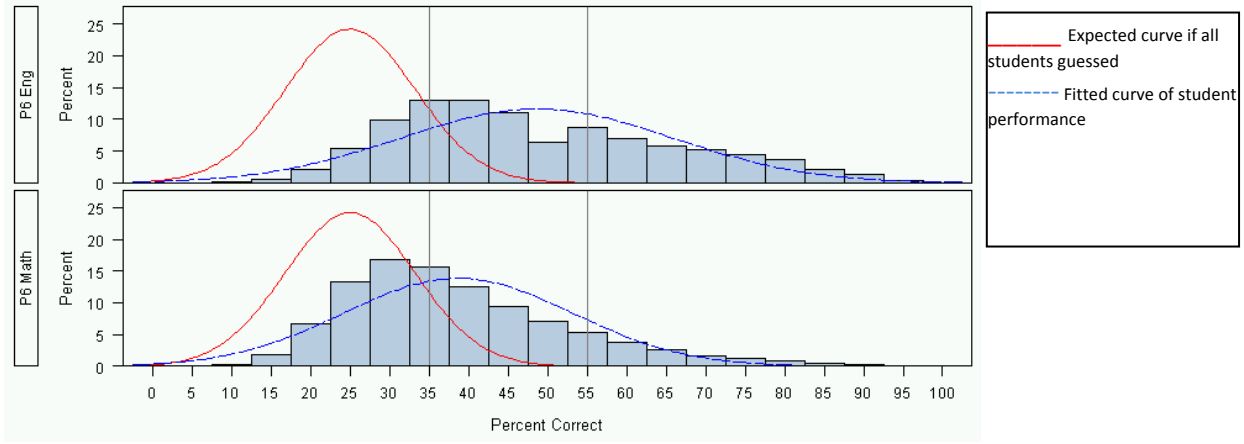
P6 Scores and Achievements

As noted previously, the P6 maths and English exams each consisted of 60 questions. If all P6 students were to guess every question, a mean score of 25%, or 15 correct answers, could be expected. If this were to happen, 43.1% of the P6 students should answer at least 15 correctly. P6 students had to correctly answer at least 21 questions to reach minimum

competency and 33 questions to reach proficiency. If all P6 students were to guess all 60 questions at random, one would expect 3.0% to reach minimum competency and only 1 student in 10 million to reach minimum competency.[38]

Figure E2 depicts the weighted histogram of the P6 English and maths scores along with a normal distribution curve that best fits the histogram (dashed line), as well as the normal distribution if the students had randomly guessed all questions (solid line). Similar to the P3 curves seen above, the dashed curved line indicates that students scored much better than if they had randomly guessed. It can also be seen that the curved dashed line in the English test is further to the right, indicating that P6 students performed better in English than in maths.

Figure E2. Score distribution of P6 students



Note: Scores are weighted.

[38] Estimates based on a binary distribution (n = 60, p = 0.25).

Annex F: Explanation of Trend Data Requirements

As mentioned in Section 3.3, in order to properly evaluate trends in student performance over time, historic item-level and item content data are needed. An explanation of why this information is needed is presented here.

The first requirement involves ***item-level data*** (i.e., question-by-question responses). "Same items"—i.e., identical or nearly identical test questions—should appear in each consecutive testing situation. They typically include a set of items that is the same across *all* testing situations.

In addition, ideally, these item-level data should exist for every year in which data were collected. These data exist for 2007 and 2011, but for 2005 and 2009, only total score data are available.

Second—again, in an ideal situation—***item content information*** (i.e., copies of test booklets with the examination questions) should be retained for all data collection years. Item content documentation is critical to establish *equating items*, or items that cover the same material. Without this documentation, comparing across years is impossible because researchers will not know, for example, if they are comparing two addition items, or an addition item with a geometry item. For the current analysis the researchers were able to obtain this item content information only for 2009 and 2011. Without the item-level data from other years to examine, these test booklets are not as useful.

Annex G: Item-Level Evaluation

Summary

As explained in Section 2.1, there were four forms (versions) of the NEA for each subject in both P3 and P6. Thus, for example, Primary 3 students were administered one test from the set of English Forms 1 through 4, and one from the set of mathematics Forms 1 through 4. Further, the forms contained the following numbers of items:

- P3 English – 10 listening comprehension plus 30 read/write
- P3 mathematics – 40
- P6 English – 15 listening comprehension plus 45 read/write
- P6 mathematics – 60

As part of the EdData II assessment, the NEA instrumentation underwent a psychometric evaluation using a method called the *Rasch model.* All the items within each subject (mathematics and English) and form (Forms 1 through 4) were assessed individually.

The Rasch model generates several types of information that can tell exam designers how well their instruments are measuring the skills they want to assess within their target population. For example, it can map test-takers' distribution along a continuum from very low to very high ability levels, based on their total scores; determine the difficulty levels of each item in every instrument; assess whether too many questions are testing at the same difficulty level; and show in aggregate how well the test-takers did on each item. Findings were consistent and comparable across the instruments by subject area and for both grade levels. For example:

- In every case, the analysis of the forms showed "normal" distributions of students among the ability levels. This means that as desired, higher numbers of students were clustered at the average ability levels in the middle, tapering down to a few students at the highest and lowest ability levels and creating a bell-shaped curve.
- For mathematics, for all four forms in both P3 and P6, the mean values for item difficulty were higher than the means for student ability. In other words, the entire form was more difficult than ideal for the target population. For English, these two mean values were generally much closer and in some cases were perfectly aligned, indicating tests that were better tailored to their audience.
- Every form in the assessment appeared to have too many items clustered at the same levels of difficulty. Packing an assessment with too many questions that test the same information or at the same difficulty level can fatigue the test-takers, and is an inefficient way to collect information about students' performance. Assessments can be streamlined by removing such redundancies. The research

team also found, however, that some items clustered at the same difficulty level were useful because they were testing quite different skills (for example, a fraction question based on an image, versus a division question). Thus, if redundant items were to be removed from future assessments, care would need to be exercised in comparing the skills they covered.

- Every form had some items that produced "unexpected responses" given the relative difficulty of the item. As an example, a surprising number of students at the middle and low end of the ability distribution might have gotten the item correct, while at the same time, some high-ability students responded incorrectly. Sometimes closer examination of the unexpected-response item revealed no particularly striking problem with the question. On the other hand, some of the items had clearly confused the students because of the way in which either the question or the answer choices were constructed. Many of these problems are described in Section 8 on limitations of the instruments.
- Form 1 of Primary 6 English was the most difficult for the test-takers, as indicated by the high number of items showing unexpected responses (five).

More detailed item-level analyses, by form, are presented further below.

Primer: How to Read Outputs of the Rasch Model

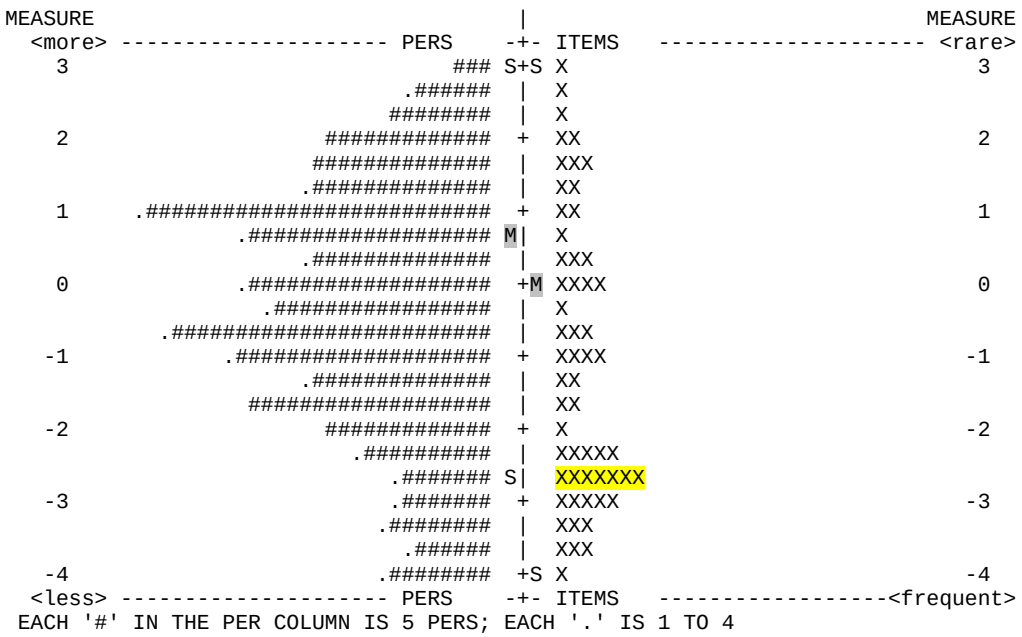
The example Wright plot below ***(Figure G1)*** is an illustration of the interval-level properties enforced on the data by the Rasch model that allow items to be examined in relation to the persons taking those items. The "Measure" axis provides the logit scale on which both persons and items are placed. Students are depicted on the left side ("#" marker) of the chart and the items are shown on the right ("X" marker). This plot can be seen as an ability continuum (for persons), and difficulty continuum (for items), with more able persons as well as more difficult items positioned near the top of the chart. In an ideal instrumentation situation, a normal distribution of individuals would appear on the left with a flat spread of one to two items for each person ability level. The item and person means ("M" highlighted in gray) are expected to be "targeted," or in close alignment. The "S" represents the designation of one standard deviation from the mean.

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

Finally, discussions of "unexpected responses" refer to items that are not eliciting the responses expected, given the relative difficulty of the item. Usually this is an outlier

issue, but it can also refer to high-ability people responding incorrectly to an easy item, or to lower-ability respondents getting a very difficult item correct.

Figure G1. Sample Wright plot of "persons" vs. test items



Item-Level Analyses by Test Form, Mathematics

Turning now to actual 2011 NEA data, the person and item targeting for **Form 1 of P3 mathematics *(Figure G2)*** is almost a standard deviation off, with the item-difficulty mean displaying higher than the person-ability mean (see shaded "M's"). This finding indicates that the P3 Form 1 was more difficult than ideal for the target population. The distribution of persons on the left is relatively normal (Gaussian), with no ceiling or floor effects (only two students with maximum scores and one student with a minimum score). However, the spread of items is rather tightly compacted, with quite a bit of measurement redundancy (yellow highlight).

Ideally, there should be a normal distribution of persons on the left, with one or two items at each level of person ability. With the current level of measurement redundancy (yellow highlight) students may be becoming fatigued; with no new information being provided about ability, the redundant items could potentially be removed. However, if there are items that require the same level of effort (i.e., same ability level) but tap different skills, the items may need to be kept to provide a well-rounded picture of ability. For example, in Figure G2, Item 7 (code: 7=C7) and Item 11 (code: 11=C11) had approximately the same item difficulty estimate, but Item 7 was an addition question, whereas Item 11 was

a measurement question. Therefore, it is likely that both items were useful, even though they produced the same level of effort.

