What assessment features affect reading comprehension scores? Experimental evidence from Rwanda and Nigeria

This study

The importance of careful measurement of reading comprehension skills

As the learning crisis in low- and middle-income countries becomes ever more apparent, policymakers and researchers worldwide have shifted their focus from educational inputs to learning outcomes. Throughout this transition, data on learning outcomes have become key in shaping the direction of educational interventions. One key literacy construct that is often measured to inform new policy action is reading comprehension. Reading comprehension is often recognized as the bridge between "learning to read" and "reading to learn." In this sense, reading comprehension is the skill that pupils need to acquire by the end of their literacy training so that they can more consistently and deeply engage with curricular content in language classes and beyond. Given the importance of this skill for pupils' academic careers, reading comprehension has been—either explicitly or implicitly— consistently measured in assessments like TIMSS, PISA, and EGRA, and it is also the key construct underlying highly visible policy indicators like the World Bank's "learning deprivation" metric. Therefore, the quantification of reading comprehension skills is at the core of diagnosing and improving learning gaps in low- and middle-income countries.

This pivotal role of reading comprehension for pupils' learning outcomes and for policymaking planning heightens the importance of its *rigorous*, *accurate*, and *consistent* quantification. Careful consideration should be paid to how different administration features of reading comprehension assessments might affect the reading comprehension scores, and ultimately the conclusions drawn from it.

Through this lens, two key features might play a large role in determining the overall reading comprehension scores for each pupil. First, the timing practices and overall amount of time that pupils are given to answer the questions could be a factor, as stricter timing practices might put more pressure on pupils that might then negatively affect the accuracy of their responses. Secondly, being allowed to reread the passage might also influence scores. Consider reading a passage that is ten lines long and on line 3 it mentions that "The name of the dog is 'Ralph'". Even if pupils understand this sentence as they are reading it, they might not remember a few minutes after reading the passage that the name was specifically 'Ralph'. This would not necessarily be a reflection of poor reading comprehension skills, but instead, it would reflect that, realistically, pupils are not expected to simply not memorize every detail in a passage after a first read.

These two administration features of reading comprehension assessments —timing and rereading practices— are particularly interesting to explore, as some of the most commonly used tools that assess reading comprehension differ on these. For instance, EGRA entails a one-on-one, timed administration of the reading comprehension battery without allowing re-reading. Instead, other assessments like TIMSS entail a large-group assessment including multiple choice questions where pupils are by default allowed to re-read passages and, within the overall time limit of the assessment, can choose how much time to spend on different questions.

Experimentally understanding how different administration features might change the measured outcomes

To better understand how these features might affect the administration and interpretation of reading comprehension assessments and outcomes at NewGlobe, we conducted a randomized controlled trial (RCT) testing the effect of these two features. We sampled 715 pupils in Primary 4-6 from KwaraLEARN and RwandaEQUIP, and administered a reading comprehension assessment consisting of an EGRA passage leveled at a P2-level and a Science passage from a P4 textbook from their context. Importantly, pupils were randomly assigned to taking the reading assessment under three different administration modalities. In particularly, we randomly assigned pupils to the following groups:

Experimental groups	Administration features of reading comprehension assessment	
	Assessment is not timed	Pupils allowed to re-read
Control group	❌	❌
Untimed group ("T1")	✅	❌
Untimed and re-reading ("T2")	✅	✅

After randomly assigning pupils to each experimental group, we gave them the same reading comprehension assessments under the three different administration conditions. Then, we compared the average reading comprehension scores across the three groups. The fact that pupils were randomly assigned to each group gives us confidence that any differences in reading comprehension performance across the three groups can be fully attributed to the type of administration of the assessment that they received.

