Working with Kano State's Local Government Officials to Assess Pupils' Literacy Skills at Scale: A Reliable Approach?

Umar Kabir
Bala Danyaro Aminu

Abstract: This paper presents an evaluation of the reliability of a large-scale pupil assessment exercise, which tested pupils' literacy skills using the UK Government's "Phonics Screening" tool. This assessment exercise was conducted in Kano State, where the assessors were local government officials. The paper presents an evaluation of the assessment training and the assessment process, based on observations of both by the authors of this paper. The research question answered by the paper is "To what extent is working with local government officials to assess pupils' literacy skills a reliable approach for understanding pupils' literacy levels at scale?"

Introduction

Understanding whether early grade pupils are acquiring essential literacy skills is vitally important, as this affects their learning across all subjects. How we can reliably assess pupils' literacy skills at scale is therefore an important question. This is particularly the case for Kano State, which has extremely high numbers of schools and pupils.

Universal Learning Solutions Initiative invited us to undertake an external evaluation of their "Phonics Screening Exercise", as part of which they trained 40 Government School Support Officers (SSOs) in how to assess pupils' letter sound and work reading abilities using a smartphone application: Jolly Monitor. These SSOs, located across different Local Government Areas then each went to 8 schools to assess Primary 1 to 3 pupils.

Methodology

Our research question for this evaluation was *"To what extent is working with local government officials to assess pupils' literacy skills a reliable approach for understanding pupils' literacy levels at scale?"*

We answered this question by observing and evaluating the quality of the assessor training, and also by observing and evaluating 16 different assessors as they conducted the assessments in schools.

Firstly, we attended the training of assessors to observe it in action. At the end of the training, we interviewed and tested the assessors, in order to evaluate the extent to which we felt that the training had been successful in giving them the knowledge and skills to effectively carry out the phonics screening exercise.

Following the training, we randomly selected assessors to observe during the assessment process. We each chose 8 officials and 8 random schools that they were visiting, and then travelled to the schools within which they were conducting assessments on the allocated days to observe the process. During the day, we completed a general evaluation questionnaire about the process, and also a questionnaire relating to the assessment of each individual child. For each assessor that we observed, we answered a number of questions pertaining to the reliability of the assessment process.

Results: Assessment Training

This section presents the results from our evaluation of the assessment training.

1.1 Out of 10, how would you rate the 2019 Phonics Screening Check training?

We interviewed all 40 SSOs and asked them how they rated the training out of 10. The average rating was 8.25/10. Figure x below splits the ratings in a pie chart. It shows that most assessors rated the training 8 or 9 out of 10. This suggests that they were generally pleased with the guidance that they received in how to successfully and reliably conduct the assessments.

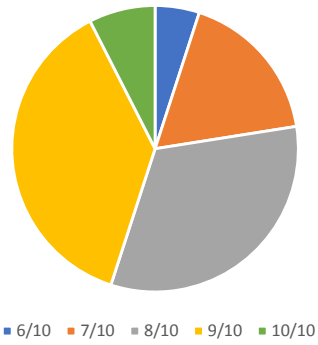

Figure 1 - Pie Chart Showing the Training Ratings by the Assessors

1.2 Out of 10, how prepared would you say that you now are for administering the Phonics Screening Check in schools?

We then asked the assessors how prepared they felt out of 10 after the training for administering the Phonics Screening Check in schools. The average score was 8.35 out of 10. Figure 2 below shows that the split of ratings. It again shows that most assessors said that they felt 8 or 9 out of 10 in terms of their level of preparedness for conducting the assessments in schools.

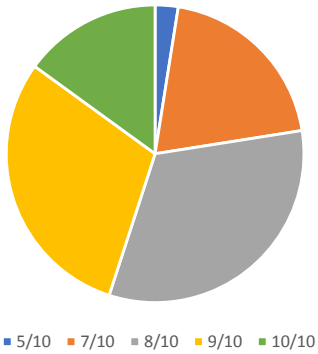

Figure 2 - Pie Chart Showing How Prepared Assessors Felt to Conduct the Assessments

1.3 How confident are you that you know all of the correct answers on the Letter Sounds test?

We then asked the assessors how confident they felt that they knew all of the correct answers on the Letter Sounds Test. Figure 3 shows that 95% of the assessors (38) felt "very confident" that they knew all of the correct answers on this test, and that the remaining 5% (2) felt "partially confident". We believe that this suggests that overall the assessors were confident with the answers on the Letter Sounds Test, as the two assessors that felt partially confident had planned to go and practice the correct answers with the guidance video that had been created.

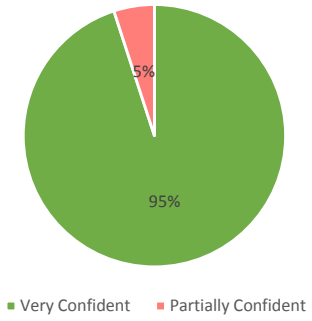

Figure 3 - Pie Chart Showing Confidence Level Split on Answers to Letter Sounds Test

1.4 How confident are you that you know all of the correct answers on the Word Reading test?

We then asked the assessors how confident they felt that they knew all of the correct answers on the Word Reading Test. Figure 3 shows that 87% of the assessors (35) felt "very confident" that they knew all of the correct answers on this test, and that the remaining 13% (5) felt "partially confident". We believe that this suggests that overall the assessors were confident with the answers on the Letter Sounds Test, as the five assessors that felt partially confident had planned to go and practice the correct answers with the guidance video that had been created.