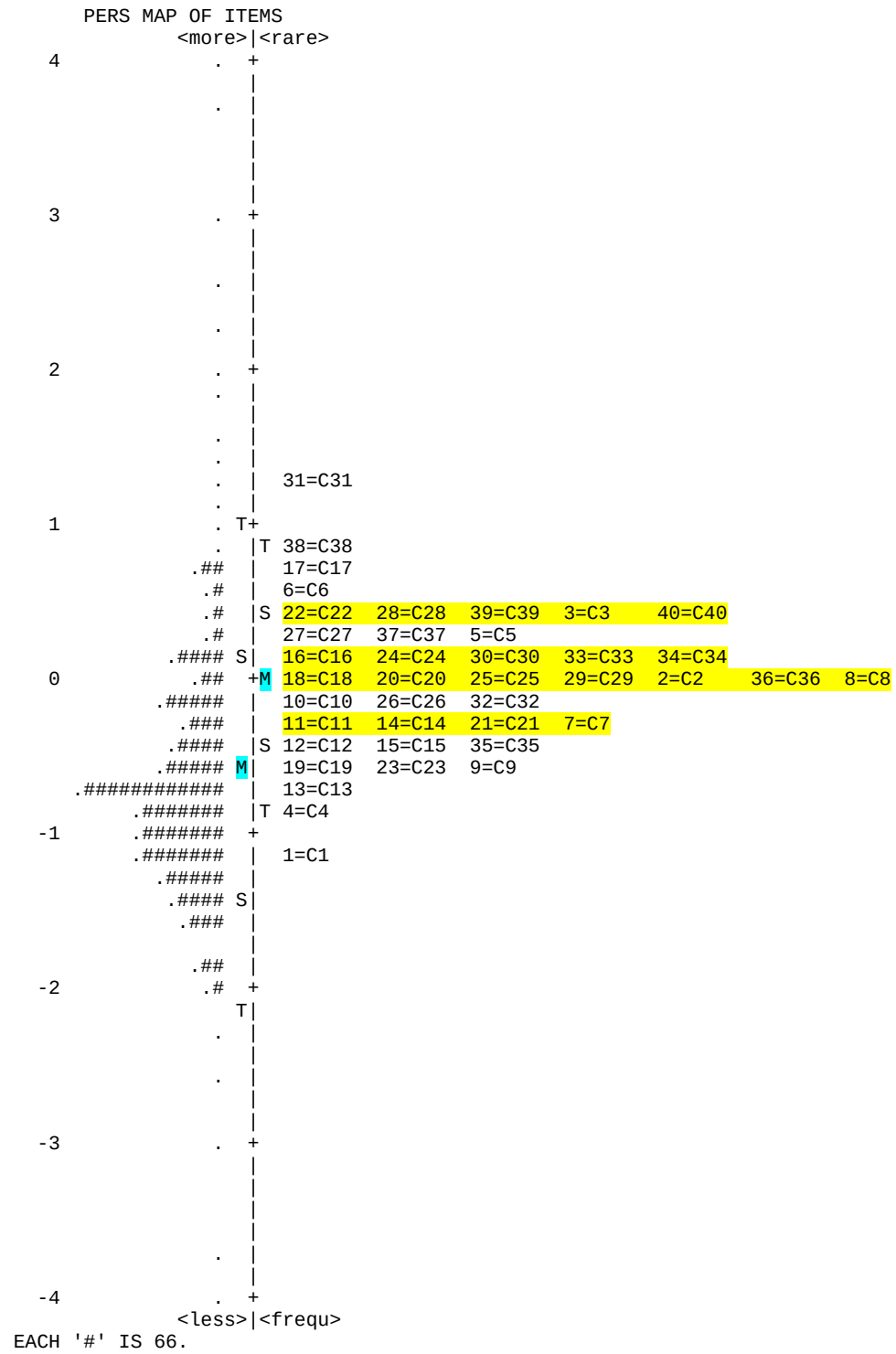
Finally, one item (Item 31) shows unexpected responses and may need to be more closely examined. In the Wright plot, the most difficult item should appear at the top of the item-difficulty distribution, on the right side). However, Figure G2 shows a surprising number of students at the middle of the ability distribution getting Item 31 correct. Looking at item content,[39] there is no apparent aspect of the item that would cause the unexpected responses.

[39] "A boy has 10 pencils. If 5 of the pencils are red, what fraction of the ten pencils is red? A. 1/6 B. 1/5 C. 1/4 D. 1/2"

Figure G2. Wright plot for NEA results, P3 mathematics, Form 1

Primary 3 – Overall Cronbach's α = 0.82

Form 1 – Cronbach's α = 0.80

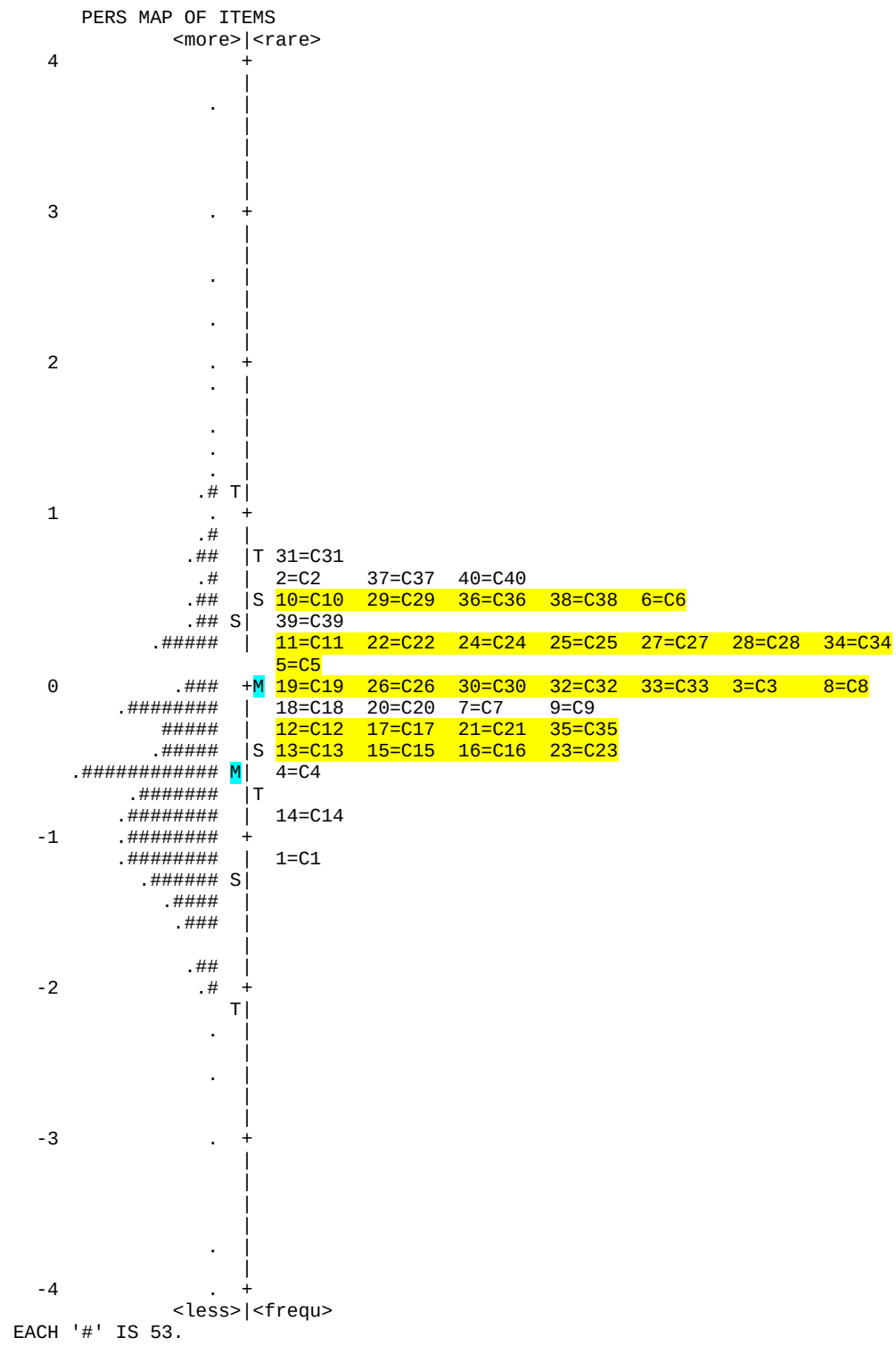


This pattern precisely repeated itself in all three of the remaining test versions for P3 mathematics—Forms 2 through 4—as well as in all four P6 test forms for mathematics. In each case, the person and item targeting was a standard deviation off (turquoise highlighting), indicating a test likely too difficult for the target population; the distribution of persons on the left was relatively normal, with only one or two students receiving a minimum score; and the spread of items was compacted, with substantial item redundancy (yellow highlighting). Thus, in the discussion that follows, readers should assume these features are relevant in addition to the text associated with the figures below, which further addresses form-specific information that is of interest.

For **P3 mathematics Form 2** (see ***Figure G3***), an example of ability-level redundancy that may *not* have been an issue concerns Item 13 and Item 23. They have approximately the same item difficulty estimate, but Item 13 was a fraction question with an image and Item 23 was a division question. Therefore, it is likely that both items were useful.