Results

1. Reading comprehension scores *were not* affected by timing the assessment or not.

Pupils' performance on the reading comprehension assessment was on average the same, whether they were timed or not, as long as they were not allowed to reread the passage. In other words, there was no significant difference in reading comprehension scores between T1 and C. This pattern held across both contexts and most passages, giving us a high degree of confidence that the timing component of the administration of the assessment does not meaningfully affect the outcomes measured.
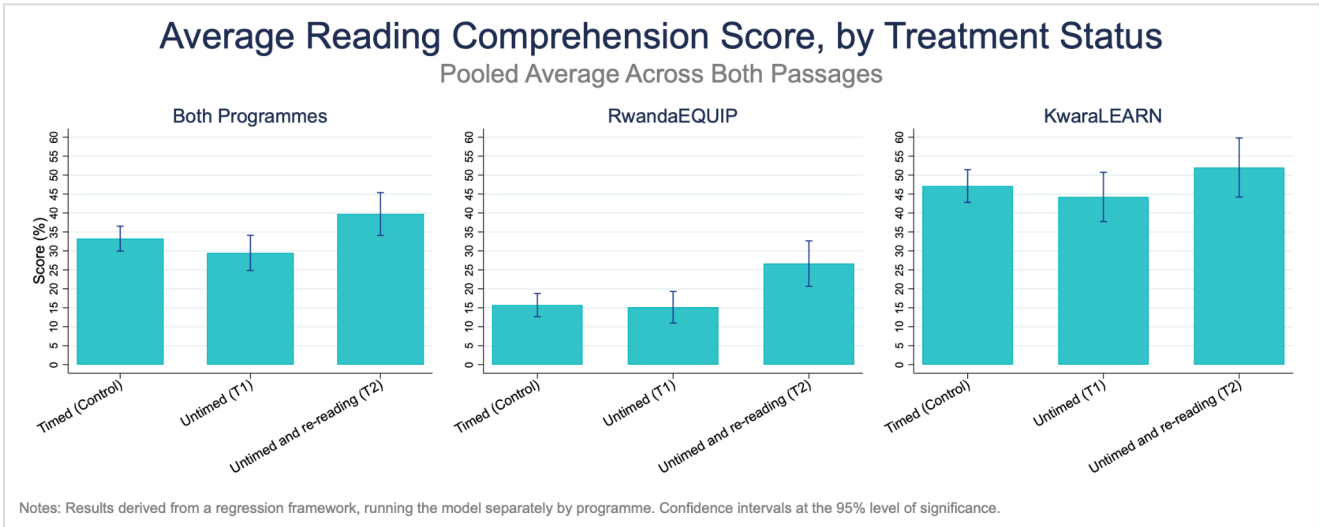
2. Reading comprehension scores *were* higher when pupils were allowed to reread the passage before answering the questions.

Being allowed to reread a passage before answering reading comprehension questions about it does lead to significantly higher outcomes. Across both Rwanda and Kwara, there was a significant increase in reading comprehension scores of 0.18 standard deviations (SD). More concretely, an effect size of 0.18 SD amounts to pupils answering, on average, 0.4 additional questions correctly out of the 10 that they were asked. Notably, this increased comprehension from rereading was not confined to direct questions but also extended to more challenging inference questions. There was also some variance in the effect of rereading by context: pupils allowed to reread passages experienced a boost of 0.29 SD in Rwanda, while pupils in Kwara only experienced a boost of 0.09 SD.

These results can be contextualized in at least two ways. First, 0.4 additional questions amount to a 12% increase in reading comprehension scores, relative to the average of 3.1 questions correctly answered by the
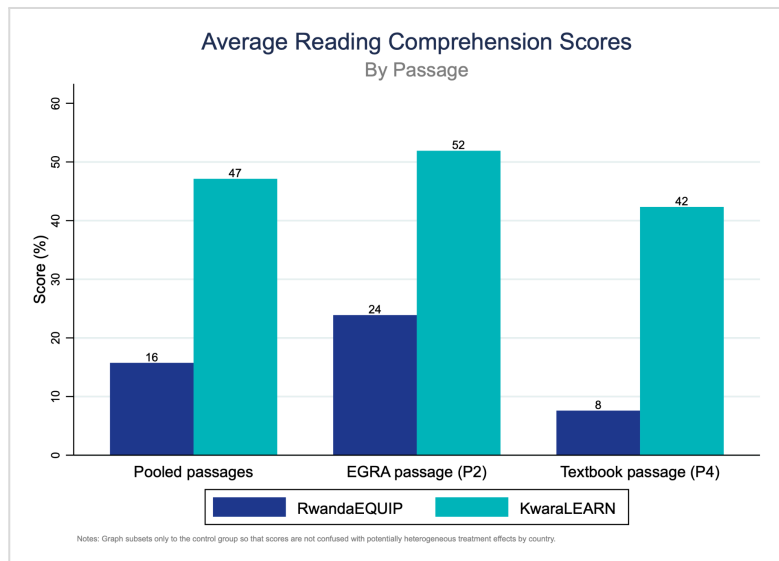
pupils who were not allowed to reread. Even more concretely, one could also define a proficiency threshold akin to 'learning deprivation' or the share of children who cannot read a text and comprehend it. Specifically, if this metric is defined with a threshold of 70% of all reading comprehension questions answered correctly, the effect of rereading a passage would mean that this sample would go from a 'learning deprivation' rate of 85% to 79%. This decrease is similar to transitioning from learning outcomes similar to those in Mali to learning outcomes similar to those in the Dominican Republic.

