National Primary Learning Assessment Report: Phase 2

For the Liberian Ministry of Education

Table of Contents

Forward

Through the transition from a content-based to a competency-based curriculum, Liberia has placed learning at the center of its education system. I am proud to announce the development of a National Learning Assessment System for primary grades as the next critical milestone in this process.

The development of the second phase of the National Learning Assessment Policy (NLAP) and Framework directly aligns with Liberia's 10-year Strategy for Education Reform, responding to the call for the need for an assessment tool to inform curriculum and teacher practice. More generally, it furthers the goals listed under Pillar 3 of the Liberia Rising Vision 2030 of eradicating illiteracy and emphasizes the focus on building human capital to reduce the risk of conflict.

The NLAP provides the Ministry of Education with key tools to accurately assess the status of learning at primary grades in the country, allowing for regional and international comparison as well as providing early diagnostics of learning outcome gaps. It thus also presents a gateway for the Ministry to embed a culture of evidence-informed decision-making and response.

This policy and framework are the outcome of the Ministry of Education partnering with Innovations for Poverty Action and the Global Partnership for Education, as well as numerous stakeholders in the education sector. I would like to express my thanks and appreciation to the World Bank as the Grant Agent for the GPE-funded Getting to Best in Education and Liberia Learning Foundations Projects and all collaborative partners for their immense contribution. The development and implementation of Liberia's first primary learning assessment system were funded under these two projects.

It is my hope that a successful implementation of the second phase of the NLAP will pave the way for the introduction of a series of targeted and continuously adapted best-fit policies in forming an education system that enables Liberian students to reach their potential and strive for excellence.

Hon. Prof. Ansu D. Sonii
Minister of Education

Executive Summary

In 2023, Innovations for Poverty Action conducted the second phase of its support to Liberia's Ministry of Education in the work to develop and pilot a National Learning Assessment Policy and Framework. In this second phase, IPA used the results of the first phase to improve the design of the assessments, and then administered the new assessments to 3,096 students across 123 schools in 27 districts and 8 counties. While certain counties were excluded due to budget limitations, a sampling and weighting strategy was implemented in this phase in order to provide reasonably representative coverage of the lower basic education Liberian student population. Written literacy and math exams were conducted with every available student in grades 3 and 6 in each school, with oral exams conducted with a subset of 3rd graders, again in each school. This report summarizes the characteristics of the students in the pilot sample, the average test scores, and some general patterns and features found in the distribution of test scores across gender, age, region, socioeconomic status, and content domain. It also describes how this work, and the results of the first phase as well, can serve as a foundation for the further development and implementation of the National Learning Assessment Policy and Framework and its potential contribution to both Liberian and global initiatives.

Key Findings

- Students were selected across a wide variety of regions and demographic groups, yielding a valid representative sample of the eight counties visited.
- The majority of students in both grades were more than two years above the standard age of entry for their grade level, and older students tended to do slightly worse on most assessments, especially the written math assessment.
- Montserrado and Nimba counties are consistently among the highest scoring on written exams for both grades and subjects, while Margibi county usually had the lowest scores.
- Compared to the first stage of this pilot in 2021, the written math and literacy assessments administered in this stage appear to be more reliable and better able to distinguish high- and low-performing students.
- The literacy tests for both grades were marginally better able to distinguish high- and low-performing students than the written math tests, with the former meeting the threshold for reliability set for this assessment.
- There is little variation in average oral assessment scores across county, gender, age, subject, and school type.
- Male and female students performed similarly across all grades and subjects, as did students at rural and urban schools.
- Students at private and faith-based schools consistently saw slightly higher average written literacy scores and similar math scores compared to public schools.

- Average scores within subject often vary significantly by content domain. For example, pre-reading and arithmetic questions receive higher scores than reading comprehension or geometry questions across both grades.
- Further analysis of student assessment results will call for the development of national and international proficiency benchmarks and learning standards.

Methodology

Sampling

Innovations for Poverty Action (IPA) used multiple data sources to develop a sampling strategy, and ultimately selected 123 schools to visit across 26 districts in eight counties across Liberia. Unlike the sample from Phase 1 of this project, schools in Phase 2 were selected to form a representative sample of 3rd and 6th grade students across region, age, and school type. Because of this important difference in survey design, this report is unable to make valid comparisons between the results of Phases 1 and 2. In each school, tests were administered with every third- and sixth-grade student present. Counties were selected to cover a wide geographical range, omitting counties that were deemed logistically infeasible due to budget and time constraints. Within counties, districts were selected randomly with probability proportional to their total estimated lower basic primary school student enrollment. These enrollment estimates were taken from the 2021 annual school census conducted by the Ministry of Education. Schools were then selected randomly within district, with a target of three schools per district, taken from a comprehensive list of schools also compiled by the Ministry of Education. At least one urban and one rural school was selected in any district that had one of each, with the third selected according to whether the majority of households in that district were considered rural in the most recent Demographic and Health Survey. As estimates of student enrollment or grades were not available, schools were selected randomly with equal probability, with additional schools randomly selected if the initially selected schools were too small to reach the target sample size. A total sample size of at least 3,000 students was targeted in response to the direct request from the MOE, with 3096 students ultimately assessed. IPA staff administered exams with every 3rd and 6th grade student available at the school on the day in which the school was visited.

Variable Construction and Analysis

In Phase 1, both written and oral assessments were administered to all students in the sample in both third and sixth grade. Given the successful administration and performance of the written exams, for example, that there were no zero scores, the team recommended that moving forward, written exams be used for both third and sixth grade, but that an oral assessment still be used with a subsample of students in third grade in order to assess pre-literacy skills and contribute to the on-going discussions in Liberia concerning oral reading fluency benchmarks. Accordingly, in Phase II, oral tests were exclusively administered to third-grade students. Design weights at the school level were calculated via inverse probability weighting to ensure that sample averages were representative of the overall student body of each county and of Liberia overall. Given the discrepancy in the number of students participating in the oral and written tests, distinct weights were computed for each test type. Standard errors and confidence intervals are clustered at the county level and calculated using Stata's standard utility for complex survey sampling designs. To aggregate test results, the percentage of correct responses was calculated by subject (literacy or math) and type of test (oral or written) for each student. The results are then disaggregated by county, student gender, urbanicity, grade level, test type, and school type.

Test Design and Performance

Methodology

Psychometric analysis was used to evaluate the performance of the assessments in relation to multiple dimensions. Analysis was conducted at two points: first, with the results of the Phase 1 pilot. This analysis was used to refine the assessments. Across all four written assessments (grade 3 literacy, grade 3 math, grade 6 literacy, grade 6 math), IPA dropped certain problematic and/or misfitting items, edited certain problematic and/or misfitting items and added in new items. Phase II administered the edited assessments. The second round of psychometric analysis was conducted with the Phase II edited assessments. This report presents findings from this second round of analysis.

By examining students' performance on individual questions, each test can be evaluated on two important dimensions: reliability and separation. A measure of reliability indicates how consistently the test measures a student's aptitude against measurement error or random variation. Conversely, a measure of separation showcases how distinctly the test can differentiate between students of differing aptitude. Reliability is measured on a scale from zero to one, with a value over 0.8 being generally considered acceptable in this case. Meanwhile, separation is reported in terms of how many distinct groups the test can identify, with any value over two being acceptable in this case. The written literacy and math tests were evaluated separately on both dimensions. The questions themselves were also evaluated for reliability and separation. Questions are considered reliable if a student's chance of answering correctly is a strong indication of their aptitude and separable if they range significantly in difficulty. Specific questions, highlighted in Appendix 4, require further examination or alterations based on the Rasch model analysis.

Literacy assessment performance

For both grades, the written literacy assessment saw a marked improvement in both reliability and separation over the first phase of the NLAP pilot. The reliability for grades 3 and 6 were estimated at 0.83 and 0.86, respectively, clearing the accepted threshold. Separation was above 2 for both grades as well. The improvement seen in the second pilot phase can be attributed in large part to the sample chosen, which captured a much more diverse group of students. The questions themselves were found to be quite reliable and well separated overall, though the grade 6 assessment may benefit from more difficult questions to better differentiate among high-performing students. Appendix 4 contains specific recommendations and suggestions pertaining to individual questions.

Math assessment performance

The Grade 3 math assessment questions exhibit high reliability and separability, differentiating across 22 levels. However, despite a noticeable enhancement from the prior stage, the tests reliability in gauging students' aptitude fell short of the recommended threshold at 0.74. This shortfall is explained by the fact that a significant portion of sampled students scored within a relatively narrow ability range. Separation also fell below the expected threshold with a value of 1.72, at least in part for the same reason. This means the current grade 3 math assessment is limited in its ability to differentiate among students of differing levels of ability. Appendix 4 identifies a number of questions that might be improved to resolve these concerns, with a general recommendation to identify questions that reliably target the “minimally acceptable” candidate student.

In the grade 6 math assessment, initial results of the Rasch model analysis led to similar conclusions as in grade 3, with insufficient reliability and separation. However, after taking additional measures to minimize the role of guessing, acceptable values of 0.83 and 2.18 were achieved. As in all other cases, the questions themselves achieved high levels of reliability and separation. Again, these values all represent improvement over the previous 2021 pilot round, but with opportunities for improvement on specific questions identified in the appendix.

Results and Analysis

Student Demographics

The second phase of the Liberia National Learning Assessment Policy and Framework assessed 3,096 students in Grades 3 and 6 from 120 schools among eight counties. The sample consisted of 1,504 students from Grade 3 (48.6%), and 1,592 students from Grade 6 (51.4%). Of these, female students composed 53.13% of Grade 3 and 53.45% from Grade 6. Most students in the sample come from urban schools, 58.51% of those in Grade 3 and 66.33% of those in Grade 6. By far, the largest share of students comes from Montserrado county (47% of the total sample), followed by Nimba (15%). Gbarpolu, the county with the smallest number of students, was combined with nearby counties during school selection, resulting in 3.84% of the sampled students in eight schools.

Table 1. Sample Composition by County and Grade

			Number of students in the sample		
County	**Number of students in the county 2020-21**	**Percentage of students by county**	**Grade 3 - Written**	**Grade 6 - Written**	**Grade 3 - Oral**
Total	452,459	100%	1,504	1,592	877
Bong	42,268	9.3%	157	109	133
Gbarpolu	7,362	1.6%	56	63	41
Grand Bassa	27,894	6.2%	70	80	48
Grand Cape Mount	13,999	3.1%	145	193	95
Margibi	41,396	9.1%	158	127	69
Montserrado	252,008	55.7%	643	826	323
Nimba	67,532	14.9%	275	194	168

Female students represent about 53% of the whole sample, with some variation across counties. Most students surveyed in urban schools were female (56%), while the gender composition in rural schools is balanced (49% female students). Further investigation into this issue could be useful to explore if in urban schools, there is higher dropout among male students than female students, for example.

In both grade levels, a significant portion of the sample for written tests consists of students attending urban schools, with 59% for grade 3 and 66% for grade 6. In contrast, the sample of students who took the oral test exhibits a more balanced distribution in terms of school urbanicity, as 49% of this sample attends urban schools.

There is a striking disparity in the percentage of students attending schools in urban areas across different counties. For instance, in Grand Cape Mount County and Gbarpolu, all surveyed students attend schools in rural areas, whereas in Montserrado County, 99% of students attend urban schools. In Bong County, approximately a quarter of the sample attends urban schools, whereas in Margibi County, three-quarters of the sample attends urban schools.

The age of the surveyed students consistently exceeds the theoretical age range for students in both grade 3 and grade 6. Liberia’s standard age of entry policy recommends students enter grade 3 at around 8 years of age. However, 93% of the surveyed sample surpasses this age, with nearly 40% of them being 12 years old or older. Similarly, for grade 6 students, who should theoretically be around 11 years old, a significant 90% of the sample exceeds this age, with over half falling within the age range of 14 to 20.

This trend holds true for the subsample of grade 3 students who participated in the oral assessment (as no grade 6 students took part). Figure 1 illustrates that a substantial 93% of the students in the sample are older than the theoretical age of 8, with the majority falling within the age bracket of 10 to 15 years old.