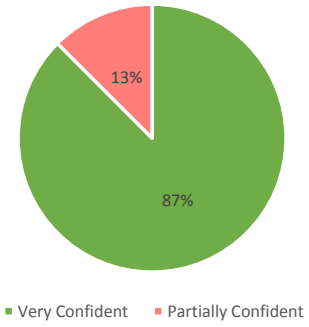

Figure 4 - Pie Chart Showing Confidence Level Split on Answers to Word Reading Test

1.5 How confident are you that you know how to score correctly on both tests?

We then asked the assessors how confident they felt about scoring answers correctly on both tests. Figure 5 shows that 90% of the assessors (36) felt "very confident" that they knew how to score correctly on both tests and that the remaining 10% (4) felt "partially confident". We believe that this suggests that overall the assessors were confident with scoring on both tests, as the four assessors that felt partially confident were again those that had felt partially confident on the correct answers on the tests and had planned to practice using the guidance videos.

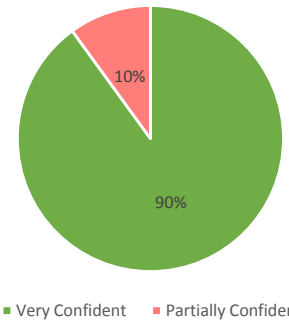

Figure 5 - Pie Chart Showing Confidence Level Split on Scoring on Both Tests

1.6 How confident are you that you know how how to sample classes and pupils in an unbiased way?

Results: Assessment Process

This section presents the results from our evaluation of the assessment process, along with a narrative of what we believe the results tell us about the reliability of the phonics screening exercise.

2.1 Was the assessor's method of gathering school, head teacher, class teacher and pupil data reliable?

First, we evaluated whether the assessor's method of gathering basic data was reliable. This included basic information about the school, head teacher, class teacher and pupils, such as school population, attendance rates, etc.

As Figure 6 displays, we deemed all of the 16 assessors' methods to be reliable, meaning that we felt that they took the time and undertook appropriate steps to capture accurate information.

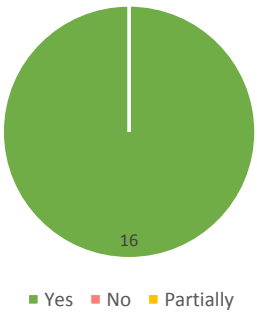

Figure 6 - Pie Chart Displaying Whether the Assessor's Method of Gathering Basic Data Was Reliable

Commented [LG1]: Finish this and more questions from the excel in folder

2.2 Was the assessor's method of choosing the classes within which to carry out the assessments unbiased?

Second, we evaluated whether the assessor's method of choosing the classes to take part in the assessments was unbiased. The assessors were instructed to randomly select one Primary 1, one Primary 2 and one Primary 3 class. The only criteria for the Primary 1 and 2 classes in the Jolly Phonics schools was that the teacher should have been trained in Jolly Phonics. They were instructed not to base their selection on any other criteria, such as choosing the best performing class, which would make the selection biased. Some of the assessors combined all classes and chose pupils from across the classes, which we also deemed to be unbiased.

Figure 7 below displays how we found 15 of the assessors to have selected classes in an unbiased way, and 1 of the assessors to

have only partially undertaken this process correctly. The reason for the "partial" rating was that they selected pupils from across several classes, but we still deemed this to be unbiased, and so did not affect the reliability of the data.

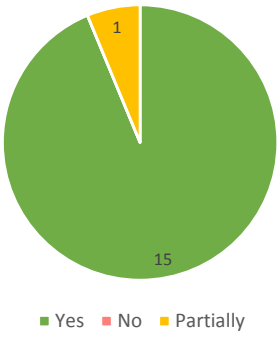

Figure 7 - Pie Chart Displaying Whether the Assessor's Method of Selecting Classes was Unbiased

2.3 Was the assessor's method of choosing a sample of pupils from each class to assess unbiased?

Third, we evaluated whether we felt that the pupil sampling procedures were unbiased. As noted above, the assessors were trained in how to randomly sample pupils.

As Figure 8 shows, we observed 13 of the 16 assessors sample pupils in this random way, but 3 assessors we felt only partially implemented this unbiased sampling procedure. However, for those that we noted only partially implemented this procedure, they still selected pupils randomly and not purposefully from the group, they just did not follow the prescribed procedure. In this respect, we felt that the sampling of all pupils was actually unbiased, and so the change in procedure did not affect the reliability of the data.