Figure G3. Wright plot for NEA results, P3 mathematics, Form 2

Form 2 – Cronbach's α = 0.81

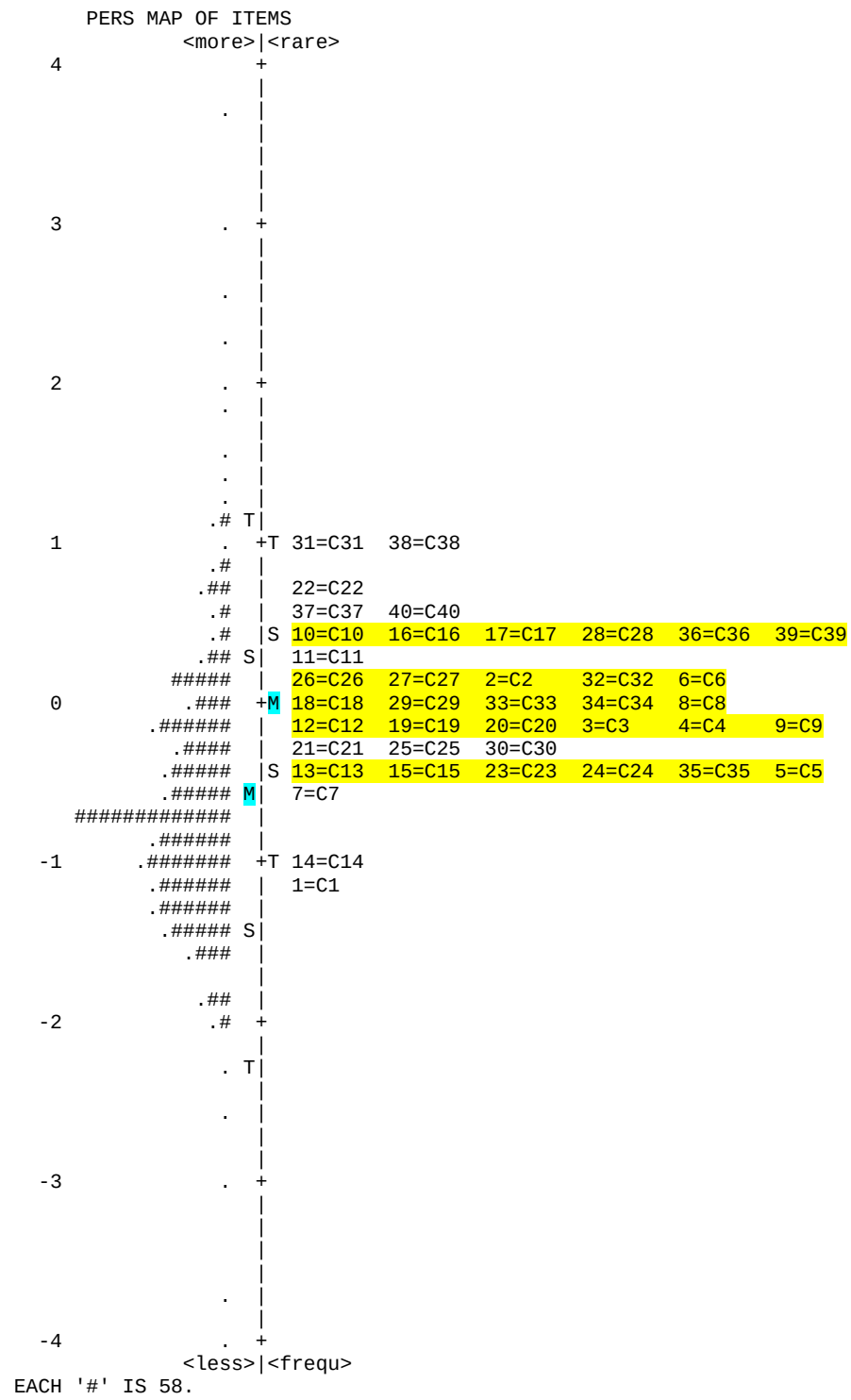


For **P3 mathematics Form 3,** although the problem of too-frequent measurement at certain skill levels again was a problem (yellow highlight), there also were a few items for which this apparent redundancy was logical. For example, ***Figure G4*** indicates that Item 5 and Item 13 have approximately the same item difficulty estimate. However, Item 5 was a quantity comparison question whereas Item 13 was a fraction question with an image. Similar to occurrences in the other test forms, then, in this case it is likely that both items were useful.

Form 3 also contained one item (Item 38 in Figure G4) that shows unexpected responses and may need to be more closely examined. A surprising number of students at the middle and low end of the ability distribution got the item correct, and in addition, some high ability students responded incorrectly. Looking at content—an addition problem that asked students to combine several values of coins—reveals no immediately apparent aspect of the item that would cause the unexpected responses.

Figure G4. Wright plot for NEA results, P3 mathematics, Form 3

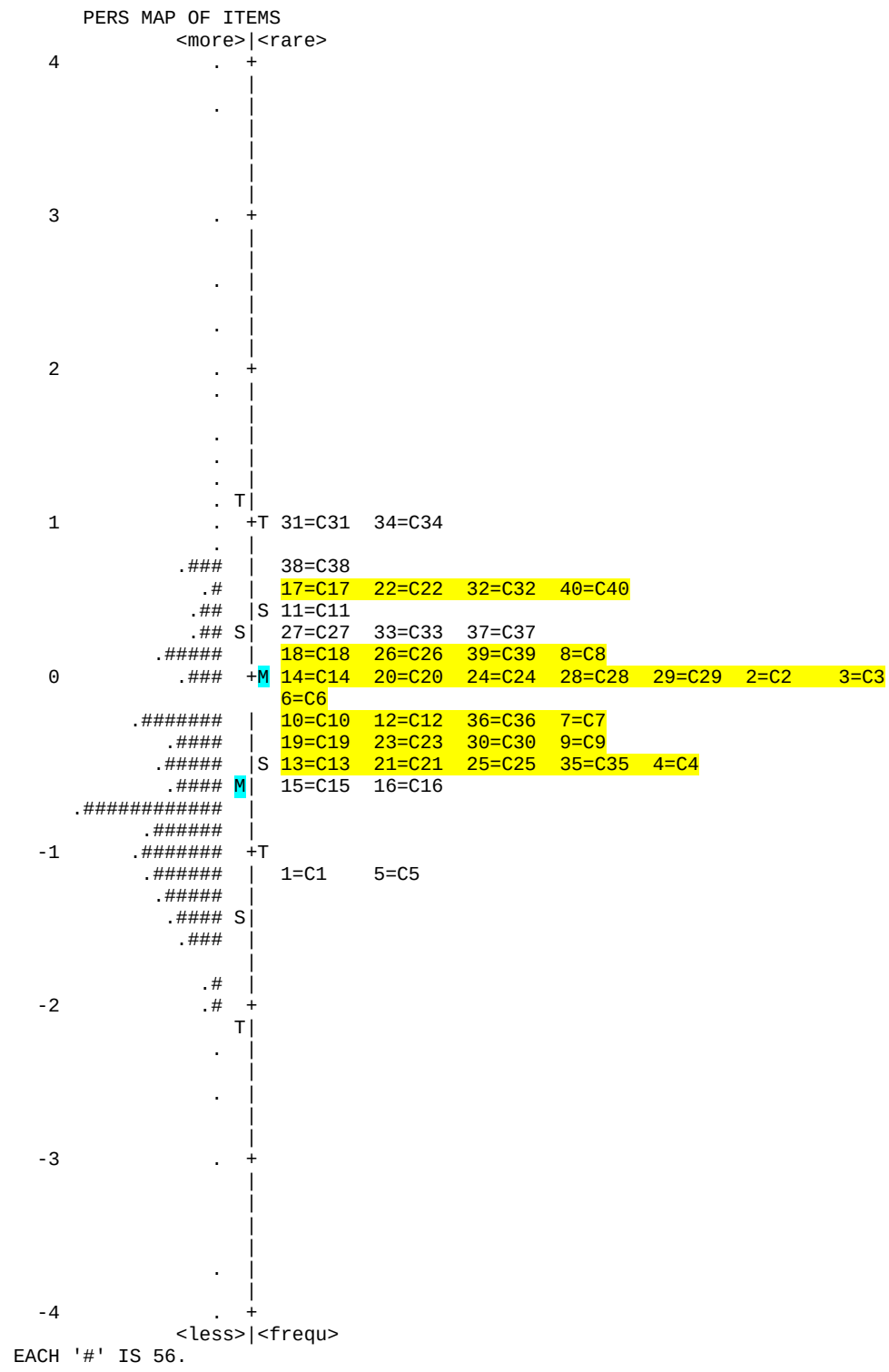
Form 3 – Cronbach's α = 0.82



Form 4 of Primary 3 mathematics also was subject to quite a bit of measurement redundancy (yellow highlight), but with some cases that could be seen as reasonable. ***Figure G5*** shows that Item 4 and Item 13, for example, have approximately the same item difficulty estimate. However, Item 4 was a number line question while Item 13 was a fraction question with an image. Therefore, it is likely that both items were useful.

Figure G5. Wright plot for NEA results, P3 mathematics, Form 4

Form 4 – Cronbach's α = 0.82



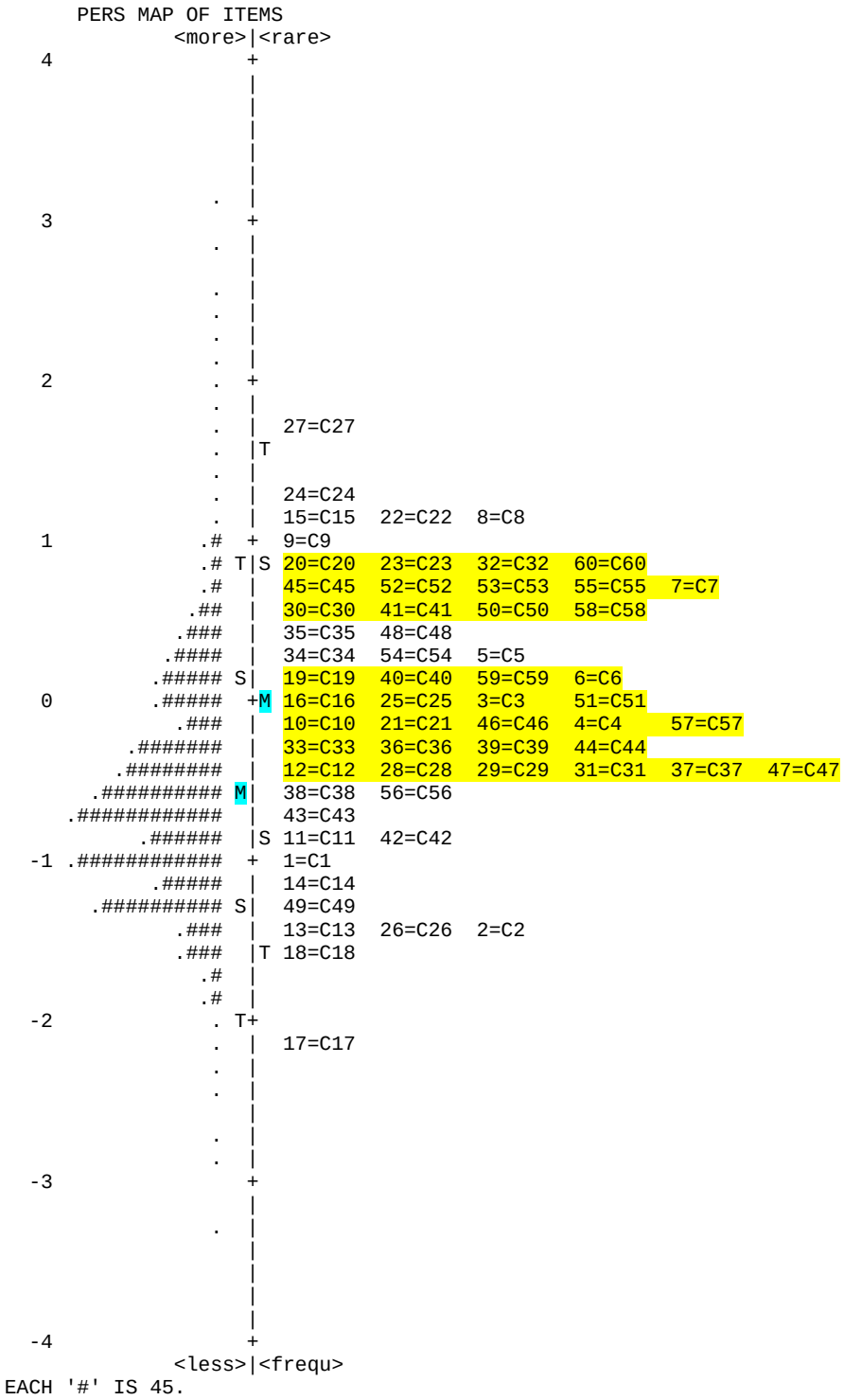
For **Form 1 of Primary 6 mathematics, *Figure G6*** indicates Item 12 and Item 47 as one of the examples of pairs of items with approximately the same item difficulty estimate, but Item 12 was a numerical value question whereas Item 47 was a compounding shape question with an image. Thus, keeping these two items could be useful even though the difficulty level was the same. In most cases, however, the yellow highlighting indicates over-measurement at those skill levels.