Table 2. Sample Composition by County, Grade, Gender and Urbanicity

	% Female students by county			**% students in Urban schools by county**		
County	**Grade 3 - Written**	**Grade 6 - Written**	**Grade 3 - Oral**	**Grade 3 - Written**	**Grade 6 - Written**	**Grade 3 - Oral**
Total	53%	53%	49%	59%	66%	49%
Bong	50%	59%	47%	21%	33%	23%
Gbarpolu	39%	43%	34%	0%	0%	0%
Grand Bassa	47%	48%	52%	10%	5%	15%
Grand Cape Mount	50%	59%	48%	0%	0%	0%
Margibi	51%	43%	49%	74%	76%	72%
Montserrado	59%	54%	53%	99%	99%	98%
Nimba	49%	56%	47%	32%	51%	18%

Figure 1: Students by Age and Grade -Written Assessment

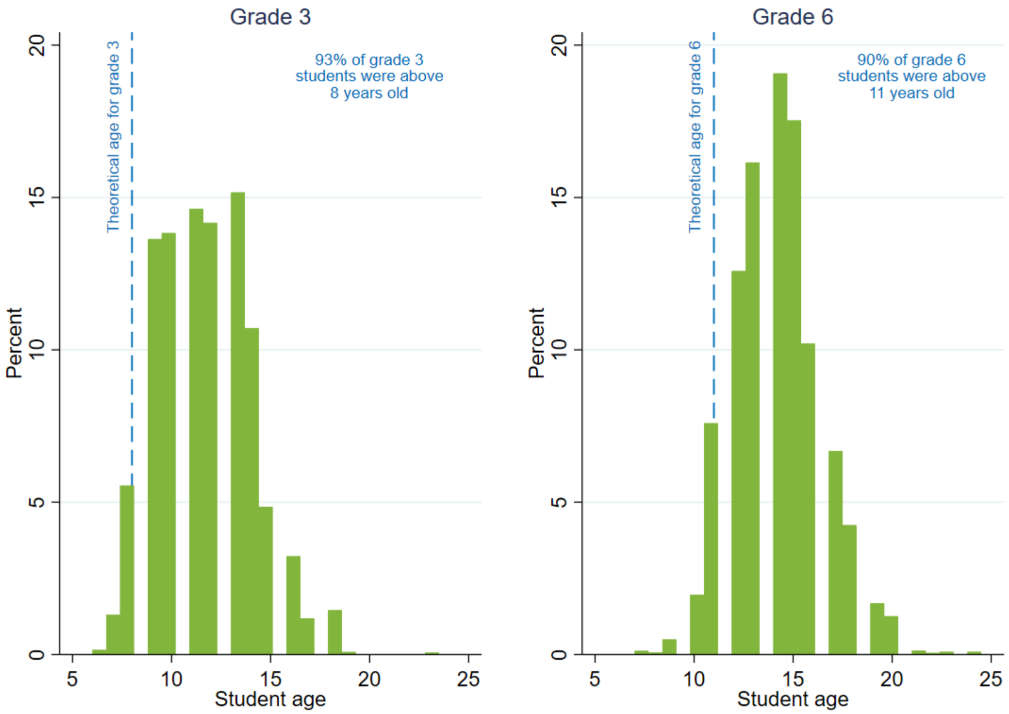
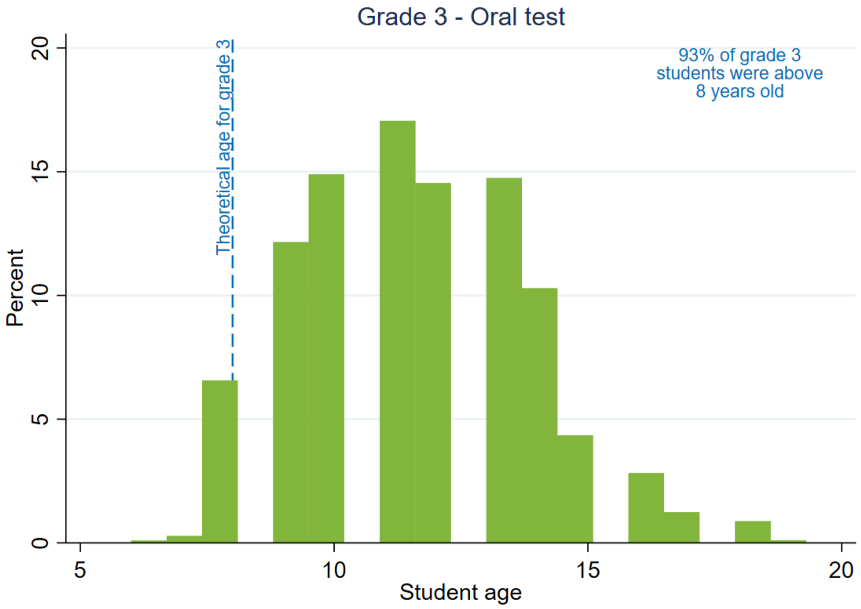


Figure 2: Distribution of Sampled Students by Age and Grade - Oral Assessment



As detailed in the National Learning Assessment Policy, performance on the assessments can in the future be linked to other data sources, such as the school census, to investigate various factors that might be associated with learning, for example, the presence of trained teachers. For Phase II, the team was interested in exploring if students could self-report a few key indicators of socio-economic status to determine how SES might be associated with learning at an individual level. Through reviewing the Poverty Probability Index and Demographic and Health data for Liberia, the team identified a few possible questions to pilot as indicators of SES that students might be able to self-report and explore the potential of this strategy to embed analysis relating to equity in the administration of the assessments.

As part of the written assessments, students were asked about various aspects of their households, including the presence of electricity and the primary material used for the main floor of their residences. Enumerators used various prompts and visuals - such as pictures of a cell phone - to support students in their answering the questions.

In total, 62% of students reported the presence of electricity in their households (with a $p<0.05$ CI of 54% to 70%). Among students attending urban and rural schools, household electrification is reported at 73% and 37%, respectively. Nimba County stands out with the highest percentage of students reporting access to electricity in their households (80%). In contrast, Grand Bassa County exhibits the lowest percentage at only 30%.

A total of 81% of students reported having concrete floors in their dwellings. This figure varies from 63% of students in Gbarpolu to 89% in Montserrado. The disparity between students attending rural and urban schools is notably smaller compared to the electricity values. Specifically, 85% of students from rural schools reported having concrete floors in their households, while 73% of students from urban schools reported the same.

Moving forward, individual indicators - such as the presence of electricity in the household - or a potential composite or index of indicators - can be used to measure SES among the student population and use these measures to investigate issues such as selection into school type as well as the potential association between SES and learning.

Table 3: Percentage of surveyed students who reported their household has electricity

Sample	Sample size	Mean	CI - low	CI - high	Std. Err.
Total	3096	62%	54%	70%	4%
Urban	1936	73%	65%	82%	4%
Rural	1160	37%	31%	42%	3%
Bong	266	36%	26%	46%	5%
Gbarpolu	119	45%	34%	55%	5%
Grand Bassa	150	30%	18%	43%	6%
Grand Cape Mount	338	64%	52%	76%	6%
Margibi	469	34%	28%	40%	3%
Montserrado	285	48%	41%	56%	4%
Nimba	1469	80%	72%	88%	4%

Table 4: Percentage of students in households with concrete floors

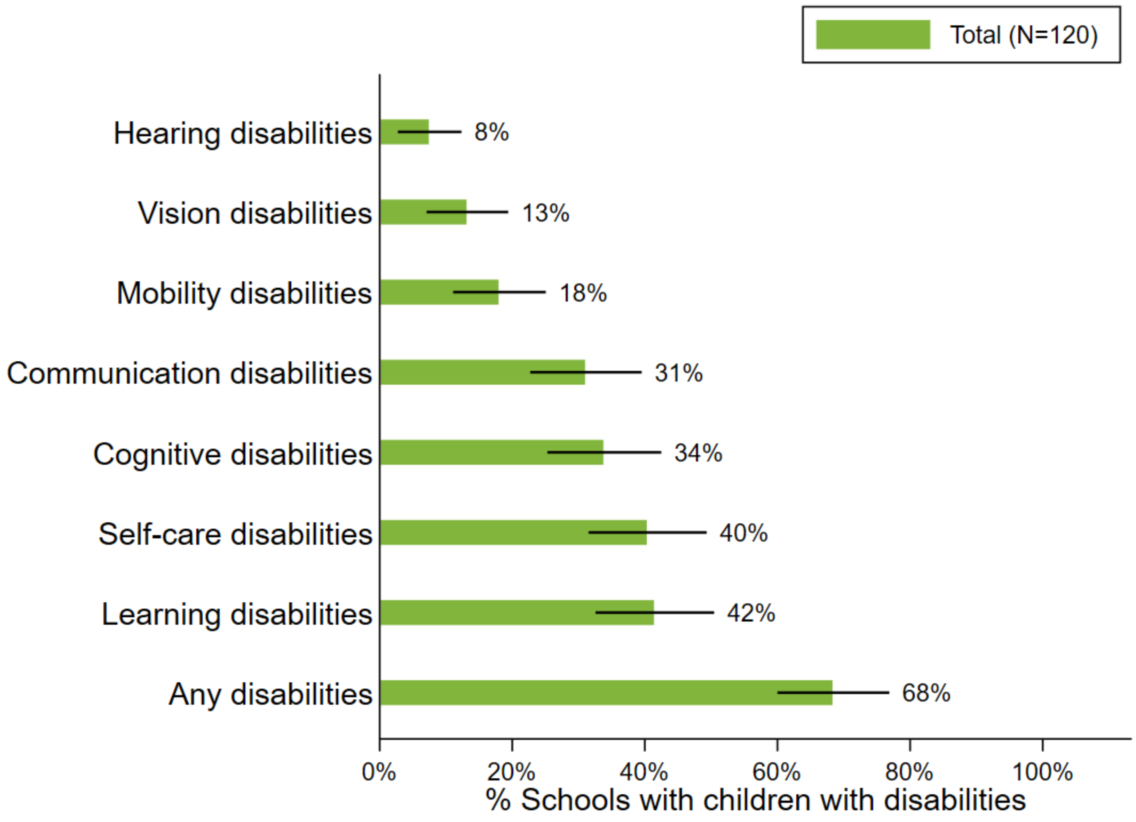
Sample	Sample size	Mean	CI - low	CI - high	Std. Err.
Total	3096	81%	77%	85%	2%
Urban	1936	85%	80%	89%	2%
Rural	1160	73%	67%	80%	3%
Bong	266	76%	66%	85%	5%
Gbarpolu	119	63%	54%	72%	5%
Grand Bassa	150	82%	67%	96%	7%
Grand Cape Mount	338	84%	79%	89%	3%
Margibi	469	71%	61%	80%	5%
Montserrado	285	89%	78%	100%	6%

Nimba	1469	84%	79%	89%	3%

As detailed in the NLAP, students with disabilities are an important subpopulation in which to investigate issues relating to learning. In Phase II, IPA consulted with various experts to try to determine a best-practice approach to identifying students with disabilities that was feasible within the constraints of the assessment exercise. Ultimately the 6-item Washington Group Short Set of Disability Questions (WGQ) was selected. In addition, the team collected qualitative data from a subsample of teachers to investigate their understanding of disability and impressions of the issues they see in their students, particularly as it relates to identifying students for useful understanding of trends in learning. These data can be used to help refine an approach to disability for the NLAP in the future.

Figure 3, displays the percentage of sampled schools which have enrolled students with disabilities. Of the 120 schools, 84 (68%) reported having at least one student with a disability present. The most common were learning, self-care, and cognitive disabilities, with 40%, 39%, and 34% of schools, respectively. Students with hearing, vision, and mobility appeared less common, being present (according to the principal) in only 7%, 13%, and 18% of schools.

Figure 3: Percentage of schools with children with disabilities



Student assessment scores

As scoring rubrics and benchmarks for each grade and subject have yet to be developed, we take each student's overall score in each subject to be the overall percentage of questions answered correctly. The distribution of overall assessment scores for each grade and subject are presented in Figure 4. Grade 3 students achieved an average of 72% of correct responses in the literacy test, whereas grade 6 students averaged 67% in correct responses. The distribution of scores on the written tests were similar across grades, but are notably different by subject. For both 3rd and 6th graders, math assessment scores were more or less symmetrically distributed around the mean, with 99% of students correctly answering between 20% and 80% of questions. By contrast, the distribution of literacy scores is skewed significantly upwards for both grades, with 36% of 3rd graders and 29% of 6th graders answering more than 80% of questions correctly.

In the context of the oral assessment , which is exclusive to grade 3 students , the average scores for both literacy and math stand at 78%. These scores are accompanied by comparable distributions across the percentage of correct responses (Figure 5). Unlike the written tests, the oral assessments in each subject also saw a number of students achieve perfect scores.