Average Reading Comprehension Score, by Treatment Status
Pooled Average Across Both Passages

Notes: Results derived from a regression framework, running the model separately by programme. Confidence intervals at the 95% level of significance.

3. Comprehension levels on EGRA passages and textbook passages —which children are actually expected to read in school— are substantially different.

Results from our reading comprehension assessment study revealed considerable divergence between pupils' performance on Early Grade Reading Assessment (EGRA) narrative passages compared to authentic textbook excerpts.

Across both Rwanda and Nigeria, comprehension was substantially lower on textbook passages taken directly from curriculum materials at pupils' grade level. In Rwanda, the disparity in performance between passage types was particularly pronounced. On 59-word EGRA narrative texts, Rwandan pupils demonstrated mean comprehension accuracy of approximately 24% correct on response items. However, when presented with 147-word excerpts drawn from Rwandan textbooks, mean accuracy declined to 8% correct. A directionally similar decline was observed in Nigeria, although the overall comprehension levels were higher.

Average Reading Comprehension Scores
By Passage

Notes: Graph subsets only to the control group so that scores are not confused with potentially heterogeneous treatment effects by country.

In other words, across assessments, Rwandan pupils exhibited even more difficulties comprehending key concepts and responding accurately to questions on textbook passages, relative to their performance on EGRA's narrative texts. This significant discrepancy raises concerns about EGRA's ability to measure reading proficiency at the level required by the syllabi in these LMICs. In other words, the fact that pupils struggled more with officially endorsed curricular materials compared to EGRA's content suggests that if EGRA, or similar passages, are utilized to gauge reading comprehension, it is likely that researchers will significantly overestimate pupils' readiness in terms of reading comprehension.

Additionally, pupils' significant challenges with comprehension of grade-level textbooks — across both Rwanda and Nigeria — indicates a mismatch between curriculum expectations and learning levels, where materials are pitched at a level that significantly surpass learners' current skills. Addressing this systemic gap between academic demands and pupils' literacy readiness deserves ongoing and focused attention. This involves not only enhancing learning outcomes in absolute terms but also ensuring the meticulous alignment of curricular texts with pupils' proficiency levels. This way, classroom instruction is not set at a level that the majority of pupils cannot engage with, and more pupils instead can benefit from their time in school.

Program Comparison of Reading Comprehension
By Passage
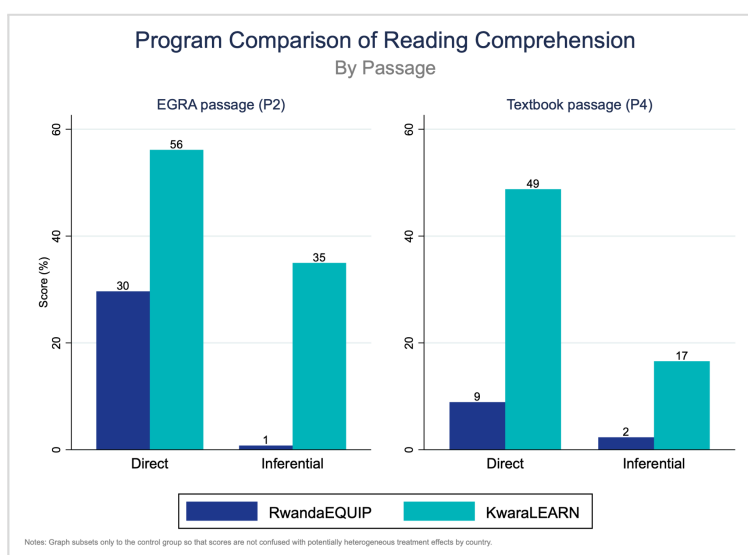
In sum, our results suggest at least two key points. First, there is a need for further research into optimal methods for measuring real-world reading comprehension. In fact, this RCT originated from our internal programmatic need to address this question, given the significant gap in the external literature on this topic. Similarly, carefully tailoring instructional texts to match pupils' demonstrated proficiency levels may help strengthen learning outcomes, as pupils might be able to access more of the content when pitched at a reading level near their current skill set.