Figure G6 also contains two items showing unexpected responses—Item 15 and Item 27—to which a surprising number of students at the middle and low end of the ability distribution responded correctly. Item 15 asked students to arrange a series of three decimal fractions in ascending order. Looking at item content, there is no immediately apparent aspect of Item 15 that would cause the unexpected responses. However, for Item 27—a word problem that required students to review three different weight measurements and find their average—there is a possibility that the close distracter response of 44.3 was confusing to participants because the correct option was 44.1.

Figure G6. Wright plot for NEA results, P6 mathematics, Form 1

Primary 6 – Overall Cronbach's α = 0.84

Form 1 – Cronbach's α = 0.84



As with the other mathematics forms, the analysis revealed that **P6 mathematics Form 2** had both significant amounts of measurement redundancy (yellow highlighting in ***Figure G7***) and a few apparent item redundancies that on closer inspection may not have been problematic. An example is Item 3 and Item 17, which have approximately the same item difficulty estimate, but Item 3 was a lowest-common-multiple question whereas Item 17 was a decimal subtraction question.

Also Item 27 in Form 2 was exactly the same weight-averaging question that appeared in Form 1, in which the close distracter response of 44.3 may have confused participants (the correct option was 44.1).

Figure G7. Wright plot for NEA results, P6 mathematics, Form 2

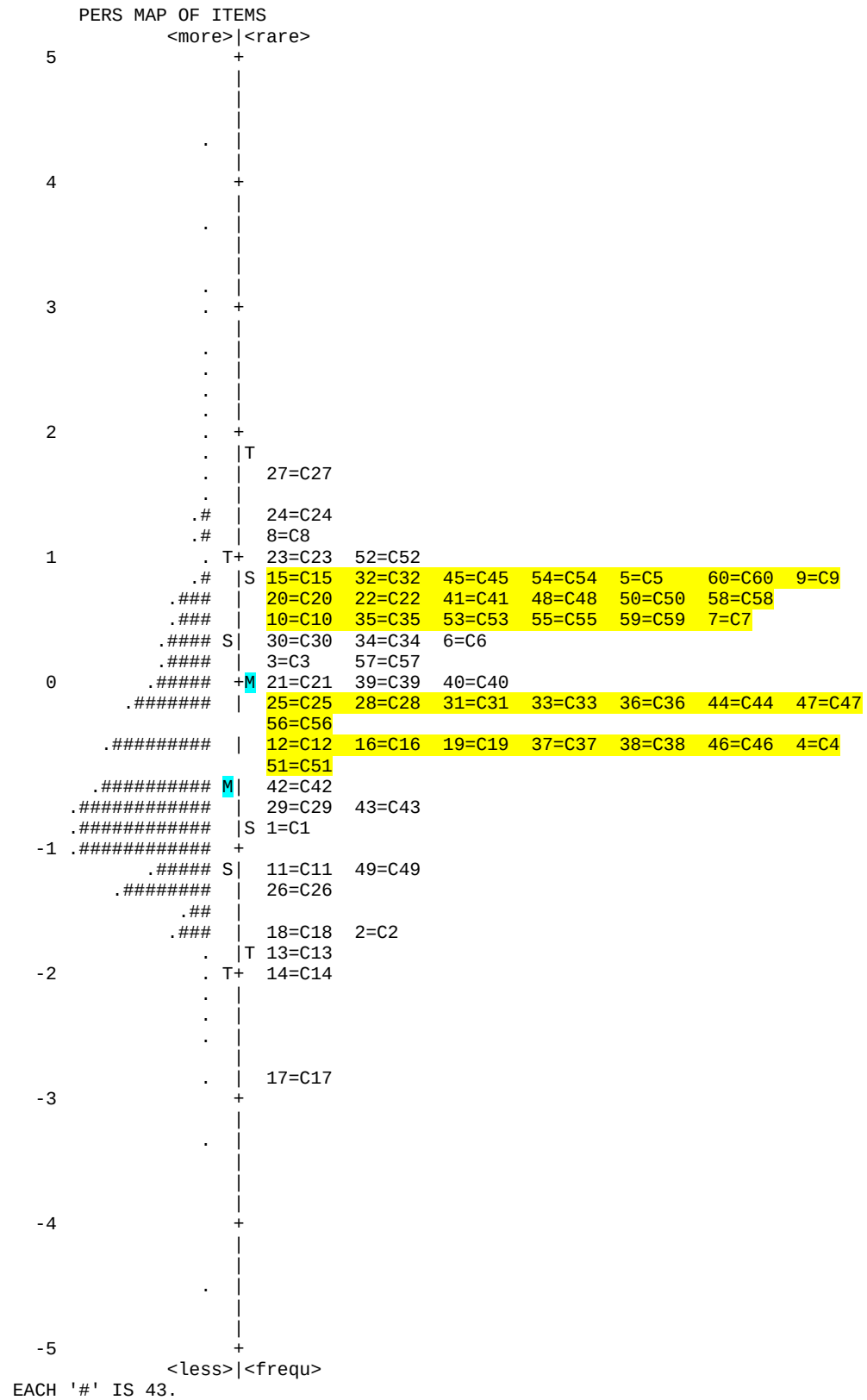
Form 2 – Cronbach's α = 0.85

Turning to **P6 mathematics Form 3,** the measurement redundancies again are clear from the yellow highlighting in ***Figure G8.*** On the other hand, Item 4 and Item 51 have approximately the same item difficulty estimate, but this is another case in which both items may nevertheless have been useful. Item 4 was a common factor question ("Find the Highest Common Factor [HCF] of 12 and 15") whereas Item 51 gave a list of items and prices in a shop, and required students to add the costs of several of the items.

Item 27 in Form 3 (averaging of weights) was identical to the Item 27 in Forms 1 and 2, and displayed the same problems with unexpected responses; this same item also recurred in Form 4, as mentioned below.

Figure G8. Wright plot for NEA results, P6 mathematics, Form 3

Form 3 – Cronbach's α = 0.85

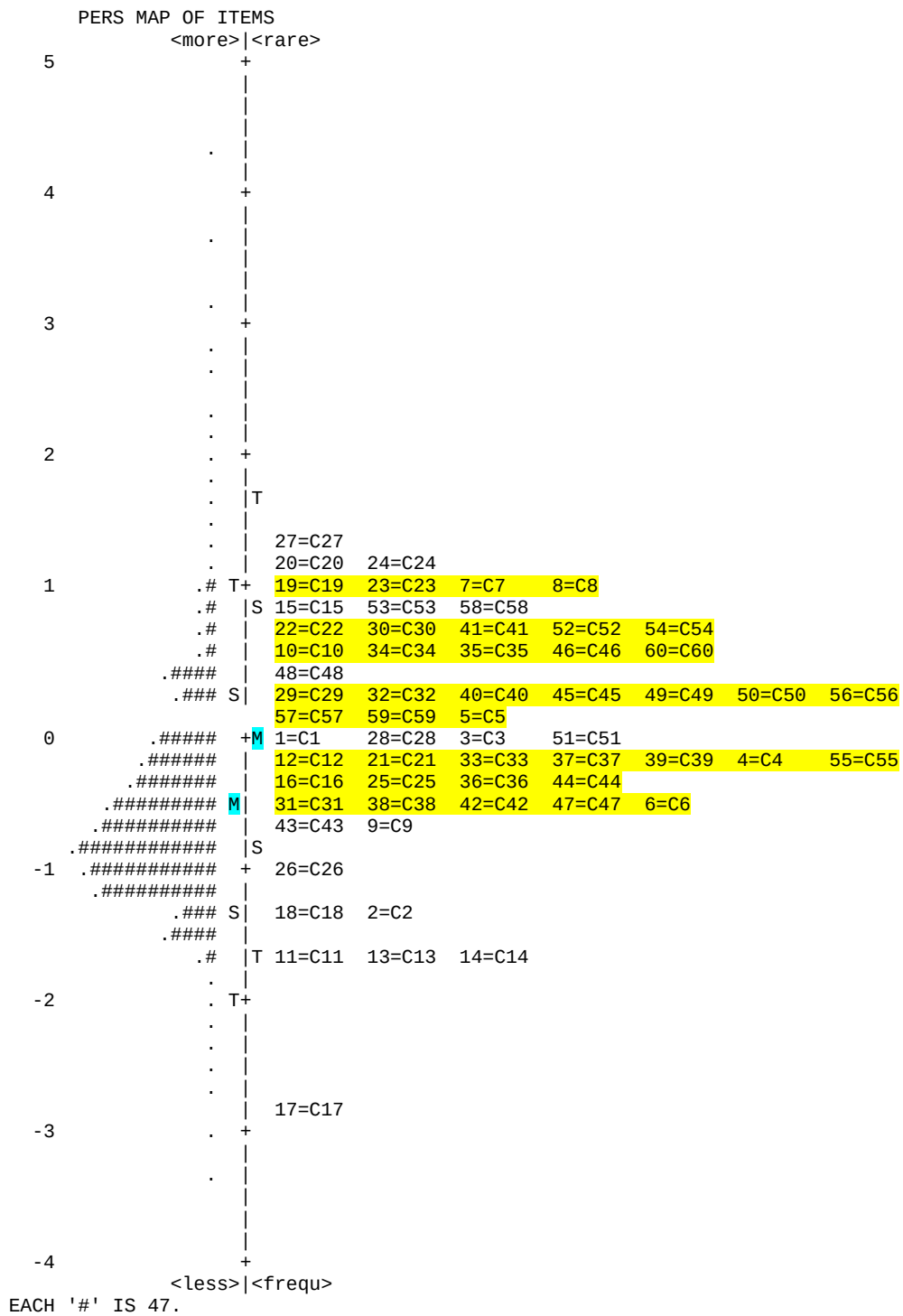


As with the other test forms, **Form 4 of Primary 6 mathematics** suffered from substantial measurement redundancy (yellow highlighting in ***Figure G9***), but also contained some items that may need to be kept to provide a well-rounded picture of ability. An example is Item 6, a fraction ordering question, which ended up at the same difficulty level as Item 31, a geometry question.

As noted earlier, the same weight-averaging problem that caused difficulties in Forms 1 through 3 also appeared in Form 4 (Item 27), with the same effect in terms of possible confusion among test-takers.