Figure 4: Performance in Written Assessments by Grade and Subject

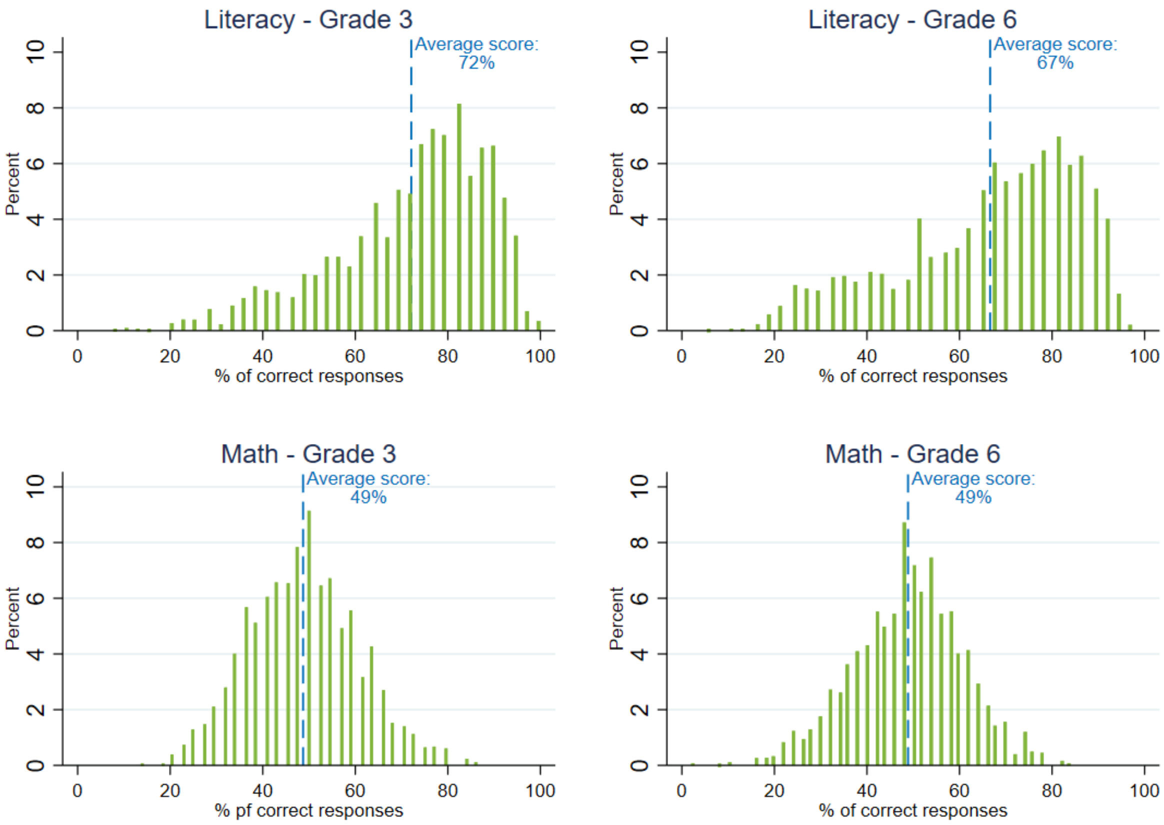
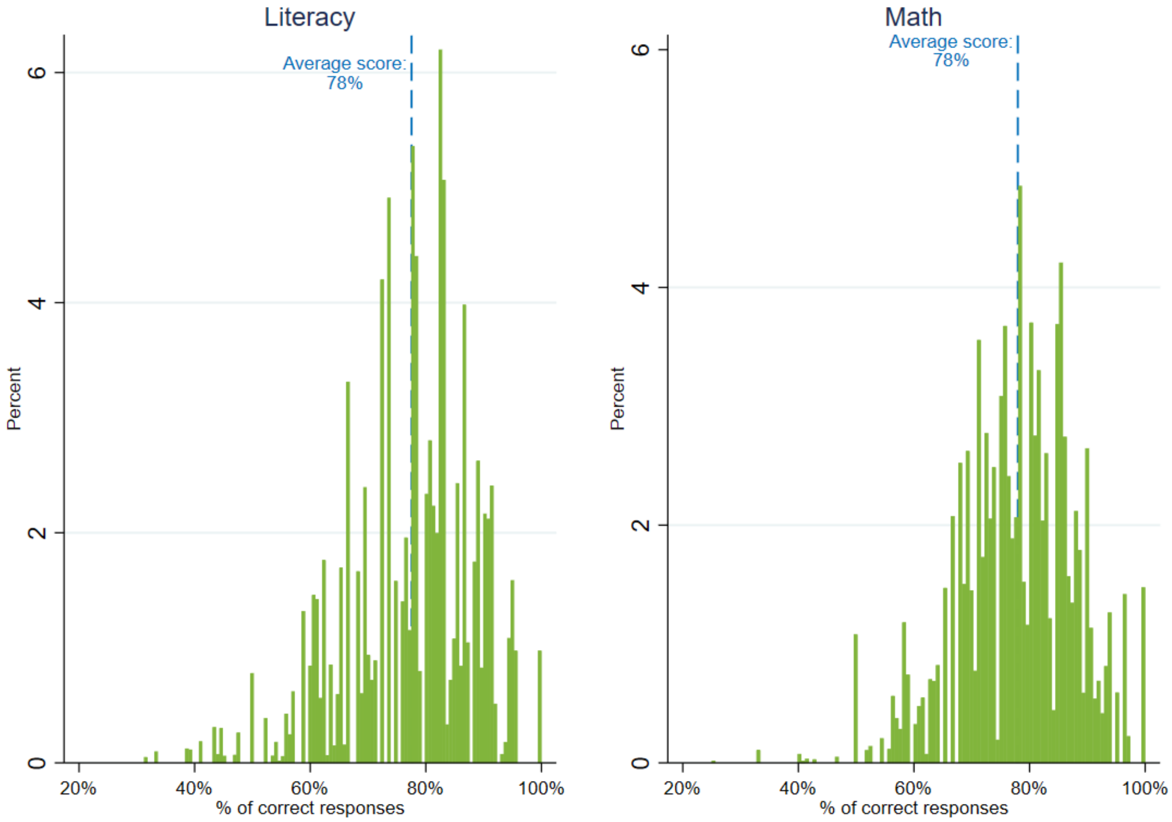


Figure 5: Performance in Oral Assessments by Subject



Geography

Average math test scores were relatively consistent across regions, varying from 44% in Margibi (both grades) to 59% in Montserrado (6th grade), with most counties averaging between 48 and 51% of correct responses. Students in Margibi county were predominantly found in rural schools and Montserrado students were almost all listed as urban. However, across all counties, there is no significant difference in math scores between rural and urban students.

Written 3rd grade literacy scores are somewhat more variable across counties, with a 23pp gap between Margibi's average of 59% and Montserrado average of 82%. The rural-urban gap is more pronounced as well, with urban students scoring on average 13pp higher in 3rd grade and 10pp higher in 6th grade.

Table 5. Performance in Math Written Assessments by County

Grade	County	N	mean	CI - low	CI - high	Std. Err.
Grade 3	Full Sample	1501	49%	46%	51%	1%
	Urban	878	49%	46%	53%	2%
	Rural	623	48%	46%	50%	1%
	Bong	157	48%	44%	52%	2%
	Gbarpolu	56	56%	51%	61%	2%
	Grand Bassa	70	50%	46%	54%	2%
	Grand Cape Mount	144	48%	45%	51%	2%
	Margibi	275	44%	42%	47%	1%
	Montserrado	158	59%	50%	68%	5%
	Nimba	641	48%	45%	51%	1%
Grade 6	FullSample	1595	49%	47%	51%	1%
	Urban	1058	49%	47%	51%	1%
	Rural	537	48%	45%	51%	1%
	Bong	109	51%	44%	57%	3%
	Gbarpolu	63	47%	39%	56%	4%
	Grand Bassa	80	51%	49%	54%	1%
	Grand Cape Mount	194	42%	39%	45%	2%
	Margibi	194	44%	40%	47%	2%
	Montserrado	127	52%	45%	59%	4%
	Nimba	828	49%	47%	51%	1%

Table 6: Performance in Literacy Written Assessments by County

Grade	County	N	mean	CI - low	CI - high	Std. Err.
Grade 3	FullSample	1501	72%	69%	75%	2%
	Urban	878	77%	74%	80%	1%
	Rural	623	64%	59%	68%	2%
	Bong	157	67%	60%	74%	4%
	Gbarpolu	56	73%	66%	80%	3%
	Grand Bassa	70	66%	60%	72%	3%
	Grand Cape Mount	144	64%	59%	69%	3%
	Margibi	275	59%	53%	65%	3%
	Montserrado	158	82%	75%	89%	3%
	Nimba	641	77%	75%	80%	1%
Grade 6	FullSample	1595	67%	64%	69%	1%
	Urban	1058	69%	66%	72%	2%
	Rural	537	59%	54%	65%	3%
	Bong	109	65%	55%	76%	5%
	Gbarpolu	63	58%	41%	75%	9%
	Grand Bassa	80	63%	58%	67%	2%
	Grand Cape Mount	194	56%	51%	61%	3%
	Margibi	194	51%	44%	58%	4%
	Montserrado	127	69%	59%	78%	5%
	Nimba	828	70%	67%	73%	2%

The 3rd-grade oral assessments also see relatively stable scores across regions. In the case of literacy, there is a 12-percentage-point difference between the lowest-scoring county, Margibi

(69%), and the highest-scoring county, Nimba (81%). However, most counties recorded an average between 75% and 78% in both subjects.

Table 7: Performance in Math and Literacy Oral Assessments by County

Subject	County	N	mean	CI - low	CI - high	Std. Err.
Literacy	FullSample	877	78%	76%	79%	1%
	Urban	434	79%	77%	81%	1%
	Rural	443	74%	72%	75%	1%
	Bong	133	75%	71%	78%	2%
	Gbarpolu	41	77%	76%	78%	1%
	Grand Bassa	48	75%	72%	77%	1%
	Grand Cape Mount	95	77%	74%	79%	1%
	Margibi	168	69%	65%	72%	2%
	Montserrado	69	77%	72%	82%	3%
	Nimba	323	81%	80%	82%	1%
Math	FullSample	877	78%	77%	79%	1%
	Urban	434	78%	77%	79%	1%
	Rural	443	78%	76%	79%	1%
	Bong	133	77%	74%	81%	2%
	Gbarpolu	41	81%	78%	84%	1%
	Grand Bassa	48	76%	73%	78%	1%
	Grand Cape Mount	95	76%	73%	79%	1%
	Margibi	168	78%	76%	80%	1%
	Montserrado	69	80%	78%	82%	1%
	Nimba	323	78%	76%	79%	1%

Gender

Male and female students received similar written scores in each subject and grade (Figure 6). Average scores for the oral assessment are nearly identical across gender and subject as well (Figure 7). In every case, the average gender gap is less than 3pp.

Figure 6. Average Written Scores by Gender, Subject, and Grade

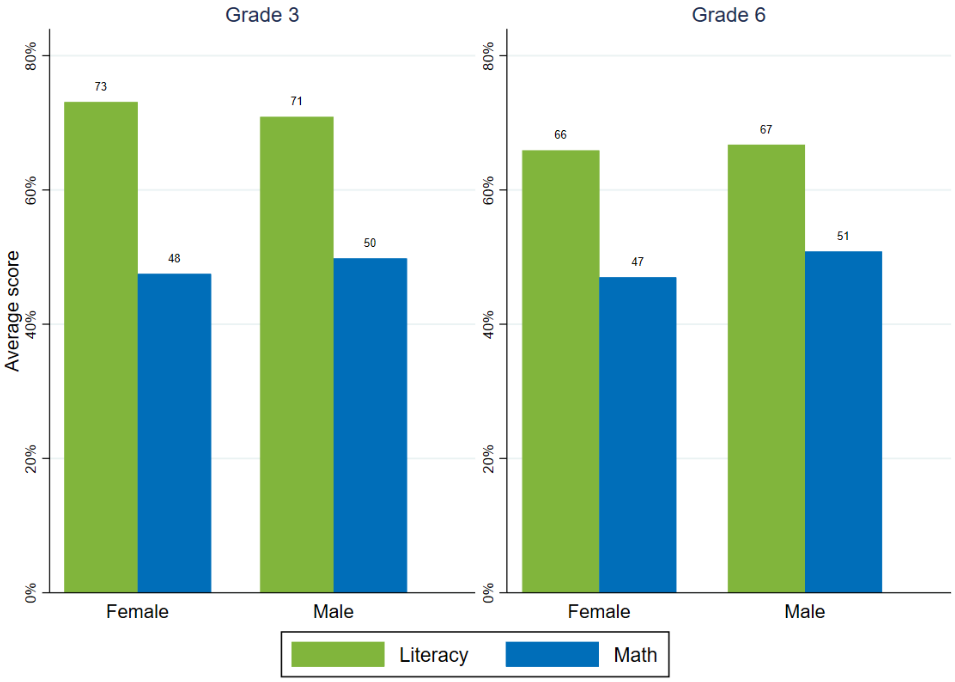
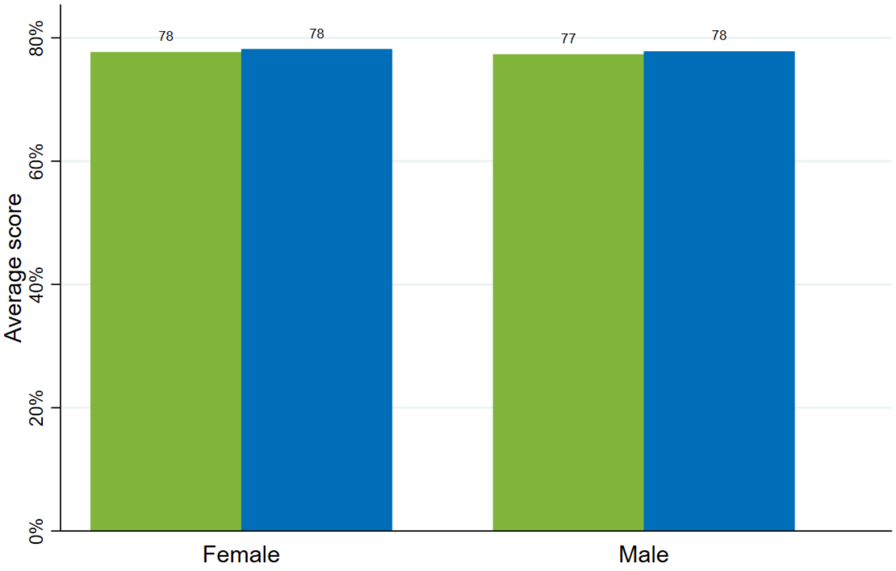