4. Reading fluency is a strong predictor of reading comprehension outcomes in both countries, but for any given level of fluency, reading comprehension was consistently lower in Rwanda — likely because of weaker oral language skills in English

Our results revealed — as expected given decades of external evidence — that oral reading fluency served as a strong predictor of comprehension outcomes across both the English-as-a-national language Nigerian context and the English-as-a-foreign-language Rwandan setting. Each increase of 10 correct words per minute in pupils' oral reading fluency was associated with comprehension score gains of approximately 0.2-0.3 standard deviations in both places.

However, noticeable differences emerged when comparing comprehension skills between groups at matched fluency levels. Specifically, Rwandan pupils, comprising primarily English learners, exhibited significantly lower comprehension scores compared to Nigerian pupils at equivalent oral fluency rates. For instance, at a fluency level of 45-55 correct words per minute, the average Nigerian pupil correctly responded to about 53% of

comprehension questions on the EGRA passage. By contrast, the average Rwandan pupil at the same fluency level answered only 20% of questions correctly *on the same passage*.
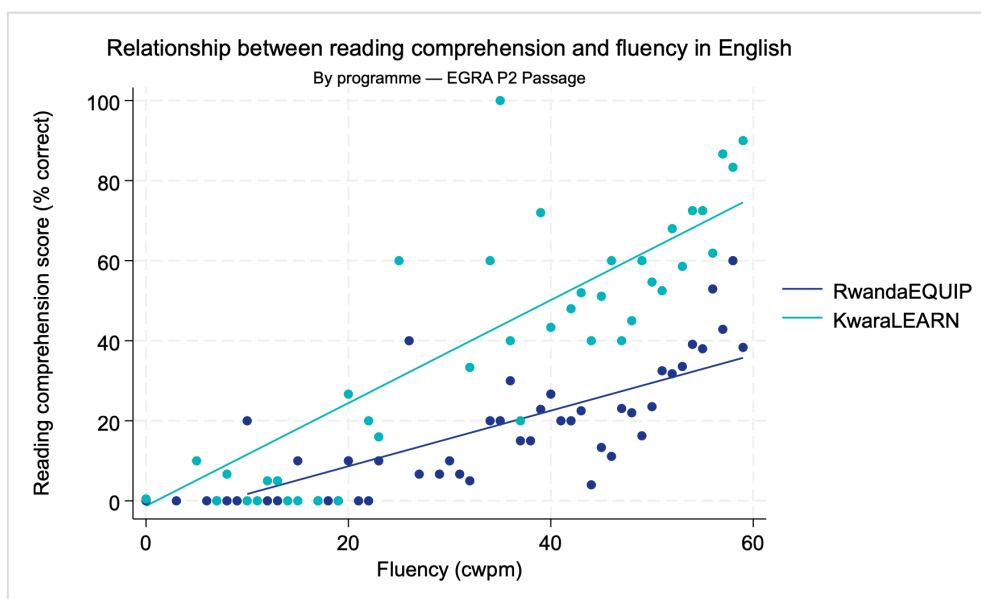
In other words, even when comparing pupils who achieved similar fluency scores across contexts, Rwandan pupils' ability to understand and interpret passages lagged behind their Nigerian peers — who read orally at the same speed and accuracy. This pattern indicates that when setting reading thresholds measuring "reading proficiency", higher benchmarks may be necessary for pupils learning English as a foreign language to overcome vocabulary and background knowledge gaps relative to native speakers. That is, additional fluency likely either provides a direct, compensatory mechanism or acts as a proxy for such, for Rwandan pupils to achieve comprehension on par with their Nigerian counterparts.