Figure G9. Wright plot for NEA results, P6 mathematics, Form 4

Form 4 – Cronbach's α = 0.84



Item-Level Analyses by Test Form, English

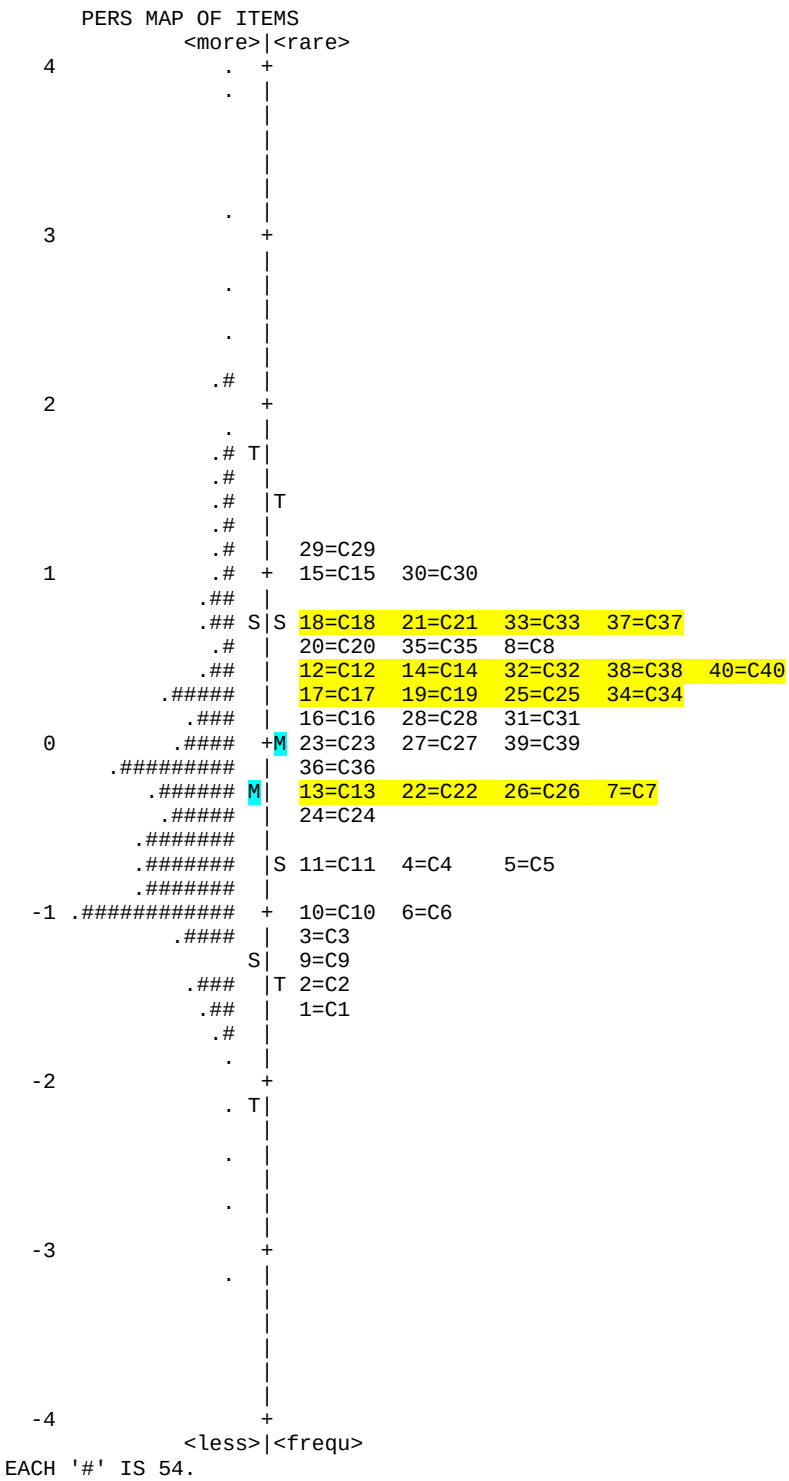
The person and item targeting (turquoise highlight) for **Form 1 of Primary 3 English** is close (see ***Figure G10***), with the item-difficulty mean only slightly above that of the person-ability mean, indicating that the test form was pretty well aligned to the target population. The distribution of persons on the left is relatively normal, with no ceiling or floor effects (only six students with maximum scores). The spread of items is fairly well distributed, with needs really centered in the upper and lower ends of the ability distribution. However, there is quite a bit of measurement redundancy (yellow highlight). Ideally, as with the mathematics forms, there should be a normal distribution of persons on the left with one or two items on the right at each level of person ability.

Again, similar to what was found with the mathematics test forms, with this level of measurement redundancy, students may have become fatigued, and with no new information being provided about ability, the redundant items could potentially be removed in future applications. However, if there are items that require the same ability level but tap different skills, the items may need to be kept to provide a well-rounded picture of ability. For example, in Figure G10, Item 7 and Item 13 have approximately the same item-difficulty estimate. However, Item 7 was a listening comprehension question that asked students to look at the face of a clock and tell the time, whereas Item 13 tested students' knowledge of English plurals by asking them to indicate the correct singular form of a plural word. Therefore, it is likely that both items were useful.

Figure G10. Wright plot for NEA results, P3 English, Form 1

Primary 3 – Overall Cronbach's α = 0.84

Form 1 – Cronbach's α = 0.85



Not unexpectedly, as with the mathematics forms, the English test forms for both P3 and P6 revealed recurring, standard patterns that match those identified above for Form 1. In each case, the person and item targeting (turquoise highlight) was close or even perfectly aligned, indicating that the test forms were pretty well aligned to the target population; the distribution of persons on the left is relatively normal in each case, with only a few students (between one and six) with maximum and minimum scores; the spread of items is fairly well distributed, with needs really centered in the upper and lower ends of the ability distribution; and there is quite a bit of measurement redundancy (yellow highlight). Thus, the discussion below assumes these features and only addresses form-specific results that are of special interest.

P3 English Form 2 displays the aforementioned measurement redundancies (yellow highlighting in ***Figure G11***), along with examples of items that may need to be kept to provide a well-rounded picture of ability. For example, Item 13 and Item 35 have approximately the same item difficulty estimate, but Item 13 was a plural knowledge question ("Choose the singular form of 'women'") whereas Item 35 was a reading comprehension question based on a brief passage. Therefore, it is likely that both items are useful.

Figure G11. Wright plot for NEA results, P3 English, Form 2

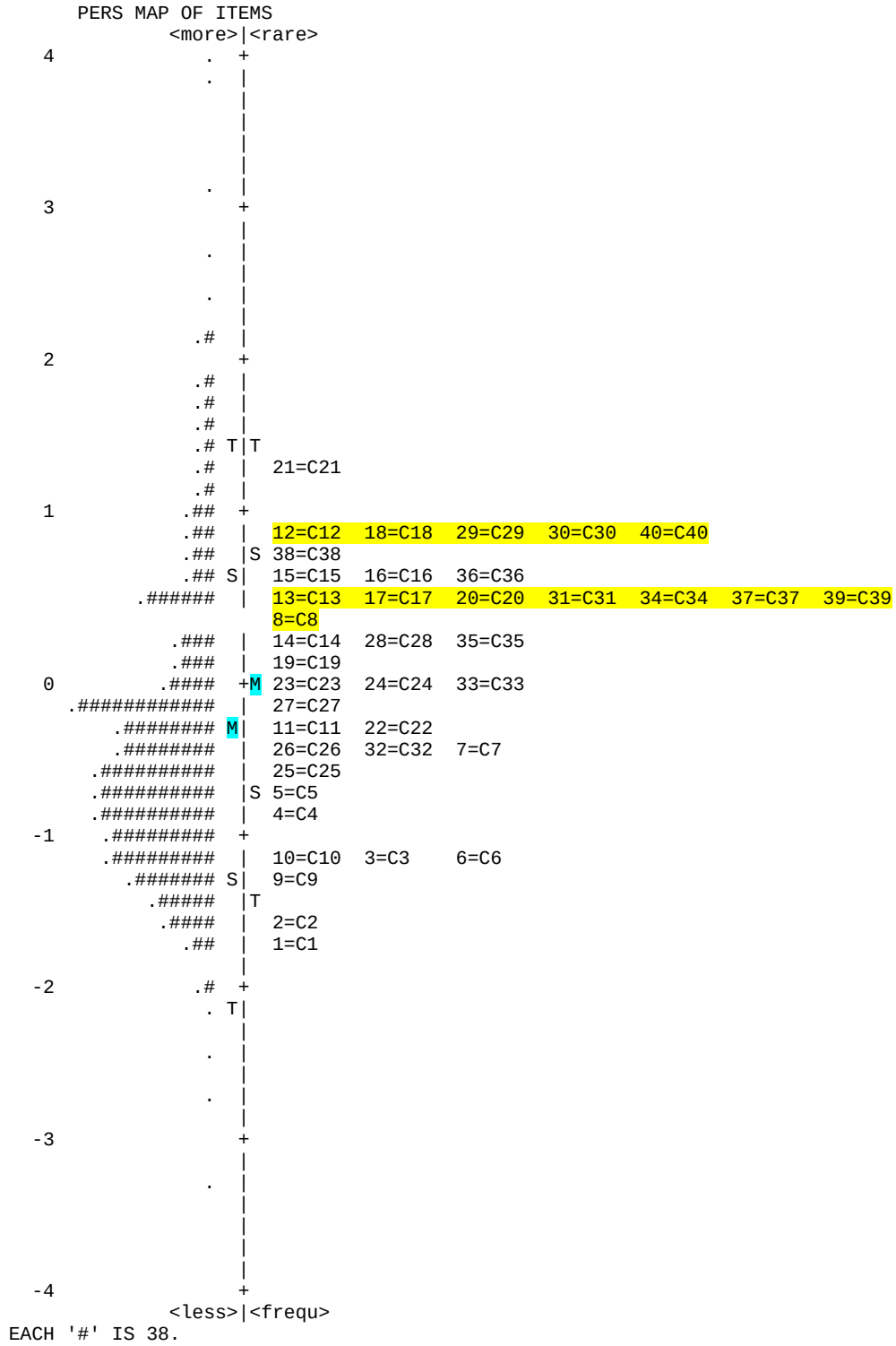
Form 2 – Cronbach's α = 0.83

Form 3 of Primary 3 English suffered from the measurement redundancy issue (yellow highlight), but with some exceptions for items that tapped different skills. For example, as shown in ***Figure G12,*** Item 13 and Item 39 have approximately the same item difficulty estimate, but Item 13 was a plural knowledge question ("Choose the correct singular form of 'feet'"), whereas Item 39 was a punctuation question ("Which sentence has the full stop in the right place?"). Therefore, it is likely that both items were useful.

Item 12 in Figure G12 shows a surprising number of students at the low end of the ability distribution getting Item 12 correct. Looking at item content ("Choose the correct plural form of 'louse'"), there is no immediately apparent aspect of the item that would cause the unexpected responses.