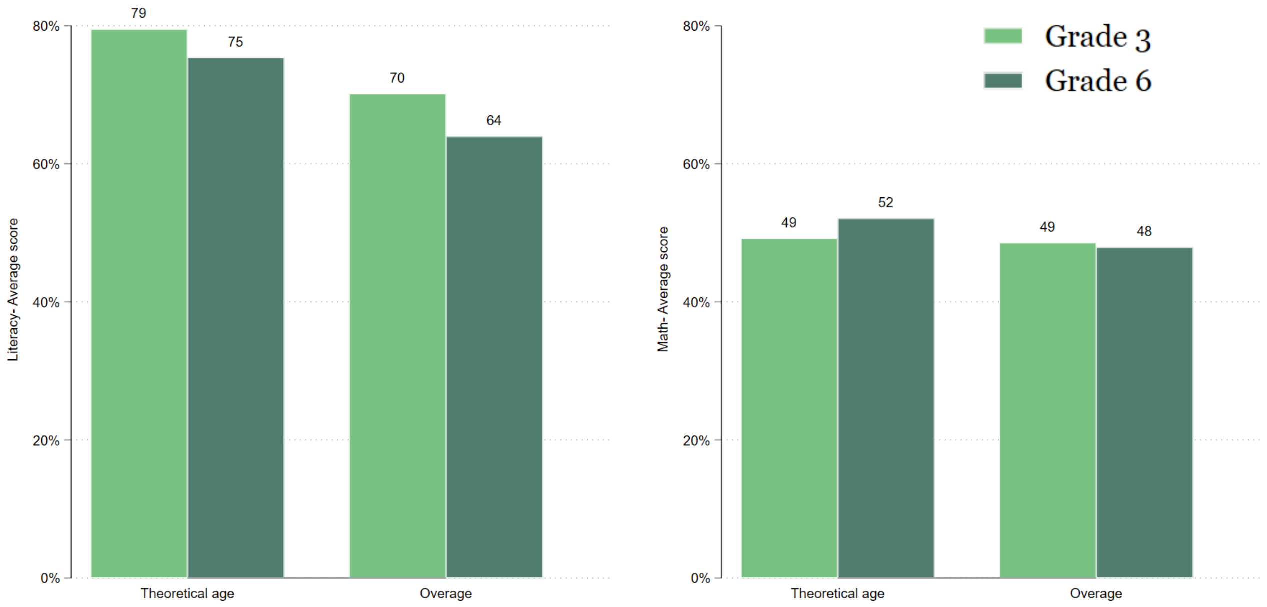
Figure 7. Average Oral Scores by Gender and Subject



Age group

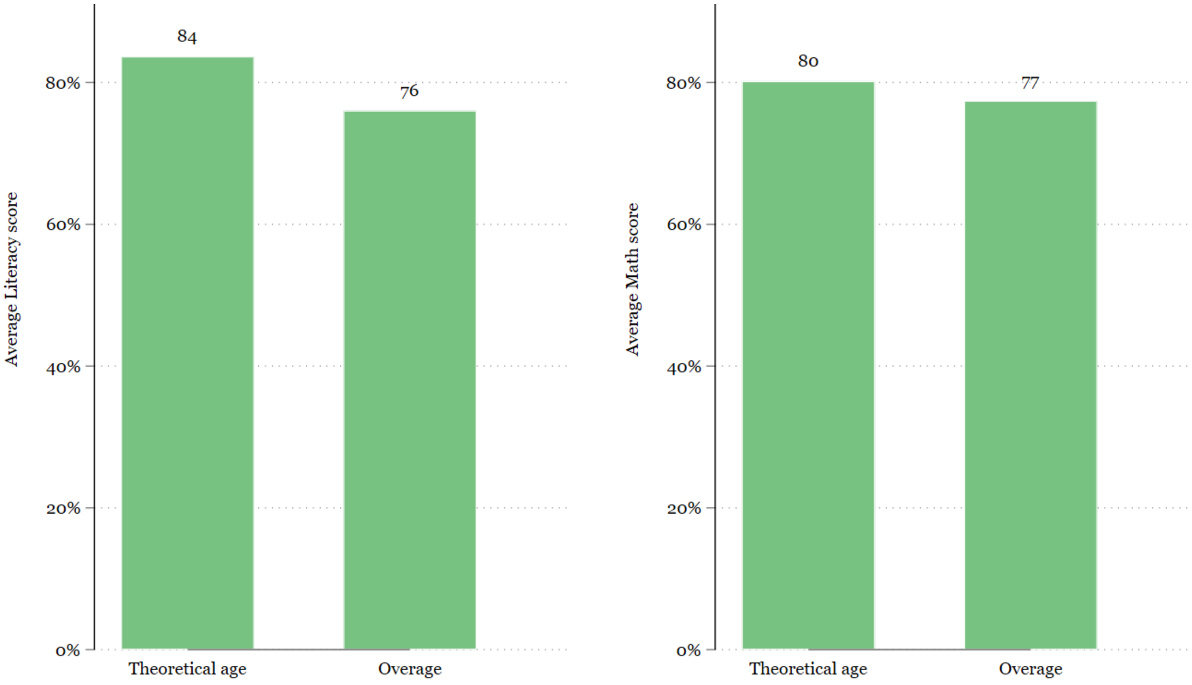
As discussed in the previous section, a significant portion of students in each grade are above the expected age based on Liberia's age of entry policy, raising the question of how assessment scores differ for those over the expected age for their grade. While the theoretical age is taken to be 8 and 11 for 3rd and 6th graders, respectively, we take a more flexible definition of 7-9 and 10-12, with those over that range defined as "overage". On average, overage students receive comparable math scores, with neither grade seeing a statistically significant difference. However, both grades see overage students receiving average scores 9pp lower on the written literacy assessment. In the case of the 3rd grade oral assessments, overage students received average scores 8pp and 3pp lower in literacy and math, respectively, albeit statistically insignificant for math.

Figure 8. Average Written Scores by Age, Subject, and Grade



***Theoretical age is considered to be 7-9 years old for grade 3 and 10-12 years old for grade 6.*

Figure 9. Average Oral Scores by Age and Subject



Test results by school type

Significant disparities in student average scores are observed based on their school type. As depicted in Figure 10, grade 3 students from public schools score between 12 and 15pp lower in the written literacy test compared to their peers in private, community, or faith-based schools. However, scores for the math written test show a narrower gap between public and private schools, with only a 3-percentage-point difference. The only case in which non-public schools underperform relative to public schools is for 6th graders in community schools, though it should be noted that only ten community schools were sampled and only in Nimba county, raising questions about the generalizability of that result.

Turning to the oral assessments, private and faith-based school students scored 8pp higher on average than public school students on the literacy assessment, but no better on the math assessment.

Figure 10. Average Written Scores by School type, Subject, and Grade

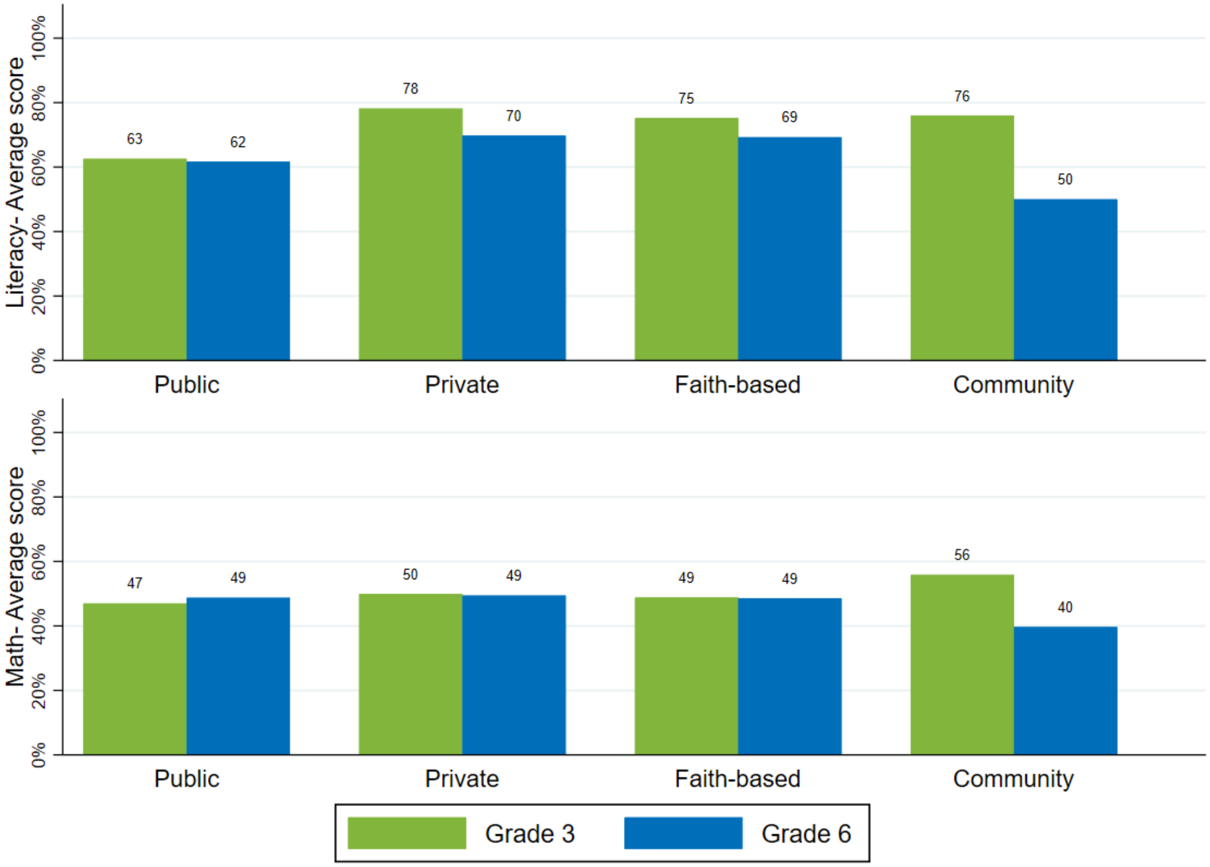
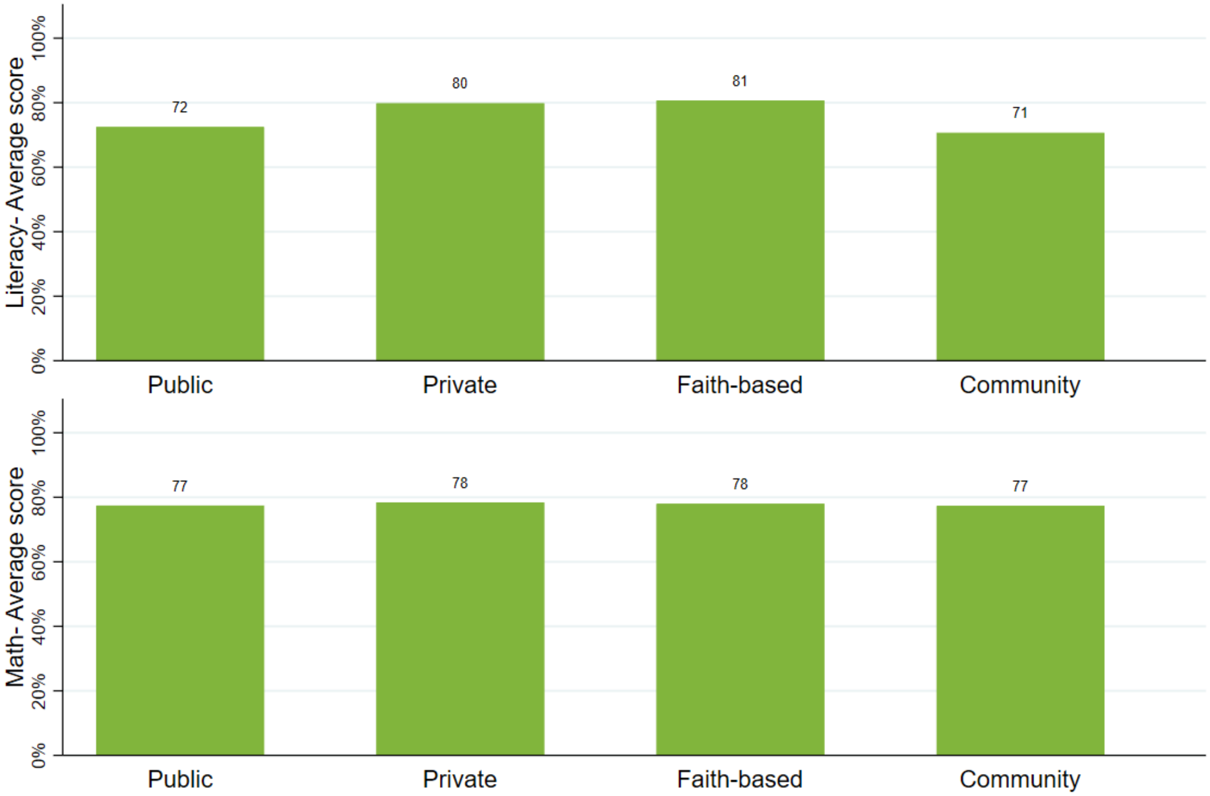