Our findings affirm oral reading fluency as a significant predictor of comprehension. Yet, these results also underscore the likelihood of differing fluency expectations and targets across linguistic contexts to enable equivalent comprehension outcomes. Further research continuing to unpack interactions between fluency, language background, and reading comprehension will prove valuable for informing effective literacy instruction for diverse learners.

Key take-aways

Our recent reading assessment study in Rwanda and Nigeria yielded several important insights into reading comprehension measurement in low- and middle-income countries (LMICs).

First, **we found that allowing pupils to re-read passages at their own pace led to substantial gains in comprehension**. On average, pupils' scores on comprehension questions increased by approximately 0.2 standard deviations when permitted to re-read in an untimed setting. This boost was consistent across different types of test questions, including straightforward factual questions as well as more challenging inferential items requiring deeper understanding. Yet, **this increase was not driven by the lack of timing on their assessment**, as children who were assessed in an untimed manner but without the possibility of re-reading the passage performed at the same level as their peers who were assessed using a timed protocol. Critically, this suggests that the standard Early Grade Reading Assessment (EGRA) procedure of reading with no re-reads likely

underestimates pupils' true reading comprehension abilities. By preventing re-reading, time constraints impose an extraneous cognitive load that may obscure pupils' comprehension, especially for harder texts.

Second, **we found substantial differences in comprehension when pupils read authentic textbook passages versus the short EGRA passages**. On textbook passages taken directly from grade-level science materials, pupils' comprehension scores were approximately 12 percentage points, or almost 0.4 standard deviations, lower on average compared to EGRA passages. This gap was especially pronounced in Rwanda, where the gap in favor of the EGRA narrative passage was 17 percentage points, relative to only 7 percentage points in Nigeria.

This finding highlights potential limitations of EGRA's passages for gauging the reading skills required by LMIC state and national curricula. Pupils struggled when attempting to read the texts they are supposed to engage with every day, suggesting EGRA-like assessments may overestimate upper-primary readiness. Our findings indicate a need for assessments incorporating authentic academic passages to better measure real-world reading proficiency. Additionally, the substantial gap between EGRA and textbook performance shines a sobering light on the sheer difficulty textbooks pose for many pupils in low- and middle-income countries. The fact that grade-level materials strain learners' reading abilities raises serious questions about the appropriateness of curricular content. Although this systemic mismatch between academic curricular demands and pupils' readiness levels has generated plenty of attention and research, it has rarely been through the lens of textbooks themselves — the materials the pupils are using every day. Hence, this finding suggests a pressing need to re-evaluate textbooks and curricula in LMICs to ensure better alignment with pupil learning levels. Without interventions to improve the accessibility of instructional materials, pupils will face content that is too difficult, and make minimal progress.

Third, our data reveal reading fluency is a robust predictor of comprehension outcomes in both the English-language Nigeria context and English-as-a-foreign-language Rwanda setting. Each increase of 10 correct words per minute in pupils' oral reading fluency was associated with comprehension score gains of approximately 0.2-0.3 standard deviations in both places. However, conditional on a fixed fluency level, Rwandan pupils' comprehension lagged far behind their Nigerian peers. For instance, at a fluency level of 45-55 correct words per minute, the average Nigerian pupil answered around 53% of comprehension questions correctly on EGRA, while the average Rwandan pupil answered only 20% correctly. This implies that a **higher fluency threshold likely enables English learners to overcome vocabulary and background knowledge deficits to achieve comprehension on par with either native speakers or pupils in places where English is more contextually prevalent**. Our findings suggest policymakers and educators may need to set more ambitious fluency goals for reading programs serving English learners to account for their linguistic disadvantages.

In summary, our study sheds light on the complex interplay between reading fluency, comprehension, language status, background knowledge, and assessment conditions in LMIC contexts. The results underscore the need to assess reading in a multi-faceted manner, using authentic texts and allowing re-reading to better capture true comprehension capacities, especially for English learners. We hope these insights can inform the design of reading assessments and classroom materials that provide policymakers with more nuanced, meaningful data. This data, in turn, can be used to improve literacy instruction and outcomes across LMICs and ensure broader accessibility of the classroom materials used to accompany classroom instruction.