Figure G12. Wright plot for NEA results, P3 English, Form 3

Form 3 – Cronbach's α = 0.84

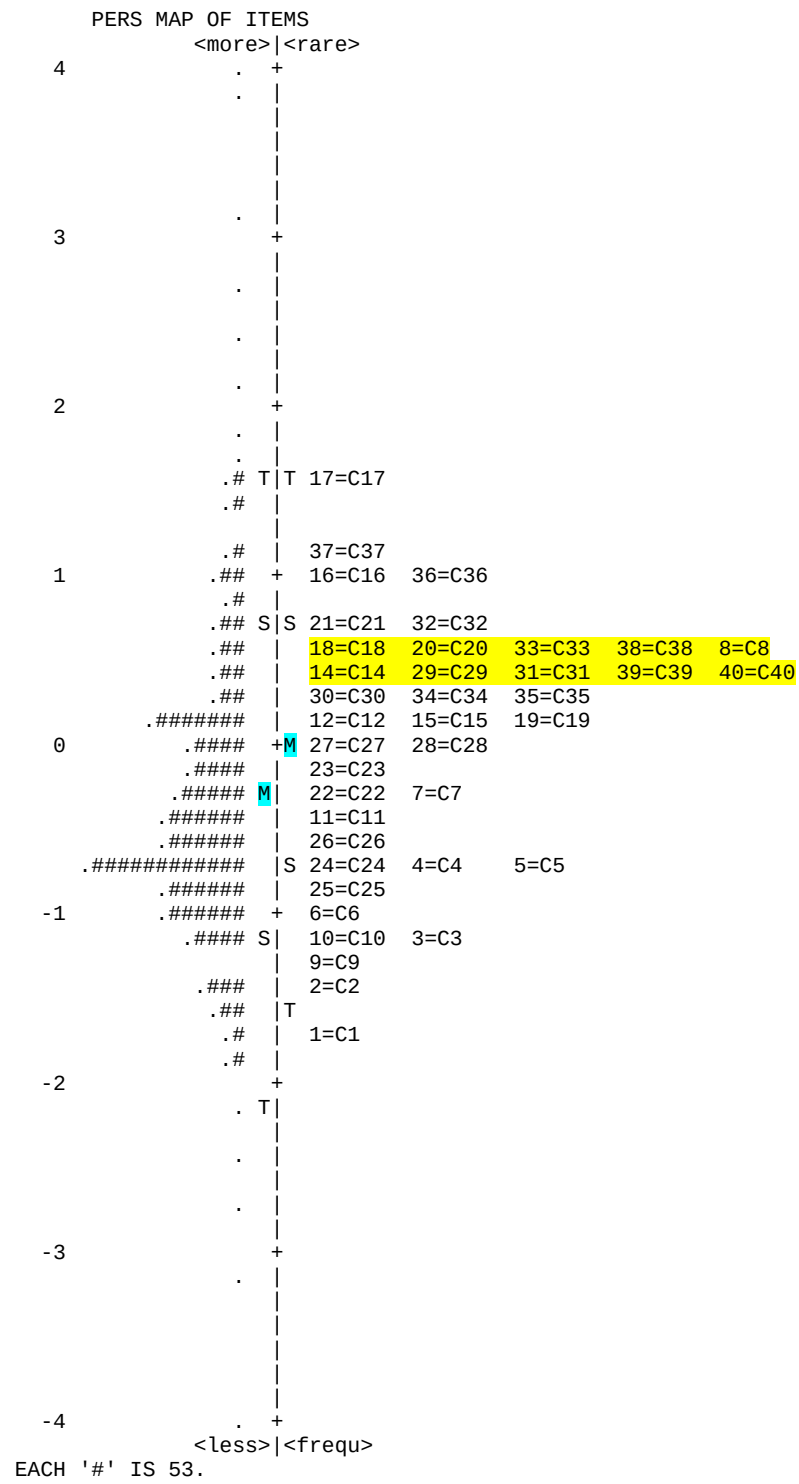


Form 4 of Primary 3 English also shows the typical item measurement redundancy (yellow highlighting in ***Figure G13***), with exceptions for some items that required the same level of effort but tapped different skills. Item 14 and Item 40 have approximately the same item difficulty estimate, but Item 14 was a verb form question ("The boys [works, work, working, is working] on their father's farm") whereas Item 40 was a punctuation question ("Which sentence has the comma in the right place?"). Therefore, it is likely that both items are useful.

Item 17 in Form 4 had a surprising number of students at the low end of the ability distribution who got the item correct. Looking at the item content ("We [run, runs, are running, were running] when the rain started"), there is a possibility that low-ability students chose the correct answer not because they had the skill, but because the correct answer was the longest of the response options. It stood out from the rest of the distracters and could have been selected by chance.

Figure G13. Wright plot for NEA results, P3 English, Form 4

Form 4 – Cronbach's α = 0.84 (C13 Deleted)



As indicated, **Form 1 of Primary 6 English** follows the pattern of the P3 forms in terms of means alignment (turquoise highlighting in ***Figure G14***), distribution of persons on the left by ability level, and measurement redundancy (yellow highlight).

Items that required the same ability level but tapped different skills included Item 7 and Item 29. Item 7 was a listening comprehension question based on a three-sentence story, whereas Item 29 was a sentence completion question ("There were [a lot of, most, all, much] people on the bus so some of us had to stand in it"). Therefore, it is likely that both items were useful.

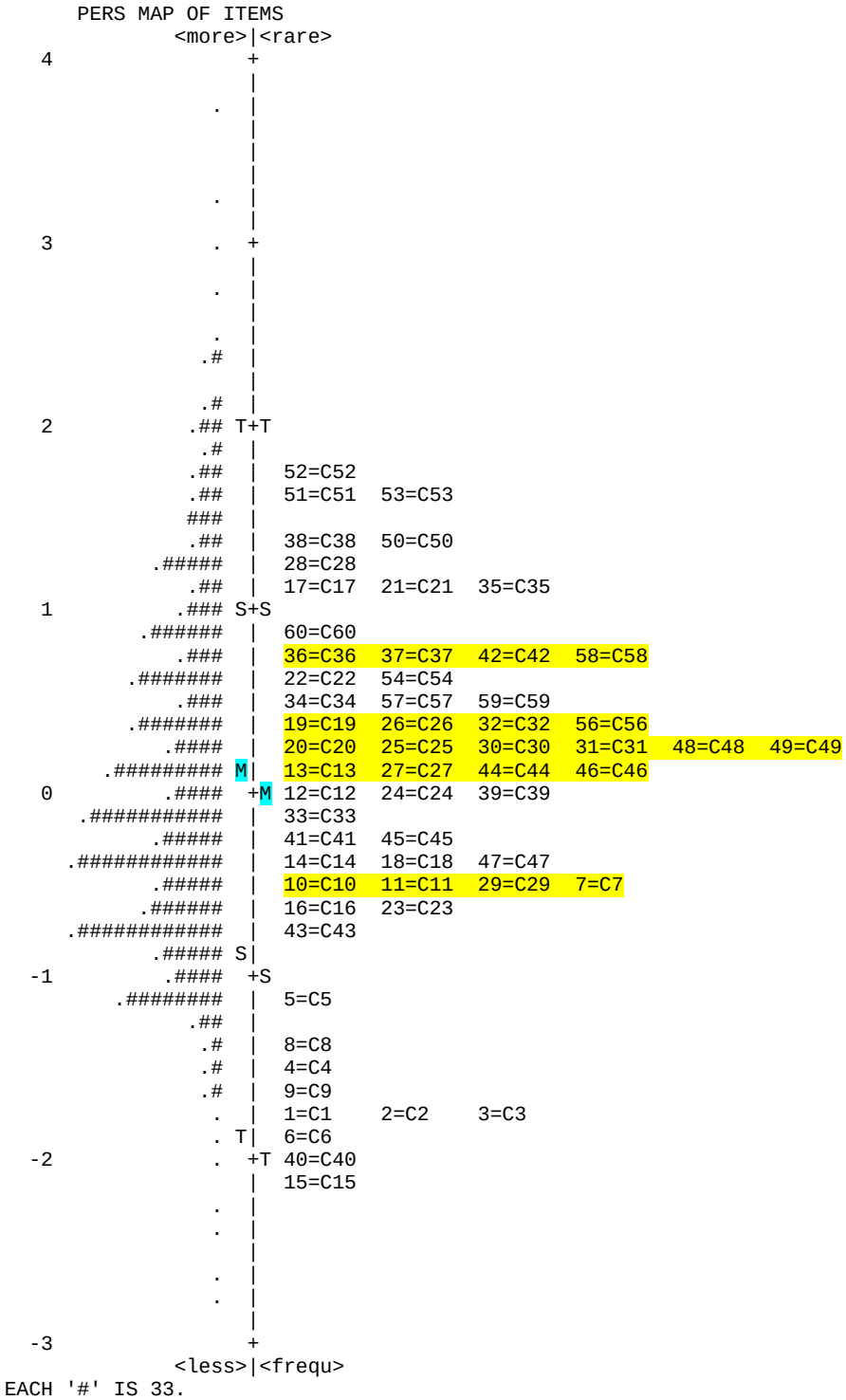
Finally, Form 1 contained five items showing unexpected responses (Item 35, Item 38, Item 50, Item 52, and Item 53). A look at the most unexpected responses reveals that for these items (see toward the top of the item difficulty distribution on the right side of the Wright plot), a surprising number of students at the low end of the ability distribution got the items correct, whereas high-ability students answered incorrectly. This type of finding indicates a severe problem with the item.

Looking at item content, Item 35, which required students to choose a phrase to fill in the blank ("Kwaku is fond of _____his friend's leg to make him cry") did not have a clear correct answer because both "stamping on" and "scratching" could be correct. Item 38 asked students to choose a word to fill in a blank ("Curious means _____"). This item also did not have a clear correct answer, which is certainly the cause of the misfit. Item 50 presented four variations on the same sentence ("'Do you like the sea?' asked the shark"), with different punctuation for each, and asked students to choose the version that was correct. There is no immediately apparent aspect of the item that would cause the unexpected responses. Item 52 ("Which of the following is the correct salutation for a semi-official letter?") did not have a clear, correct answer due to capitalization and punctuation differences in the responses, as well as a lack of definition of "semi-official." Item 53 ("Which of these is the best way of ending a semi-official letter?") had issues similar to those of Item 52, with confusing response options likely causing the misfit.

Figure G14. Wright plot for NEA results, P6 English, Form 1

Primary 6 – Overall Cronbach's α = 0.89

Form 1 – Cronbach's α = 0.90



The person-item targeting (turquoise highlight) for **Form 2 of Primary 6 English** is perfectly aligned ***(Figure G15)***. The distribution of persons on the left is normal with no ceiling or floor effects. The spread of items is fairly well distributed, with needs in the extreme upper and lower ends of the ability distribution. However, as in all the other test forms, there is quite a bit of measurement redundancy (yellow highlight).

Among the items that are at the same ability level but tap different skills are Item 7 and Item 42. Item 7 was a listening comprehension question based on a three-sentence story, whereas Item 42 was a reading comprehension question related to a short dialogue ("Why did the goat cross the bridge?"). Therefore, it is likely that both items were useful.