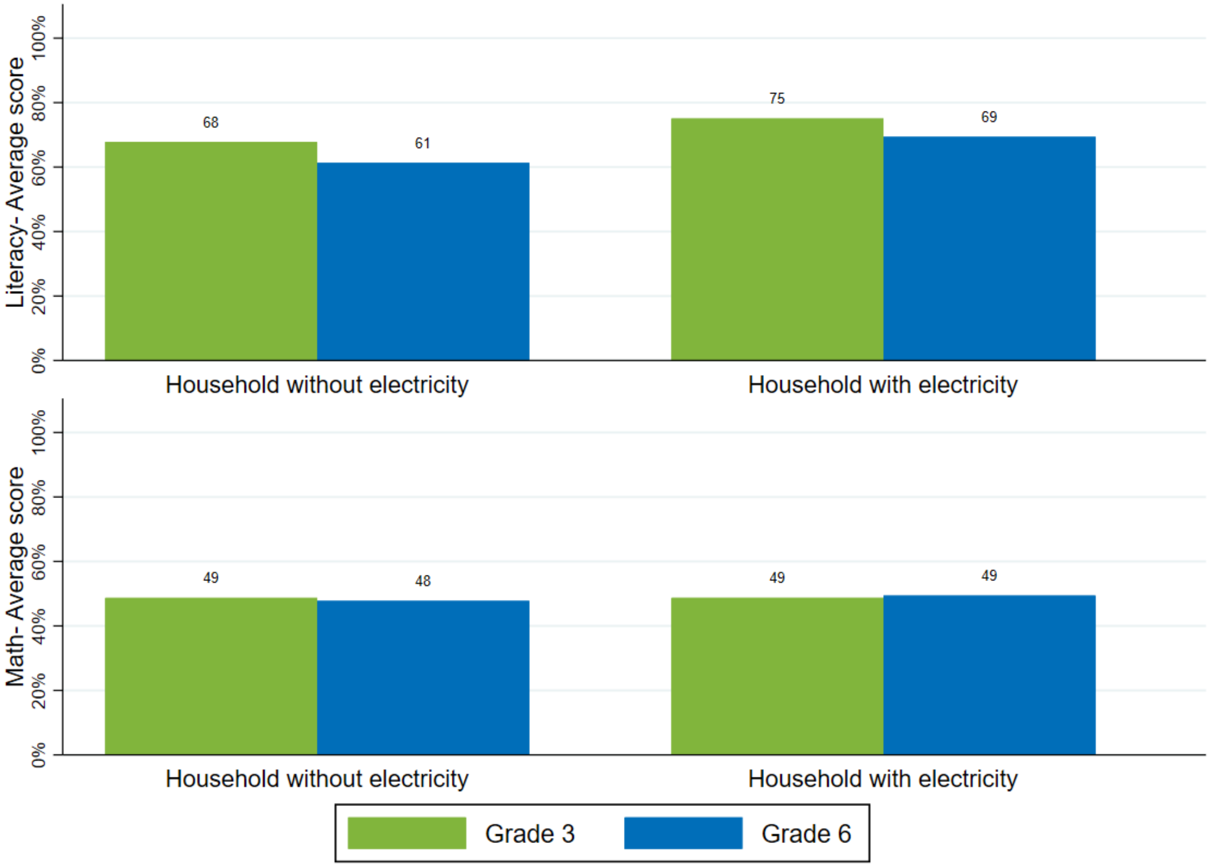
Figure 11. Average Oral Scores by School type, Subject, and Grade



Test results by household electrification

During the assessment, students were asked a short set of questions about dwelling characteristics and household assets that they were likely to be able to answer and which were potentially correlated with socioeconomic status. Among these, access to electricity at home proved most informative, and so is used here as a proxy for household living conditions. The absence of electricity in the household may also be a barrier to at-home study, making it particularly relevant to educational outcomes. While average written and oral math scores do not vary with electricity access, it is associated with 7pp higher scores on average on the written literacy exam. This holds true for both grade 3 and grade 6 and for written and oral tests.

Figure 12. Average Written Scores by Electricity presence, Subject, and Grade



Assessment Scores by Content Domain

When designing the written math and literacy assessments, each subject was separated into narrower content domains. The math subject covers three content domains. The “number” domain deals with basic arithmetic and recognition of numbers and patterns. The “measurement and geometry” domain covers the identification and analysis of geometric shapes and physical quantities. Finally, the “data” domain covers the reading, interpreting, and representation of data, along with using data to solve problems. The literacy subject covers two content domains: “pre-reading” and “reading comprehension”. Each content domain has a set of items that measure the students’ knowledge and skills in that domain.

While overall subject-level scores are a central focus of this report, a brief analysis of student performance at the domain-level highlights how policy makers can use more detailed and specific information about students’ strengths and weaknesses to identify areas of concern and design curricula to be receptive to the needs of Liberian students. By analyzing the content domain scores, the ministry of education can identify the gaps and challenges in students’ learning outcomes and design appropriate interventions and policies to address them.

Meanwhile, individual schools will be able to use domain-level insights to identify at-risk students and tailor instruction to specific identified needs.

Figure 13 presents national-level assessment averages across the three math content domains for both grades. While 3rd graders appear to have had equal scores for numeracy and data analysis, marginally more errors were made in measurement and geometry. Meanwhile, 6th graders were able to answer a significantly higher portion of questions around numeracy than measurement or data analysis. Figure 14 presents the same domain-level scores for the two literacy domains. While average scores are lower in 6th grade than 3rd for both domains, the gap in average scores between the two domains narrows significantly between the two grades.

While these averages are difficult to interpret without national benchmarks or proficiency standards to provide context, Figures 13 and 14 highlight how policymakers might use properly benchmarked domain-level assessment scores to identify specific areas in which students are failing to learn or improve. For example, the fall in data analysis scores between grades 3 and 6 raises the question of whether instruction in this area is inadequate.

Figure 13: Average math score by content domain

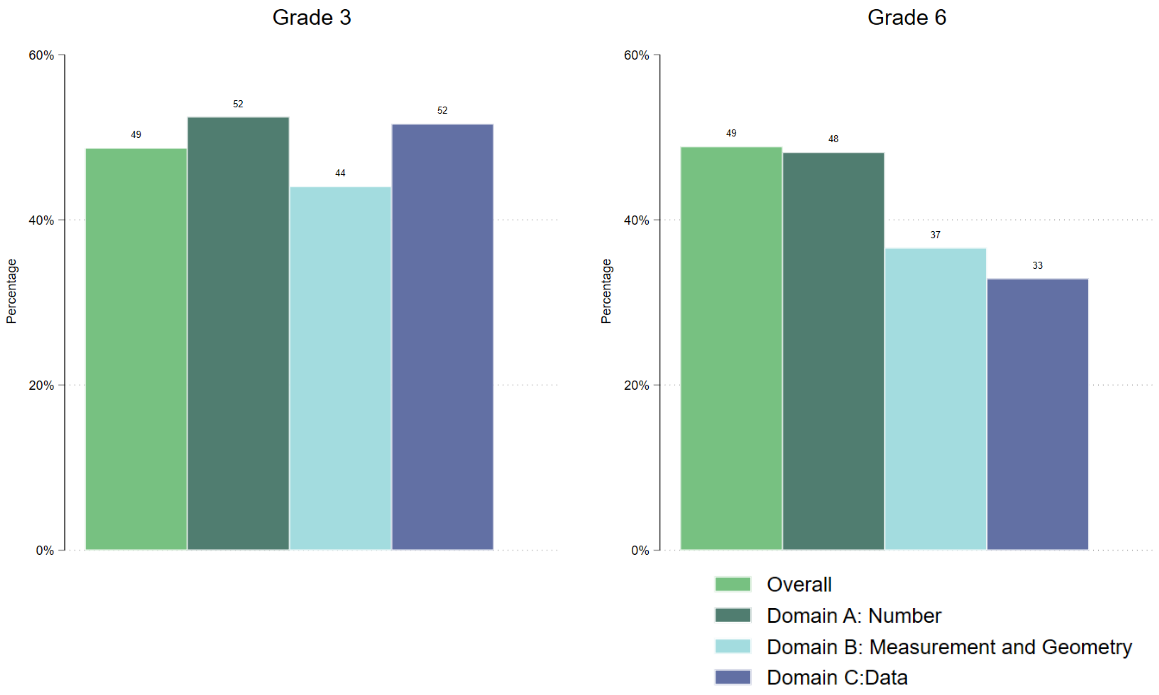


Figure 14: Average literacy score by content domain

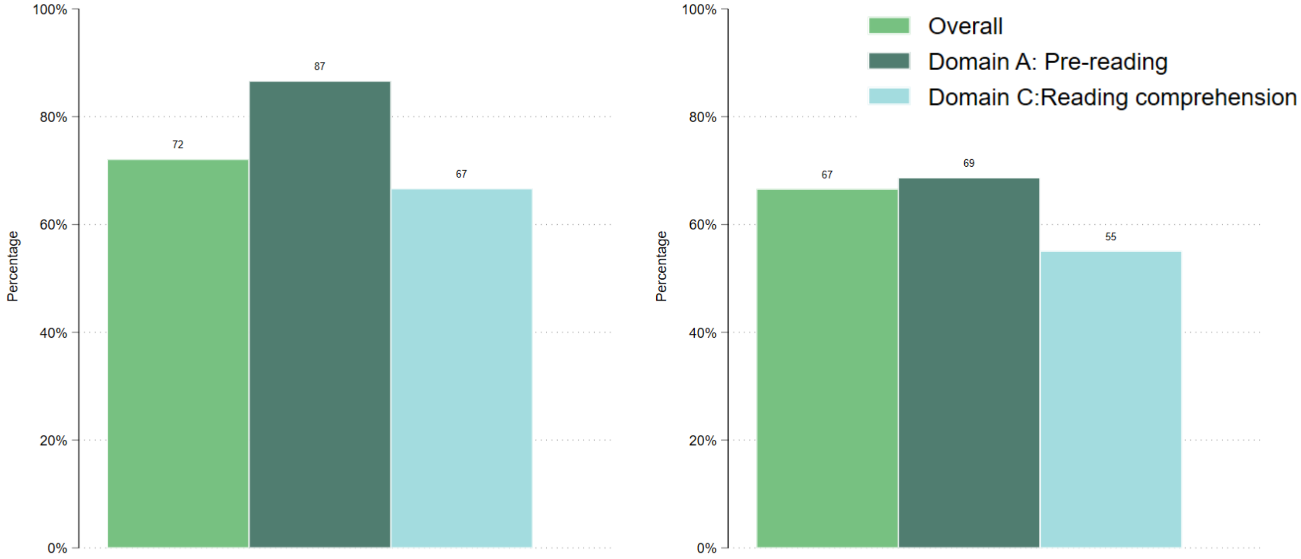


Table 8: Scores and number of questions by content domain and grade

Subject	Content domain	3rd Grade		6th Grade	
		Number of questions	**Average % Correct**	**Number of Questions**	**Average % Correct**
Math	**Overall**	**44**	**49%**	**50**	**49%**
	Domain A: Number	15	52%	17	48%
	Domain B: Measurement and Geometry	9	44%	10	37%
	Domain C: Data	1	52%	3	33%
Literacy	**Overall**	**39**	**72%**	**37**	**67%**
	Pre-reading skills	4	87%	18	69%
	Reading comprehension skills	21	67%	6	55%

Conclusions & Future Work

Proficiency Frameworks and Benchmarking

As described in the National Learning Assessment Policy, a primary objective for a national learning assessment is to provide relevant data on students' performance in relation to grade-level standards and benchmarks at both national and global levels. As Liberia has not yet finalized grade-level standards and benchmarks, our goal for the current assessments was to contribute to the on-going discussion on this topic at the national level.

Liberia's 2022-2023 Academic Calendar includes two proposed benchmarks by grade: a fluency benchmark and a reading comprehension benchmark. As fluency needs to be assessed in an oral exam, and the oral exam was conducted only with a subpopulation of third grade students, the current analysis cannot assess learning or growth against these benchmarks across grades. In the Academic Calendar, Grade 3 is assigned the Reading Fluency Benchmark of 45 correct words per minute and a Comprehension Benchmark of 60%. The target is a projected 10% of learners meeting these benchmarks.

Among 3rd graders who took the oral assessment, children were able to read on average 14 correct words per minute, ranging across counties from 19 correct words per minute in Margibi to only 5 words per minute in Nimba. In this sample, only 3.3% of students were able to read 45 words per minute or more, meeting the Reading Fluency target. It should be noted that while the Reading Fluency target is developed to measure fluencing in reading passage, this oral assessment relied on a familiar word grid of 45 words derived from USAID's Early Grade Reading Assessment. While no accurate words-per-minute estimate is available for the reading passages in the oral exam, preliminary analysis suggests that students read more fluently in these passages than in the familiar word grid. Meanwhile, 53% of the students achieved the Reading Comprehension target. These results are provided in detail in Appendix 3. As the Liberian curriculum continues to be defined and reformed, aligning the National Learning Assessment with those developments will be critical as well.

In addition to national standards and benchmarks, the National Learning Assessment was designed to align with international initiatives as well. Liberia's results can contribute to global goal-setting and monitoring of progress. As an illustration, we provide a few examples here of how the Phase II results can be interpreted in relation to the Global Proficiency Framework (GPF), an initiative developed by UNESCO Institute for Statistics and other actors to define for both reading and mathematics the minimal proficiency levels children are expected to obtain at the end of each of grades two through six.

As an example how how the GPF can be used to interpret performance in this exam, one can consider the reading domain "Reading Comprehension" and construct "Retrieve information", which contains the subconstruct "Locate explicitly stated information". The GPF identifies two skills for Grades 3 and 4: Locate prominently-stated information in two consecutive sentences (Grade 3) and Locate prominently-stated information within a single paragraph (Grade 4). In the third grade written literacy assessment, children were asked to reach a simple passage and answer four questions about that passage that concerned identifying prominently stated information. 53% of children were able to answer all four questions correctly. Performance on each of the four individual questions is illustrated by Figure 14 below.