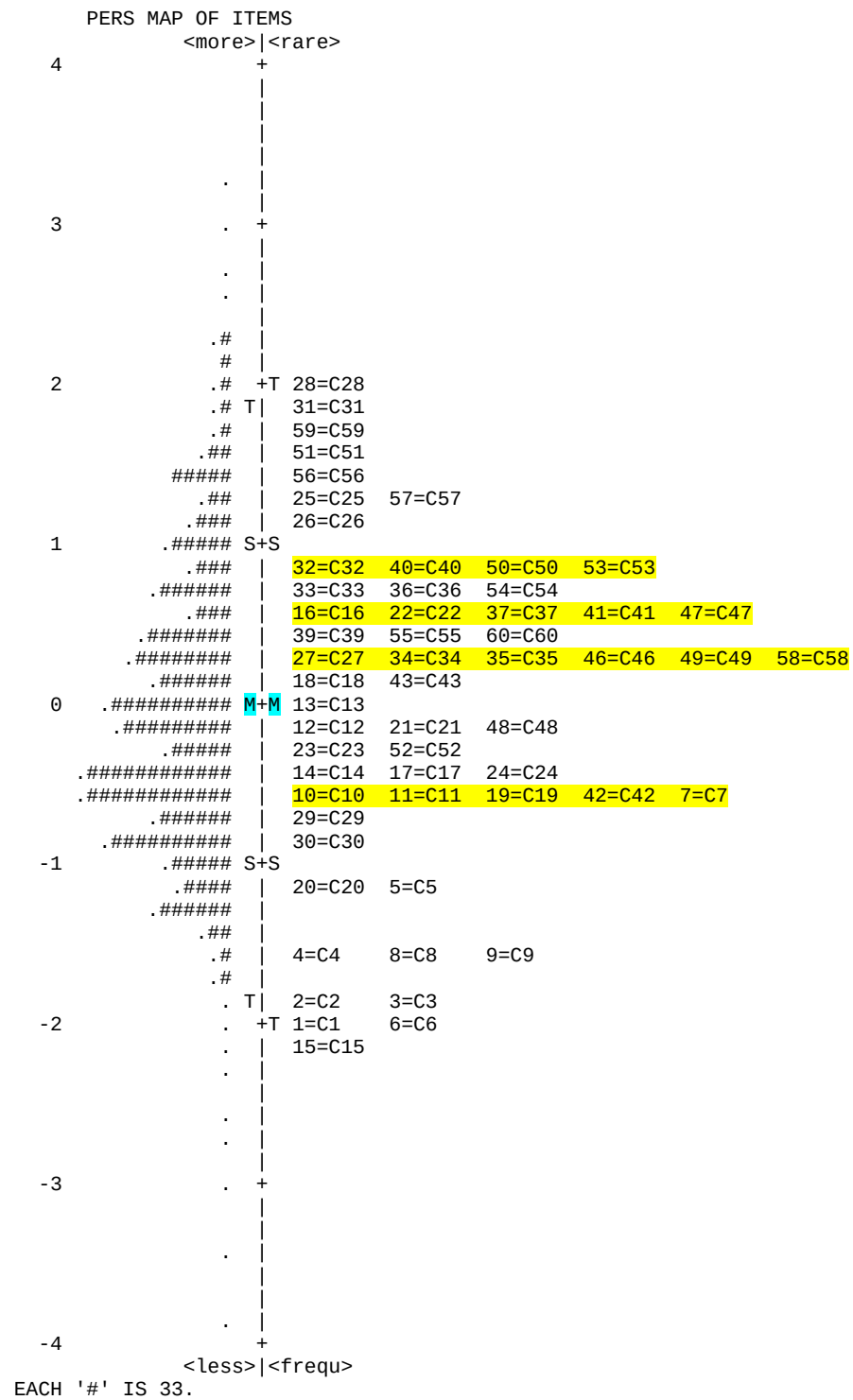
Finally, two items show unexpected responses: Item 31 and Item 59. In this case, a surprising number of students at the middle and low end of the ability distribution got the items correct.

Looking at item content, the instructions for Item 31 were "Choose the correct answer that means the same as the underlined idiomatic expression." However, the question was worded such that students had to choose a phrase to fill in a blank ("The thief ____ to escape punishment"); there was no underlined expression. In addition, several of the multiple-choice responses made sense for this particular sentence, and the one that contained an "idiomatic expression" should have read "told a tall tale" instead of "told a tall story."

Item 59 presented several possible advertising phrases and directed students to "Choose the best advertisement." The students taking this test may have been confused because the responses were all quite similar.

Figure G15. Wright plot for NEA results, P6 English, Form 2

Form 2 – Cronbach's α = 0.89



As with Form 1, the person-item targeting (turquoise highlight) for **Form 3 of Primary 6 English** is perfectly aligned ***(Figure G16)***, the distribution of persons on the left is relatively normal, and the spread of items is fairly well distributed. Likewise, there is quite a bit of measurement redundancy (yellow highlight).

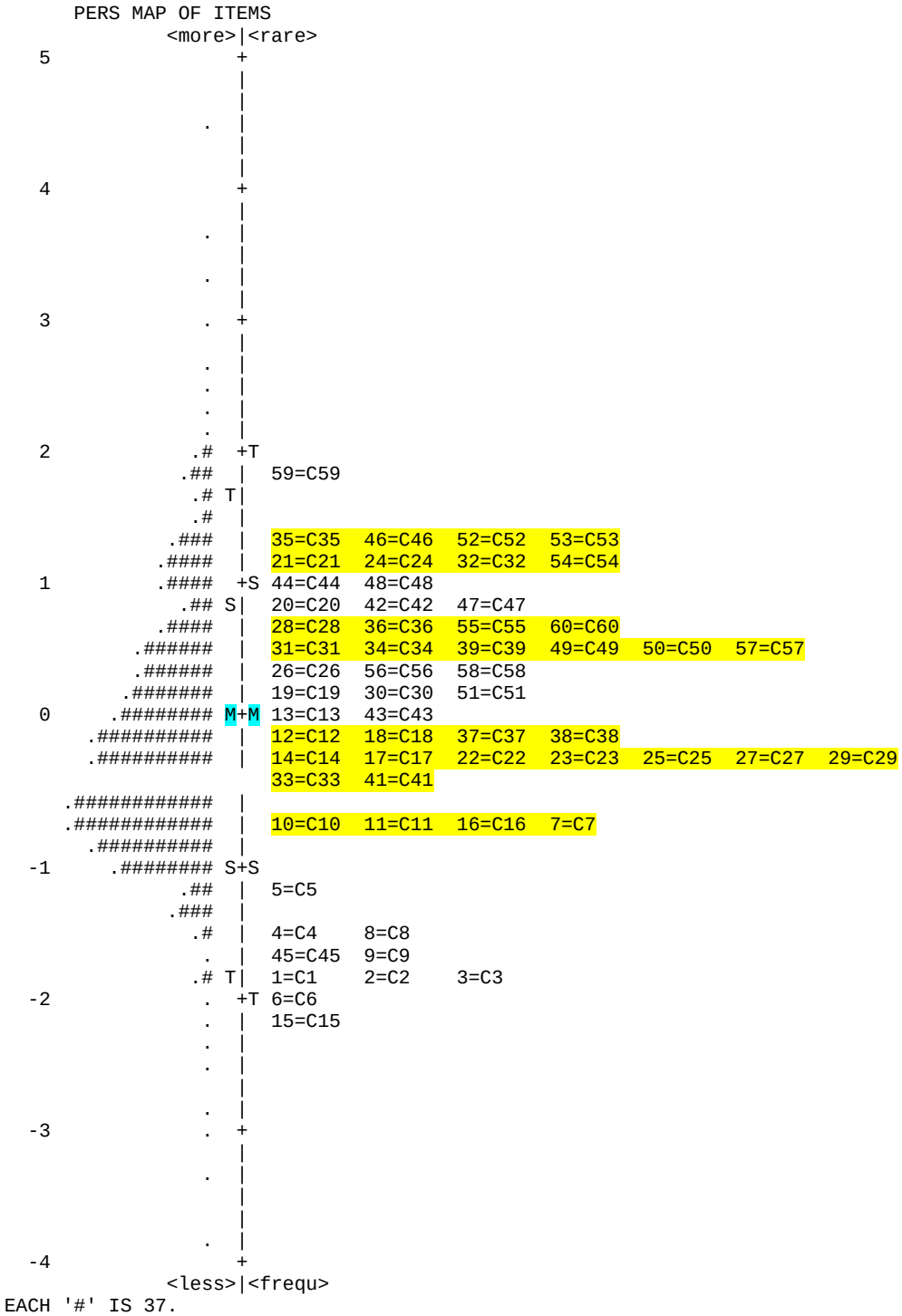
An example of items at the same ability level that tapped different skills is Item 7 and Item 16. Item 7 was a listening comprehension question based on a three-sentence story, whereas Item 16 was a grammar question ("It's nearly nine thirty. We [has, must, were, have] work hard"). Therefore, it is likely that both items were useful.

Three items show unexpected responses: Item 46, Item 53, and Item 59. For the very difficult items (see near the top of the item-difficulty distribution on the right side of the Wright plot), a surprising number of students at the middle and low end of the ability distribution got the items correct.

Looking at the content for Item 46, which required students to read a poem and then choose the best answer to the question "What is the speaker talking about?", there is no immediately apparent aspect that would cause the unexpected responses. Item 53 ("Which of the following is the correct salutation for a semi-official letter?") did not have a clear correct answer due to capitalization and punctuation differences in the answer choices, as well as a lack of definition of "semi-official." Item 59, which contained advertising phrases similar to those constructed for Form 2, also had some response options that were likely confusingly similar to the students taking this test.

Figure G16. Wright plot for NEA results, P6 English, Form 3

Form 3 – Cronbach's α = 0.88



Form 4 of Primary 6 English has the item-difficulty mean only slightly above that of the person ability (turquoise highlight), normal distribution of persons and spread of items, and the usual measurement redundancy (yellow highlight).

Items aimed at the same ability level but tapping different skills include Item 12 and Item 56. Item 12 was a listening comprehension question based on a five-sentence passage, and Item 56 was a question that tested knowledge about traffic law. Therefore, it is likely that both items were useful.

Finally, six items showed unexpected responses: Item 27, Item 34, Item 52, Item 53, Item 57, and Item 59. Surprising numbers of students at the middle and low end of the ability distribution got the items correct, and high-ability students answered incorrectly.

Looking at item content, there were confusing response options on Item 27, which required students to read a sentence and then choose the correct wording for the same sentence as converted to reported speech; and Item 34, which required filling in a blank with the correct phrase ("The boys saw the headmaster and____").

Item 52 ("Which of the following is the correct ending of a friendly letter?") seems to have had two correct answers, causing unexpected responses.