The GPF for math identifies the skill: Demonstrate fluency with multiplication facts to 10 x 10 and related division facts for Grade 4 as part of the Domain: Number knowledge, Construct: Operations, and Subconstruct: Multiply & divide quantities concretely & symbolically. In our sixth grade assessment, 66 percent of 6th graders were able to answer more than half of multiplication & division questions correctly, illustrated in Figure 15 below.

Figure 15. Third Grade Performance on Reading Comprehension Passage

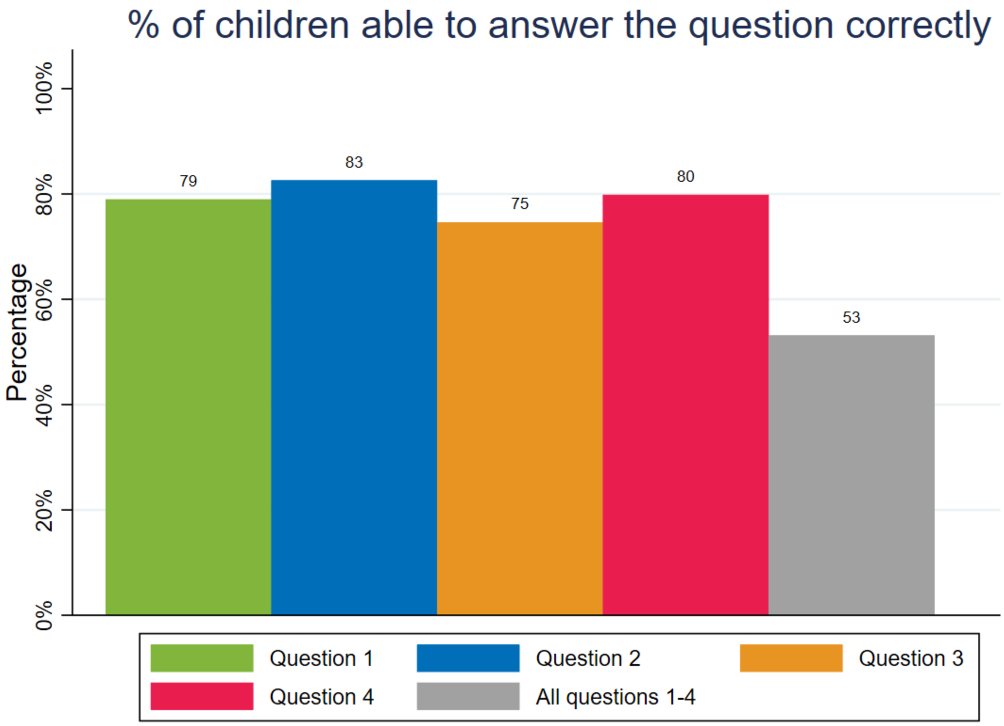
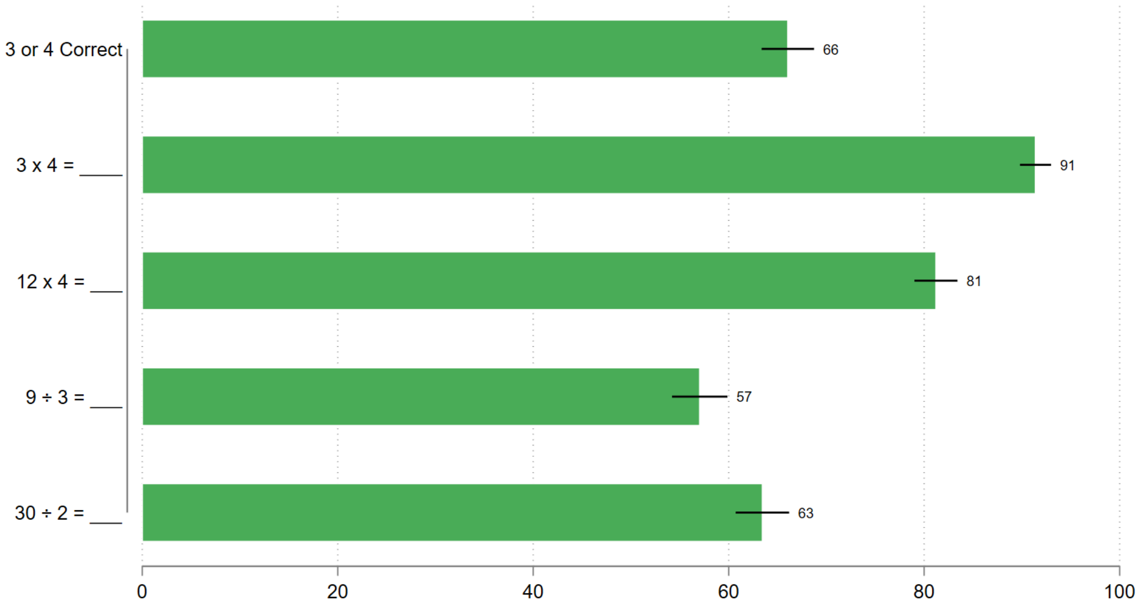


Figure 16. Sixth Grade Performance on Simple Multiplication and Division Items



Moving forward, Liberia’s National Learning Assessment can both inform and be responsive to global initiatives to define benchmarks and standards and monitor progress.

Conclusions & Next Steps

Following up on a pilot assessment in phase one of the National Learning Assessment Policy (NLAP) framework, this study implemented improved math and literacy assessments among a well-balanced sample of primary school students across eight counties. A psychometric analysis of the pilot exam in Phase 1 presented several opportunities to improve on the design of the assessments. Subsequent psychometric analysis suggests that for both grades and both subjects, the test design and more diverse sample of students yielded significantly more informative results in Phase 2, improving the reliability of test questions and their ability to distinguish high-performing students from median and low-performing students. Analysis of the test results highlights the potential value of these national student assessment exercises while also suggesting ways to build on this experience to provide important and actionable insights to educators and policy makers moving forward.

At the national and regional levels, this report provided aggregated comparative analysis of subgroups of students across demographic and socioeconomic dimensions. Insights into which types of students are doing well and which are falling behind offer researchers and government officials a nuanced evidence-based perspective on where to focus efforts to support learning at the national level. Looking in more detail at assessment results, the report then separates the math and literacy subjects into five content domains, which define the general categories of skills in which students should be learning and expanding competency. These domain-level results may prove informative during curriculum development, while even more granular subdomain and item-level analysis can inform specific school officials and teachers about where instruction and learning can be improved.

While these statistics are themselves informative, a crucial next step for interpreting assessment results is the development of specific proficiency benchmarks and learning standards. National standards will be valuable for comparing performance across regions or other student subgroups, while international standards will be valuable for comparing Liberia's progress to those of other West African nations. Perhaps most important, analysis of student proficiency across grade levels and over time allow for concrete assessment of the learning process and which students are being adequately served or left behind.

The need to measuring student learning leads as well to the report's final recommendation: improved and continued data collection. The psychometric item-level analysis of the Phase 2 instrument in the Appendix offers specific recommendations for how to improve the questions asked of students to ensure highly reliable results that clearly separate students by ability. Finally, continued collection of assessment data that is comparable with this Phase 2 data, and covering with a wider array of students in all Liberian counties, will offer consistent, timely, and actionable insights to inform curriculum design and opportunities to improve learning among primary school students.

Appendices

Appendix 1. Definitions for statistical calculations

N: The number of observations included in the group
Mean: The average outcome for the number of observations in the group
CI-low: The lower bound of the 95% confidence interval.
CI-high: The upper bound of the 95% confidence interval.
Std Err.: Standard error.

Appendix 2. Disaggregated Math Oral test scores

Indicator	Group	N	mean	CI - low	CI - high	Std. Err.
Addition score (number of correct responses out of 4)	FullSample	845	2.76	2.68	2.85	0.04
	Urban	412	2.81	2.69	2.93	0.06
	Rural	433	2.67	2.57	2.76	0.05
	Bong	130	2.61	2.44	2.78	0.09
	Gbarpolu	41	3.05	2.75	3.34	0.15
	Grand Bassa	46	2.53	2.39	2.67	0.07
	Grand Cape Mount	92	2.89	2.68	3.11	0.11
	Margibi	161	2.50	2.40	2.61	0.05
	Montserrado	68	2.82	2.56	3.09	0.13
	Nimba	307	2.87	2.74	3.00	0.07
Subtraction score (number of correct responses out of 4)	FullSample	780	2.80	2.69	2.92	0.06
	Urban	391	2.76	2.61	2.91	0.08
	Rural	389	2.90	2.75	3.05	0.08
	Bong	120	2.80	2.55	3.05	0.13
	Gbarpolu	40	2.73	2.38	3.07	0.17

	Grand Bassa	45	3.01	2.69	3.33	0.16
	Grand Cape Mount	83	2.73	2.35	3.11	0.19
	Margibi	134	2.86	2.65	3.06	0.10
	Montserrado	62	2.88	2.40	3.37	0.25
	Nimba	297	2.76	2.58	2.93	0.09
Multiplication score (number of correct responses out of 4)	FullSample	561	1.18	1.06	1.30	0.06
	Urban	273	1.20	1.03	1.37	0.09
	Rural	288	1.14	0.99	1.29	0.08
	Bong	88	1.07	0.81	1.33	0.13
	Gbarpolu	29	1.41	0.80	2.02	0.31
	Grand Bassa	37	0.88	0.81	0.95	0.04
	Grand Cape Mount	63	1.12	0.65	1.58	0.24
	Margibi	85	1.05	0.84	1.27	0.11
	Montserrado	44	1.75	1.40	2.10	0.18
	Nimba	215	1.17	1.01	1.34	0.08
Division score (number of correct responses out of 4)	FullSample	289	1.53	1.37	1.69	0.08
	Urban	133	1.52	1.30	1.74	0.11
	Rural	156	1.55	1.33	1.76	0.11
	Bong	50	1.59	1.27	1.92	0.16
	Gbarpolu	22	1.81	0.91	2.71	0.45
	Grand Bassa	21	1.31	0.78	1.84	0.27
	Grand Cape Mount	22	2.14	1.51	2.77	0.32
	Margibi	43	1.42	1.21	1.63	0.11
	Montserrado	24	1.83	1.28	2.38	0.28

	Nimba	107	1.49	1.25	1.74	0.12

Appendix 3. Disaggregated Literacy Oral test scores

Indicator	group	N	mean	CI - low	CI - high	Std. Err.
Number of correct letters per minute	FullSample	877	79	76	81	1
	Urban	434	80	77	83	2
	Rural	443	76	73	79	1
	Bong	133	75	69	80	3
	Gbarpolu	41	82	76	88	3
	Grand Bassa	48	82	78	85	2
	Grand Cape Mount	95	75	70	80	3
	Nimba	168	68	64	72	2
	Margibi	69	87	81	92	3
	Montserrado	323	81	77	84	2
Number of correct words per minute	FullSample	715	13	10	15	1
	Urban	434	15	12	17	1

	Rural	443	8	7	10	1
	Bong	133	7	5	10	1
	Gbarpolu	41	12	10	13	1
	Grand Bassa	48	10	7	12	1
	Grand Cape Mount	95	12	9	15	2
	Nimba	168	5	4	6	0
	Margibi	69	19	12	27	4
	Montserrado	323	15	12	18	2
% of students who were able to read 45 words per minute or more	FullSample	715	3.3%	1.3%	5.3%	1.0%
	Urban	434	4.4%	1.6%	7.2%	1.4%
	Rural	443	0.9%	-0.5%	2.3%	0.7%
	Bong	133	0.3%	-0.3%	1.0%	0.3%
	Gbarpolu	41	0.0%	.	.	.
	Grand Bassa	48	0.0%	.	.	.
	Grand Cape Mount	95	1.9%	-1.8%	5.5%	1.8%
	Nimba	168	0.0%	.	.	.

	Margibi	69	9.6%	3.1%	16.0%	3.2%
	Montserrado	323	4.3%	1.1%	7.5%	1.6%
Average score for reading comprehension task 5 (% of correct answers)	FullSample	509	53%	47%	59%	3%
	Urban	277	56%	50%	62%	3%
	Rural	232	45%	35%	55%	5%
	Bong	75	45%	29%	62%	8%
	Gbarpolu	28	57%	49%	65%	4%
	Grand Bassa	25	25%	6%	44%	10%
	Grand Cape Mount	59	58%	50%	65%	4%
	Nimba	53	46%	35%	56%	5%
	Margibi	55	42%	26%	58%	8%
	Montserrado	214	60%	54%	66%	3%
% students who were able to answer 60% of the reading comprehension questions correctly or more	FullSample	509	51%	42%	59%	4%
	Urban	277	56%	46%	66%	5%
	Rural	232	38%	27%	49%	6%
	Bong	75	38%	24%	52%	7%

	Gbarpolu	28	57%	50%	64%	4%
	Grand Bassa	25	15%	-6%	36%	11%
	Grand Cape Mount	59	59%	48%	69%	5%
	Nimba	53	33%	19%	48%	7%
	Margibi	55	32%	10%	55%	11%
	Montserrado	214	63%	54%	72%	5%

Appendix 4. Test and Question Reliability

Stage 2 Grade 3 Maths Test - Rasch Model

1. Stage 2 Summary Statistics – Rasch Model – Reliability and Separation

Items	The 44 items have a reliability of 1.00 and create 22 levels of separation
Persons	The reliability is 0.74 and creates 1.72 levels of separation.