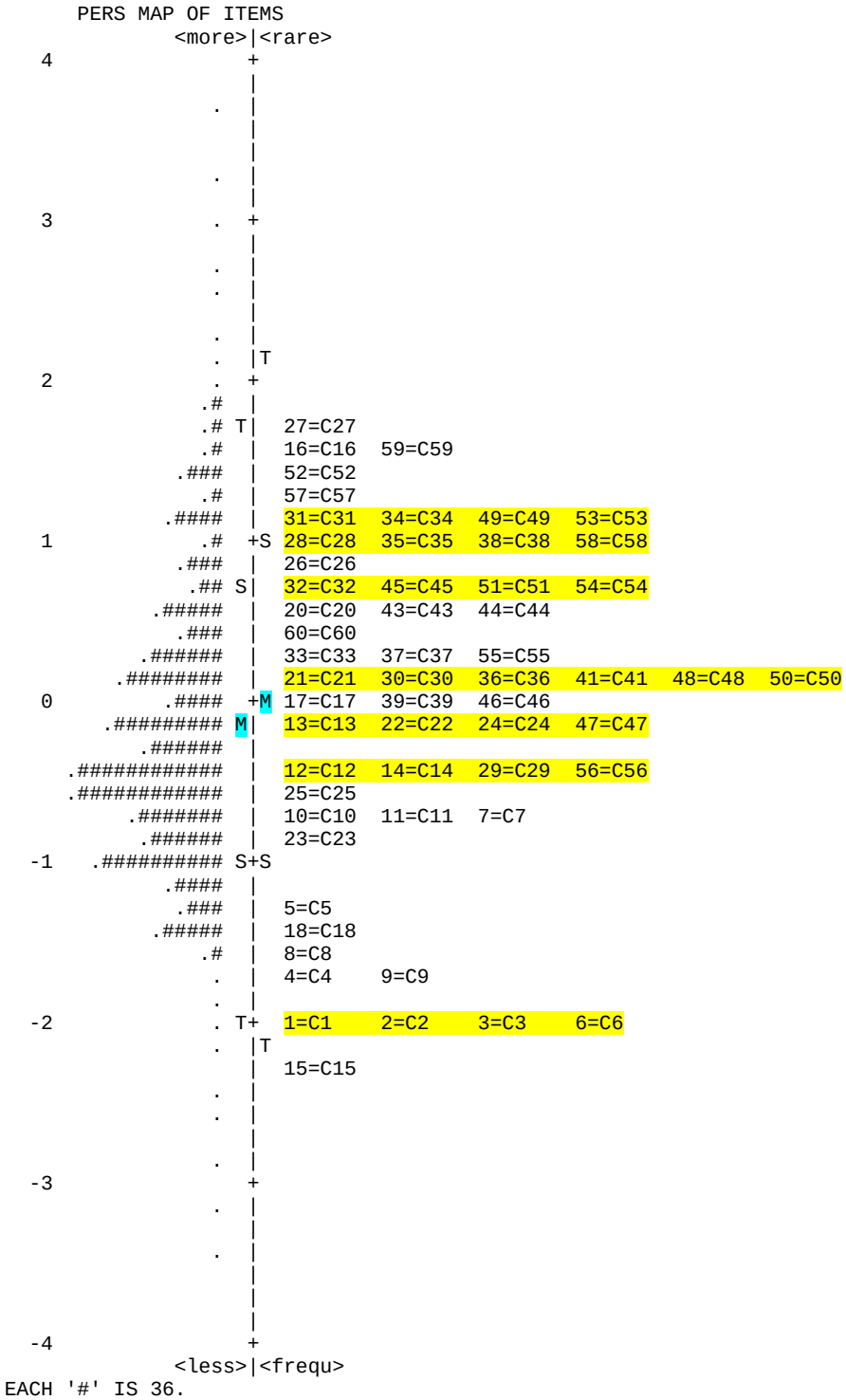
Item 53 ("Which of the following is the correct salutation for a semi-official letter?") did not have a clear correct answer due to capitalization and punctuation differences among the answer options, as well as a lack of definition of "semi-official."

Grammatically, all of the phrases given as response options to Item 57 ("Pedestrians must _____") could be correct in a given situation.

Finally, Item 59, which presented advertising phrases similar to those in the other P6 English forms, had response options that were all quite similar and thus likely confused the students taking this test.

Figure G17. Wright plot for NEA results, P6 English, Form 4

Form 4 – Cronbach's α = 0.88



Annex H: Logistic Regressions Adjusted for Demographic Variables

In order to see how demographic variables correlate with test achievement outcomes for public and NALAP pilot schools, the RTI research team ran a logistic regression model containing all demographic variables for each of the four examinations. Including all demographic variables in the model means that the reported odds ratios (ORs) are adjusted for all other demographic variables (see ***Tables H1 and H2***). For instance, given students' gender, region, school type and test form, P3 students in urban areas were 65% more likely than rural students (OR=1.65) to achieve minimum competency in maths. Similarly, P3 urban students were more than twice as likely (219%) to achieve proficiency than rural students given their gender, region, school type (only public and NALAP schools), and test form number. Girls were 16% less likely (OR=0.84) to achieve MC35 than boys in maths, but were only 3% less likely to achieve MC35 in English (OR=0.97). Greater Accra was designated the reference point for regional performance, and as shown in Table H1, other regions were about half as likely (OR range: 0.41–0.55) to achieve the minimum maths and English competency levels achieved by Greater Accra maths *and* English (OR range: 0.29–0.63).

P3 students in the NALAP schools fared only slightly better than the public schools in maths MC35 (OR=1.18) and significantly (statistically speaking) better in English MC35 (OR=1.38).

Table H1. Weighted odds ratios (adjusted by core demographic variables) of P3 students achieving minimum competency and proficiency levels, by subject (excluding private schools)

Variable		Maths: Odds of achieving (95% confidence interval)[a]		English: Odds of achieving (95% confidence interval)[a]	
		Minimum competency	Proficiency	Minimum competency	Proficiency
Urban/rural status	Rural[b]	1	1	1	1
	Urban	1.65 (1.35, 2.02)***	2.19 (1.64, 2.91)***	1.84 (1.45, 2.33)***	2.43 (1.78, 3.31)***
Gender	Male[b]	1	1	1	1
	Female	0.84 (0.78, 0.91)***	0.89 (0.8, 0.99)*	0.97 (0.9, 1.04)	0.99 (0.91, 1.08)
Region	Ashanti	0.49 (0.31, 0.77)**	0.38 (0.2, 0.71)**	0.51 (0.29, 0.9)*	0.34 (0.16, 0.72)**
	Brong Ahafo	0.52 (0.33, 0.8)**	0.3 (0.16, 0.56)***	0.52 (0.29, 0.93)*	0.28 (0.13, 0.6)**
	Central	0.49 (0.31, 0.77)**	0.31 (0.16, 0.62)***	0.53 (0.31, 0.92)*	0.29 (0.13, 0.65)**
	Eastern	0.53 (0.33, 0.86)*	0.35 (0.17, 0.73)**	0.52 (0.3, 0.93)*	0.31 (0.14, 0.68)**
	Greater Accra[b]	1	1	1	1
	Northern	0.41 (0.26, 0.65)***	0.36 (0.18, 0.72)**	0.53 (0.3, 0.94)*	0.43 (0.19, 0.97)*
	Upper East	0.50 (0.31, 0.79)**	0.42 (0.21, 0.84)*	0.37 (0.21, 0.65)***	0.25 (0.11, 0.57)***
	Upper West	0.43 (0.27, 0.69)***	0.32 (0.16, 0.63)**	0.29 (0.16, 0.54)***	0.27 (0.11, 0.66)**
	Volta	0.42 (0.27, 0.66)***	0.27 (0.14, 0.51)***	0.35 (0.2, 0.61)***	0.22 (0.1, 0.49)***
	Western	0.55 (0.33, 0.89)*	0.5 (0.24, 1.05)	0.63 (0.35, 1.13)	0.41 (0.17, 0.94)*
School type	Public[b]	1	1	1	1
	NALAP pilot school	1.18 (0.95, 1.47)	1.03 (0.75, 1.42)	1.38 (1.04, 1.84)*	1.03 (0.68, 1.54)
Test form number	1[b]	1	1	1	1
	2	1.30 (1.18, 1.43)***	1.36 (1.17, 1.57)***	0.71 (0.65, 0.78)***	0.56 (0.49, 0.65)***
	3	1.23 (1.12, 1.36)***	1.35 (1.18, 1.55)***	0.88 (0.8, 0.96)**	0.83 (0.72, 0.95)**
	4	1.35 (1.22, 1.49)***	1.51 (1.3, 1.76)***	0.98 (0.89, 1.08)	0.84 (0.74, 0.96)**

[a]Weighted odds ratio estimates and 95% confidence intervals were calculated using Taylor linear series to account for the sample design's cluster effect.
[b]Reference class.
*$p < .05$.
**$p < .01$.
***$p < .001$.
Note: Significance was calculated based on Pearson's Chi-squared test.

Table H2. Weighted odds ratios (adjusted for core demographic variables) of P6 students achieving minimum competency and proficiency levels, by subject (excluding private schools)

Variable		Maths: Odd of achieving (95% confidence interval)[a]		English: Odd of achieving (95% confidence interval)[a]	
		Minimum competency	Proficiency	Minimum competency	Proficiency
Urban/rural status	Rural[b]	1	1	1	1
	Urban	1.99 (1.58, 2.5)***	3.1 (2.13, 4.52)***	2.65 (2.02, 3.48)***	2.87 (2.09, 3.95)***
Gender	Male[b]	1	1	1	1
	Female	0.75 (0.69, 0.81)***	0.64 (0.55, 0.74)***	0.87 (0.79, 0.95)**	0.92 (0.83, 1.01)
Region	Ashanti	0.52 (0.27, 0.98)*	0.39 (0.15, 0.97)*	0.36 (0.18, 0.69)**	0.34 (0.17, 0.7)**
	Brong Ahafo	0.39 (0.21, 0.74)**	0.23 (0.09, 0.62)**	0.2 (0.11, 0.39)***	0.21 (0.1, 0.47)***
	Central	0.36 (0.2, 0.67)**	0.22 (0.08, 0.58)**	0.25 (0.13, 0.47)***	0.19 (0.09, 0.39)***
	Eastern	0.55 (0.27, 1.09)	0.44 (0.17, 1.17)	0.27 (0.14, 0.53)***	0.34 (0.15, 0.75)**
	Greater Accra[b]	1	1	1	1
	Northern	0.24 (0.13, 0.45)***	0.11 (0.04, 0.3)***	0.19 (0.1, 0.37)***	0.16 (0.08, 0.33)***
	Upper East	0.46 (0.24, 0.87)*	0.4 (0.15, 1.09)	0.22 (0.11, 0.42)***	0.21 (0.1, 0.46)***
	Upper West	0.45 (0.24, 0.84)*	0.34 (0.14, 0.87)*	0.25 (0.13, 0.47)***	0.17 (0.08, 0.36)***
	Volta	0.43 (0.23, 0.8)**	0.39 (0.15, 1)*	0.28 (0.15, 0.52)***	0.26 (0.13, 0.52)***
	Western	0.53 (0.26, 1.06)	0.56 (0.17, 1.86)	0.29 (0.14, 0.58)***	0.26 (0.1, 0.64)**
School type	Public[b]	1	1	1	1
	NALAP pilot school	1.43 (1.11, 1.85)**	1.39 (0.98, 1.98)	1.78 (1.37, 2.33)***	1.55 (1.15, 2.08)**
Test form number	1[b]	1	1	1	1
	2	0.87 (0.79, 0.97)**	0.9 (0.79, 1.03)	0.97 (0.88, 1.09)	0.79 (0.71, 0.89)***
	3	1.22 (1.1, 1.35)***	1.22 (1.05, 1.41)**	0.88 (0.8, 0.98)*	0.66 (0.59, 0.73)***
	4	1.04 (0.94, 1.15)	0.89 (0.76, 1.03)	0.77 (0.7, 0.85)***	0.5 (0.44, 0.56)***

[a]Weighted odds ratio estimates and 95% confidence intervals were calculated using Taylor linear series to account for the sample design's cluster effect.
[b]Reference class.
*$p < .05$.
**$p < .01$.
***$p < .001$.
Note: Significance was calculated based on Pearson's Chi-squared test.