Commentary	The test in stage 2 has performed better with this sample than the test used in stage 1 and the related sample. The revised instrument has an increased person reliability but has not met the minimally acceptable criteria of 0.80 and as a result is still not able to separate candidates into 2 levels. However, the increase is noticeable from 0.58 to 0.74. The lower-than-expected person reliability is likely due to the narrow range of ability present in the stage 2 sample. Although the sample is 3 times larger than the stage 1 sample, the current sample is mainly clustered between -1.5 to 1.5 logits, with over 55% of the candidates scoring around the 0 logits. This would lead to lower variance and bring down the reliability estimate.
	The items have perfect reliability and have nearly twice the separation than the previous stage 1 test.

2. Item Fit – Rasch model

** Fit ranges of 0.8 to 1.2 as acceptable for measurement as used in stage 1 analysis*

** For stage 2 sample size of 1500 candidates, it is recommended to use fit ranges 0.95 to 1.05. However, the more liberal range used in stage 1 was applied again as the assessment is still in early stages. Once a larger bank of items has been created, more conservative ranges can be applied.*

Problematic Items	**Item 23**
	4 items were identified as misfitting
	All items were towards the extreme ability ends of the latent continuum
Misfitting Items Outfit	**3, 22, 23, 33, 42**
Misfitting Items **Infit**	**23**
Item polarity	**22, 23, 31, 33, 42**

Commentary	**Item 3** This is an easy item was had several lower ability candidates answer incorrectly unexpectedly.
	ITEM 22 This item was the 5th hardest item and was answered correctly by 9% of the sample. 69% of persons chose option C. This was a rectangle shape. This item did not have a noticeable number of unexpected correct answers. It is likely the issue is with the item prompts. The images are blurred, and the correct answer D does appear to have slightly rounded edges, especially on the left of the shape. **This item needs revising.**
	ITEM 23 This item had both infit and outfit misfit. This item has 1.3 logit difficulty. The mean person measure was -.12 logits. Most candidates had a 25% chance of success on this item. Should the item be this difficult for the sample? It is likely the wording in the prompt is ambiguous. 67% of the sample chose option A which is 156 or from the prompt 28 and 128 together. The use of the word ‘sum’ in the prompt likely caused this. **This item needs revising as it was on the border of acceptable infit or remove.**
	ITEM 33 This item had several unexpected correct answers from lower ability persons. This item has 1.6 logit difficulty. The mean person measure was -.12 logits. Most candidates had a 15% chance of success on this item. Should the item be this difficult for the sample? It is likely the wording in the prompt is ambiguous. 57% of the sample chose option A which is are the numbers in the prompts. Additionally, the words have been mis ordered for the actual number they represent. **This item needs revising as it was on the border of acceptable infit.**
	Item 42 This item has a high difficulty of 2.5 Logits. Should this item be this difficult for the sample? 50% of persons chose option C, and 29% chose option A. Are the distractors too appealing to the candidates

	because of the images used in the prompts? **This item needs revising**

Discrimination – Rasch empirical discrimination statistics

Item discrimination	**Item 21 violates discrimination requirements**
	Despite the item difficulty falling below the mean person ability and relatively easy it did not discriminate well between high and low candidates. This item had significant infit and overfit but not above the 1.2 criteria used. However, the ZSTD values were significant. While these are typically affected by sample size, given the seemingly simple content of the item, this item is behaving unusually. This item is qualitatively different from the other items in the test. The image is also blurred with a similar proportion of persons choosing the 3 incorrect distractors (11%, 15%, 12%). **Remove this item.**

3. Item Coverage

Item Coverage	· Mean item and person logits were very close. · The spread of items appears well targeted for the sample. Most of the items matched ability present.
	· There were 5 items above the most able candidate with the most difficult item [item 31] around 2 logits higher than the most able persons who would have only 12% chance of success. Such an item could be repurposed. · Less than 10 persons fell below the easiest item. These candidates are extreme and are around 2 S.D. away from the person and items means. Thus, they may not represent the target sample. · There are some gaps within 1 S.D. of mean person ability that require items. This may improve person reliability estimates.

4. Unidimensionality and local independence - Rasch model

Item Unidimensionality	Unexplained variance in first contrast has a 2.35 Eigen value. This exceeds the recommended limit of 2 – meaning there are roughly two items in the set that are multidimensional. On investigation of the Standardized residual loadings for item, items 28 and 30 were identified as above the recommended limit of 0.40. These items were also identified below as having local dependence on each other. Once removed, the Eigen Value fell to 2.10. On further inspection items 8 and 13 had unusually high loadings. This is surprising as these were simple arithmetic sums. However, once these items were also removed the Eigen value fell to 1.86 – well below the threshold. **Recommend remove items 8,13,28,30.**
Local independence of items	**Items 28 and 30**

See appendix 4 – Table 5 and 6

ITEMS and Gender DIF	No items met substantial thresholds to be a threat to measurement

5. Effect on item removal – rash model

4 options were modelled to assess the impact on item and person reliability, while maintaining unidimensionality. The effect on DIF in four areas were also considered.

	Item reliability	**Person Reliability**	**Separation**	**Unidimensionalit y and Independence**	**DIF - Gender**

Removing: **High infit/outfit [23]** **Local dependence [8,13,28,30]** **Low Disc [21]** **Too difficult [31]** **TOTAL Items: 37**	1.00	0.73	1.63	Yes / yes	No

6. Recommendations

Recommend remove	1. To improve item fit: remove item23– threat to measurement. 2. To improve dimensionality: remove items 8,13,28,30 – unclear why from item content. 3. To allow replacement of items for better measurement of central ability candidates if this is aim of test: removing item 31. 4. Remove item 21 as poor discrimination and questionable content.
Investigate	Items 22,33,42 for content/prompt issues.
Suggestion	To increase person reliability: 1. Wider range of person ability or more clearly define target persons i.e. identify minimally acceptable candidates and target this area of the latent continuum. 2. Increase number of items to 50. 3. Possible consider polytomous items – items with several categories.

Stage 2 Grade 3 Written Test - Rasch Model

1. Stage 2 Summary Statistics – Rasch Model – Reliability and Separation

Items	The 44 items have a reliability of 1.00 and create 16 levels within the domain
Persons	The reliability is 0.83 which creates separation of 2 levels
Commentary	The test in stage 2 has performed better with this sample than the test used in stage 1 and the related sample. The revised instrument has an increased person reliability above the minimally acceptable criteria of 0.80 and as a result has in this sample been able to separate candidates into 2 levels. The items have perfect reliability and have been separated across 16 levels. This is twice more than the previous stage 1 test which achieved 8 levels of separation.

2. Item Fit – Rasch model

** Fit ranges of 0.8 to 1.2 as acceptable for measurement as used in stage 1 analysis*

** For stage 2 sample size of 1500 candidates, it is recommended to use fit ranges 0.95 to 1.05. However, the more liberal range used in stage 1 was applied again as the assessment is still in early stages. Once a larger bank of items has been created, more conservative ranges can be applied.*

Problematic Items	**Item 31**
	4 items were identified as misfitting
	All items were towards the extreme ability ends of the latent continuum
Misfitting Items Outfit	**31 / 15 / 5 / 6 -** a result of candidates of differing ability to the items getting them either unexpectedly correct or incorrect
Misfitting Items	**31** – this is a result of persons at the item ability level unexpectedly getting the item incorrect

Infit	
Item polarity	**5, 6,** and **15** have a low polarity.
	This is likely caused by high candidates getting items 5 and 6 unexpectedly wrong and low candidates getting item 15 correct.
Commentary	**ITEMS 5 and 6** Items 5 and 6 likely have a high outfit due to carelessness of higher candidates. On inspection of the misfitting persons to the item, high ability candidates were unexpectedly getting the item wrong. This is further supported by using distractor analysis. While the items were written responses to oral input and not MCQ items, the varied given answer responses were analysed and found that 11 high candidates wrote the wrong answer in comparison to 1462 weaker candidates who were of lower ability who answered correctly. In addition, the standardized residual correlations for items 5 and 6 appear to be more highly correlated than all other items suggesting a violation of item independence. The value of 0.49 falls just below the recommended limit of 0.50. An alternative explanation could be that test taking conditions. The candidates were required to write down single letters of the alphabet which were read to them by examination staff. Were letters repeated and equal number of times across all administrations. Could all candidates hear the prompt. The low correlations of 0.07 and 0.10 could indicate a dimension other than that intended to be assessed.
	ITEM 15 This item was the hardest item on the test and saw many unexpected correct responses by weaker candidates. This explains the high outfit. However, this item also had a low pt-measure correlation of 0.12 which indicates there may be another dimension to the item other than literacy.

	ITEM 31 This item had both high infit and outfit. The biggest concern is the infit since the item persons unexpectedly getting the item incorrect were at the same ability as the item. This likely the result of the item content or prompt wording or answer choices. **This item needs revising.**

Discrimination – Rasch empirical discrimination statistics

Item discrimination	**Item 31 violates discrimination requirements**
	9 items were identified as under discriminating using Pearson's Correlation using ranges >0.3 and <0.7 as acceptable. Items 5, 6, and 31 had correlation values between 0.08 to 0.11 which are substantially below the minimum of 0.30. However, it should be noted that items 5 and 6 were extremely easy items and showed high candidate carelessness with high outfit estimates, thus explains why these figures are so extreme.
	However, using a Rasch empirical discrimination statistic, this enabled triangulation of several data points to identify 3 items that may degrade measurement. Items 23, 24, and 31 showed low levels of discrimination. All items had much lower discrimination slopes and were shown to be substantial with lower asymptotes > 0.10. However, only item 31 showed misfit at 1.3 infit. When taken together, item 31 is a threat to measurement.

3. Item Coverage

Item Coverage	· The mean person ability is 1.43 logits (0.3 SEM) higher than the mean item ability. · With this sample there is a mismatch between item difficulty and person ability. · The test was too easy for overall sample of persons if test conditions and scores are to be accepted as genuine. · Item coverage for the test in stage 1 was better than this set of items in stage 2 for the samples the tests were administered to. · Item error was greatest for lower difficulty items as very few candidates scored at this ability. · Person error was greatest amongst higher ability candidates as few items targeted these candidates.
	· Item precision is highest below mean person ability. · There is good item coverage for lower ability persons. · Potentially, redundant items around mean item difficulty. · Items around -2.5 logits meet very weakest persons who have a 10% chance of success at items at mean ability.

Commentary	If the aim of the test is to measure across the spectrum of ability then many more harder items are required if this sample is a more accurate reflection of candidate ability. A mean person ability of 1. 43 logits means that the average candidate has am 80% probability of success on items lower than their ability. Only 5 items in the test were had higher difficulties than the average candidate. This explains why the most misfitting persons were generally the more able candidates. To improve item coverage, for example candidates who score at -1.4 logits have a 20% chance of success on items of mean difficulty. Any items below this may be better repurposed for higher level candidates if the aim is to keep the test around 40 items in length and measure the full spectrum of ability. However, if the aim is to measure around the cut point of minimally acceptable literacy ability, then the item coverage is suitable for the purpose and the misfit for higher candidates is to be expected as the essential accuracy and score precision around lower abilities is met with this test and 44 items. For example, 7 items targeted only a maximum of 8 persons of the 1500 test takers. Items 4,5,6,7, 12, 20, 36. These items could be repurposed. The average person has a 98% chance of success on item 5, for example. This contributes very little to measurement. A clearer definition of test purpose and expected candidates is required.

4. Unidimensionality and local independence - Rasch model

Item Unidimensionality	Unexplained variance in first contrast less than 2 indicates likely unidimensional measure of literacy.
Local independence of items	No items reached substantial thresholds to be of concern.

ITEMS and Gender DIF	No items met substantial thresholds to be a threat to measurement

5. Effect on item removal – rash model

4 options were modelled to assess the impact on item and person reliability, while maintaining unidimensionality. The effect on DIF in four areas were also considered.

	Item reliability	**Person Reliability**	**Separation**	**Unidimensionality and Independence**	**DIF - Gender**
Removing Q31 **TOTAL Items: 43**	1.00	0.83	2.2	Yes / yes	No
Removing - **Q31** - **easiest items** **[4,5,6,7,20]** **TOTAL Items: 38**	1.00	0.83	2.2	Yes / yes	No

Removing: **- Q31** **- Easiest [4,5,6,7,20]** **- Items out of synch with cluster** **[12,17,36,44,44]** **TOTAL Items: 33**	1.00	0.83	2.2	Yes / yes Removing items 12 and 43 reduced loading on 1st contrast – improving unidimensionality	No
Removing: **- Q31** **- Easiest** **- Items out of sync with cluster** **- Lowest discriminating [23,24]** **TOTAL Items: 31**	1.00	0.83	2.2	Yes / yes	No

6. Recommendations

Recommend remove	1. To improve item fit: remove item 31 – threat to measurement. 2. To improve dimensionality: remove items 12 + 43 – unclear why from item content. 3. To allow replacement of items for better measurement of higher ability candidates if this is aim of test: removing 13 items mentioned in report sections above does not affect overall measurement of persons.

Investigate	Item 15 – why so difficult Item 12 – what additional dimension might this item have? Items 17 and 43 – why are these items much Harder than the other items in their curriculum objectives? Item 12 was much easier than the other picture prompt items. Why? Have the items located on the difficulty continuum as expected?
Suggestion	Are items 4 to 6 necessary? How can the test conditions of these items be assured to be consistent? Item 8 was much more difficult than item 7 – cat vs light written response. There is almost a 3-logit difference in difficulty. A 3-logit gap from a person ability can reduce a person's chance of success on the item to about 5%.

Stage 2 Grade 6 Maths Test - Rasch Model

1. Stage 2 Summary Statistics – Rasch Model – Reliability and Separation

Persons (original)	The reliability is 0.78 and creates 1.88 levels of separation.
Commentary	The test in stage 2 has performed better with this sample than the test used in stage 1 and the related sample. The revised instrument has an increased person reliability but has not met the minimally acceptable criteria of 0.80 and as a result is still not able to separate candidates into 2 levels. However, the increase is noticeable from 0.54 to 0.78. The lower-than-expected person reliability is likely due to the interaction between test item difficulty, which for 23 items are of higher difficulty than most of the persons, and the candidates' strategies to accommodate this discrepancy given the older age of the G6 students than in G3. A large amount of unexpected correct answers is observed

	in the data for candidates of lower ability than the difficulty of the items they answered correctly on.
Items **(original)**	The 50 items have a reliability of 1.00 and create 22 levels within the domain
Commentary	The test in stage 2 has performed better with this sample than the test used in stage 1 and the related sample.
Persons (Cutlo)	Once the cutlo method had been applied to reduce the impact of guessing: **The reliability is 0.83 and creates 2.18 levels of separation.**
Commentary	To reduce the impact of guessing the cutlo method was used to better estimate item quality and person reliability. This method eliminates off-target responses where persons have a low expectation of success. The criteria used was 1 logit. This means that responses by candidates with 1 logit difference between their ability and the item difficulty were removed from the analysis. This amounts to persons with a 25% chance of success for items 1 logit above their ability. The item/person targeting will be more accurate and truer person and item ability is provided.
Items **(Cutlo)**	The 50 items have a reliability of .97 and create 6.21 levels within the domain
Commentary	The item reliability and item separation index have fallen, but are still well above minimally acceptable criteria.

2. Item Fit – Rasch model

*** Item fit information is based on the application of cutlo= -1 logit.**

** Fit ranges of 0.8 to 1.2 as acceptable for measurement as used in stage 1 analysis*

** For stage 2 sample size of 1500 candidates, it is recommended to use fit ranges 0.95 to 1.05. However, the more liberal range used in stage 1 was applied again as the assessment is still in early stages. Once a larger bank of items has been created, more conservative ranges can be applied.*

Problematic Items	**28, 50,**
Misfitting Items Outfit	**1,14**
Misfitting Items **Infit**	**No items to report – items targeting at a person of a similar level are psychometrically sound**
Item polarity	**31 –** negative correlation to other items
Commentary	**ITEM 50** is so difficult that no person was within 1 logit of the item difficulty thus no candidate response was available after applying cutlo. This item is simply too difficult for this sample and no measure was provided. **Remove item**
	ITEM 28 in the original data set before cutlo was applied had an item difficulty of 2.78 logits. Only 5 candidates in the person sample were within 1 logit of this item, thus the item error is 1.90 logits and no accurate measure could be provided that is useful for measurement. This item is far too difficult for the sample. **Remove item**

	ITEM 31 This item has a negative correlation to other items as it suffers from disordered distractor responses. The mean person ability of the candidates choosing the correct answer is lower than the mean person ability of persons choosing another option. 26% of persons chose the correct option A. However, their mean ability was 0.63 logits. However, 57% of persons chose option D and had a mean person ability of 0.67 logits. The difference in ability is minimal and the item has appropriate infit, thus the item can be left in the test. However, the prompt diagram is possible blurry and option D make look like 8/10 which may explain why 57% of candidates chose this option. **Review this item for content**
	ITEM OUTFIT **Items 1 and 14** are easy items and have several unexpected mistakes by more able persons. This is termed carelessness. **These items function as expected. No need for review.**

Discrimination – Rasch empirical discrimination statistics

Item discrimination	**Items 14, 17, 33 violate discrimination requirements**
Commentary	**ITEM 14** is a simple subtraction question. The poor discrimination is likely due to unexpected correct responses by higher candidates as discussed in outfit above.

	This item can remain.
	ITEM 17 has the lowest discrimination on the test. This item was flagged as of concern in the G3 Maths analysis. The images are blurred, and the correct answer D does appear to have slightly rounded edges, especially on the left of the shape. **This item needs revising.**
	ITEM 33 has lower discrimination than the most items and is classified as substantial. Again, this item has a blurry image. This item needs revising.

See appendix 2 - Table 3

3. Item Coverage

Item Coverage	· Too many items were too difficult for the sample. · 5 items (including item 28 and 50) were above the most able person. · 3 items were matched at the best candidate, causing redundancy. · Item coverage appears to cover most of the spectrum of ability in the sample. However, this is misleading. · The mean person ability was -.24 logits. · 16 items were 1 logit or more above the mean person. Thus, the average person would have a 27% and less chance of success on these items. · 10 items were more than 2 logits easier than the mean person. Thus, the average person would have an 88% chance of success on these items. However, they still contribute to measurement. · There is a gap in items around mean ability. **Remove items 28, 42, 45, 46, 50 – too difficult.** **Remove items 1, 3, 7 – too easy** **Remove items 2 and 6 – redundant at low level.** **Remove redundant items – where there are 3 or more items at the same difficulty level.** **Add items around mean person ability to fill gaps.**
	· Despite the issues described above, the group of items in the test have managed to generate a person reliability of 0.83 as they do overall cover the spread of ability within the sample.
Commentary	*The decision must be made more clearly on the intended or focus person on the test. As discussed for Grade 3 Written test, if the test is to measure a wide spectrum of ability, then more difficult items must be added.

	However, if the lower end of the ability spectrum is the focus, then this selection of items together would work well, especially if there was a cut score around or below the mean person ability.

4. Unidimensionality and local independence - Rasch model

Item Unidimensionality	Unexplained variance in first contrast less than 2 indicates likely unidimensional measure of literacy.
Local independence of items	Item 25 has a correlation of 0.69 with item 27, thus item local independence has been violated. It is unclear how they may be dependent on each other, but they have a deterministic relationship. This is a threat to measurement.

ITEMS and Gender DIF	No items met substantial thresholds to be a threat to measurement

5. Effect on item removal – rash model

4 options were modelled to assess the impact on item and person reliability, while maintaining unidimensionality. The effect on DIF in four areas were also considered.

	Item reliability	**Person Reliability**	**Person Separation**	**Unidimensionality and Independence**	**DIF - Gender**

Removing: **Local independence [25, 27]** **Too easy [1,3, 4, 7]** **Easy and redundant [2, 6]** **Too difficult [28, 42, 45, 46, 50]** **TOTAL Items: 37**	1.00	0.83	2.18	Yes / yes	No

6. Recommendations

Recommend remove	To ensure item local independence: remove items 25 and 27 Remove items far to easy for sample and consider removing persons below 3 logits from sample as they have 5% probability of success for items at mean difficulty: 1, 3, 4, 7. Remove items too difficult for sample: 28, 42, 45, 46, 50 Look to remove redundant items that are either at the very easy end of the continuum such as items 2 and 6 and look at the more difficult items that have few persons at their level such as item 47.
Investigate	Review items 17, 31, and 33 for content improvement

Stage 2 Grade 6 Written Test - Rasch Model

1. Stage 2 Summary Statistics – Rasch Model – Reliability and Separation

Items	The 41 items have a reliability of 1.00 and create 15 levels within the domain
Persons	The reliability is 0.88 which creates separation of 2 levels. This is likely caused by a wide spread of abilities that have mostly been met by well-targeted items.
Commentary	The test in stage 2 has performed better with this sample than the test used in stage 1 and the related sample. The revised instrument has an increased person reliability above the minimally acceptable criteria of 0.80 and as a result has in this sample been able to separate candidates into 2 levels. The items have perfect reliability and have been separated across 15 levels. This is twice more than the previous stage 1 test which achieved 8 levels of separation.

2. Item Fit – Rasch model

** Fit ranges of 0.8 to 1.2 as acceptable for measurement as used in stage 1 analysis*

** For stage 2 sample size of 1500 candidates, it is recommended to use fit ranges 0.95 to 1.05. However, the more liberal range used in stage 1 was applied again as the assessment is still in early stages. Once a larger bank of items has been created, more conservative ranges can be applied.*

Problematic Items	**34, 37**
Misfitting Items Outfit	**34,35,36,37,40,41**- a result of candidates of low ability guessing
Misfitting Items **Infit**	**34** – this is a result of persons at the item ability level unexpectedly getting the item incorrect
Item polarity	No item to report

Commentary	**ITEM OUTFIT** Items reported with high outfit over 1.2 MNSQ and ±2 ZSDT were investigated by removing low ability candidates with a less than 27% chance of success to reduce the effects of guessing on the items. All reported items improved in quality to either below required thresholds or marginally above. See Table 2a and 2b for comparison. As a result, items 34 and 37 are still suggested for review as they have outfits of 1.21 and 1.22 MNSQ respectively. These items are likely impacted by guessing of persons closer to the item abilities.
	ITEM 37 A result of reducing the impact of guessing saw the difficulty of the item increase from 2.82 to 3.73 logits. The average person measure was 1.29 logits. Before guessing was reduced the mean person had a 15% chance of success on this item. Once a more accurate difficulty measure was established, the mean person would now have an 8% chance of success. Should this item be this difficulty for the sample? especially given it is a simple date question – described in the IPA reading framework as retrieving explicit information from informative texts. Possibly the content is misleading and would cause communication issues as a text in itself. Why are the dates not placed in chronological order in the text prompt? **Review this item for content**

	ITEM 34 **Again, once the impact of guessing had been reduced, both infit and outfit improved for this item. The Infit misfit reduced from a concerning 1.31 MNSQ to an acceptable 1.16 MNSQ. The Outfit remained higher than the 1.2o threshold, though very marginally.** **This item is the most difficult item on the test. In regard to quality measurement this item is fine.**

Discrimination – Rasch empirical discrimination statistics

Item discrimination	**Item 34 violate discrimination requirements**
	Items with low discrimination: ***34, 37***
Commentary	**ITEM 34** **The initial discrimination was very low at 0.28 (the ideal is a slope of 1). This was substantial as indicated by the lower asymptote of 0.11 (>10 is substantial). Again, once guessing had been reduced discrimination improved, but still much lower than other items and was substantial (slope of 0.51 / l.a. of 0.11).** **This item could be removed if opportunities for item efficiency were important.**
	ITEM 37 **This item should be viewed as above, but the lower discrimination of 0.58 is just below the substantial criteria (slope of 0.58 / l.a. 0.08)** **Given the discussion of the content issues of this item, it should be removed or revised.**

3. Item Coverage

Item Coverage	· Item coverage was skewed towards easier items for the sample. · The mean person ability of 1.29 was much higher than the mean item ability. 28 items have a mean difficulty of 0 or lower which means the average candidate would have a 75% chance of success on 66% of the items in the test. · Nearly 10% of the persons were above the most difficult item (item 37) and thus will have higher standard errors. · Item 4 have a difficulty measure of -5 logits. The least able person in the sample had a 90% chance of success on this item. · There are multiple items of the same difficulty especially with 1 s.d. of mean person ability. This may have artificially inflated the person reliability. A number of items suffer from overfit as a result, but this does not degrade measurement as much as underfit. **Remove item 4.** **Remove redundant items.** ***Add more difficult items for more able persons**
	· 90% of the persons were well-targeted by the items.

Commentary	The item coverage can be described as well-targeting overall. *The decision must be made more clearly on the intended or focus person on the test. As discussed for Grade 3 Written test, if the test is to measure a wide spectrum of ability, then more difficult items must be added. However, if the lower end of the ability spectrum is the focus, then this selection of items together would work well, especially if there was a cut score around or below the mean person ability. Item 4 should be removed. It adds nothing to measurement of the sample.

4. Unidimensionality and local independence - Rasch model

Item Unidimensionality	Unexplained variance in first contrast has a 2.20 Eigen value. This exceeds the recommended limit of 2 – meaning there are roughly two items in the set that are multidimensional. On investigation of the Standardized residual loadings for item, items 27 and 36 were identified as above the recommended limit of 0.40. ITEM 27 was the only item to show significant overfit – it behaved too predictably and provided little additional information to measurement. ITEM 36 was an inference question and may require students to perform cognitive tasks outside of the prompt text. Is this content and cognitive domain essential to the construct definition? Once removed, the Eigen Value fell to 1.97. **Recommend remove items 27 and 36.**
Local independence of items	**No items to report**

ITEMS and Gender DIF	No items met substantial thresholds to be a threat to measurement

5. Effect on item removal – rash model

4 options were modelled to assess the impact on item and person reliability, while maintaining unidimensionality. The effect on DIF in four areas were also considered.

	Item reliability	**Person Reliability**	**Person Separation**	**Unidimensionality**	**DIF - Gender**

				and Independence	
Removing: **Dimensionality [27,36]** **Too easy [4]** **TOTAL Items: 38**	1.00	0.88	2.65	Yes / yes	No

6. Recommendations

Recommend remove	To improve dimensionality: remove items 27 and 36 To allow replacement of items for better measurement remove item 4 as far too easy for sample.
Investigate	Review items 34 and 37 for content improvement
Suggestion	Add more difficult items if the target persons are across the ability continuum.