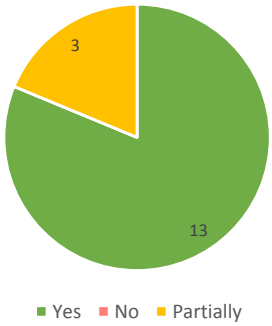

Figure 8 - Pie Chart Displaying Whether the Assessor's Pupil Sampling was Unbiased

2.4 Did the assessor carry out the assessments in a quiet, private and comfortable environment?

Fourth, we evaluated whether the assessor conducted the assessments in an appropriate environment. An appropriate environment was deemed to be where it was quiet, private and comfortable for both the assessor and pupils.

As Figure 9 displays, we found 10 out of the 16 assessors to have chosen an appropriate environment within which to conduct the assessments, 5 partially did and we felt that 1 chose an environment that was not appropriate.

To explain further, some of the partially appropriate environments were not private, in that they were conducted in an open classroom where people could enter at will, and others were affected by noise from neighbouring classrooms or outside. Overall, however, we believe that the

slightly noisy and non-private environments would have had a negligible impact on pupils' scores, because the assessments could still be conducted and the pupil did not seem affected in their performance by these factors. In this respect, we do not believe that the environments affected the reliability of the data collected by the officials.

The one environment that we deemed to be inappropriate was where the assessments were conducted outside. However, we do not think that this would have affected the pupils' scores enough to have their data removed from any evaluation.

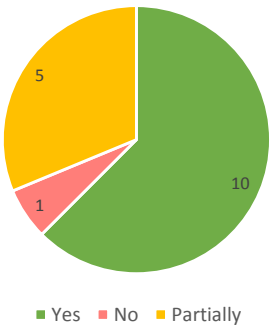

Figure 9 - Pie Chart Displaying Whether the Assessor Carried Out the Assessments in an Appropriate Environment

2.5 Are you confident the pupil understood what was being asked of them after the assessor explained the process to them?

Fifth, we judged whether we felt that each pupil fully understood what was being asked of them after the assessor explained the assessment process. Our decision was based on the reaction of the pupil, rather than any questioning of the pupil, meaning that we were not 100% certain that the pupil did or did not understand.

Figure 10 shows that we were confident that 97 out of 144 pupils (67%) fully understood what was being asked of them, that we felt that 38 (26%) partially understood, and with 9 pupils (6%) we were not confident that they understood what was being asked of them.

With the partially confident pupils, they were still able to complete the tests, but perhaps would have benefitted from better instruction.

Overall, we therefore felt that the pupils mostly understood what was being asked of them, so we believe that the data is good enough to provide a reliable evaluation of pupil performance.

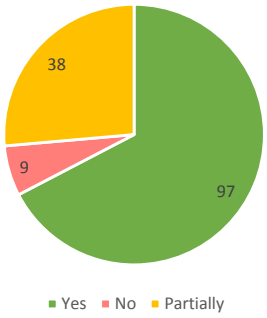

Figure 10 - Pie Chart Displaying Whether Pupils Were Thought to Have Fully Understood Assessment Instructions

2.6 Letter Sounds Test – did you observe any possible errors in the scoring of the letter sounds test?

We then observed the sounds test being conducted with each pupil in order to see if we observed any errors in the scoring of pupils by assessors. As Figure 6 shows, out of 144 pupils that we observed being assessed, we felt that the assessor scored correctly for 137 pupils, which is 95%.

For the 7 pupils (5%), we observed what we thought was a minor error in the scoring for one sound. For six pupils, they slightly pronounced a single vowel sound incorrectly and the assessor scored it correct. For example, /oo/ was pronounced more like /o/ by one pupil. For one pupil, we saw that they pronounced the /ar/ sound correctly, but the assessor scored it incorrect.

Overall, however, we felt that these were very minor errors in the scoring of pupils' letter sound knowledge, and so did not affect the overall scores significantly enough to mean that the results were unreliable.

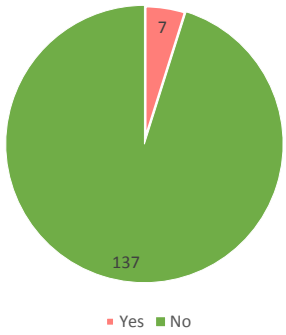

Figure 11 - Pie Chart Displaying Whether the Sounds Test Was Scored Properly by Assessors

2.7 Word Reading Test – did you observe any possible errors in the scoring of the word reading test?

We then observed the word reading test being conducted with each pupil in order to see if we observed any errors in the scoring of pupils by assessors. As Figure 7 shows, out of 144 pupils, we again felt that the assessor scored correctly for 137 pupils, which is 95%.

For six pupils, we felt that they mispronounced some words slightly, but they were scored correct. For example, 'slirt' was read 'slet' by one pupil, and 'modern' was read 'mode' by another. This could have been the assessor mishearing the pupil, but also could have been that they were slightly biased in their scoring, wanting pupils to do well. For one pupil, we observed the assessor helping the pupil with the sounds in one word, which helped them to pronounce the entire word.

Overall, however, we again felt that these were very minor errors in the scoring of pupils' word reading knowledge, and so did not affect the overall scores significantly enough to mean that the results were unreliable.

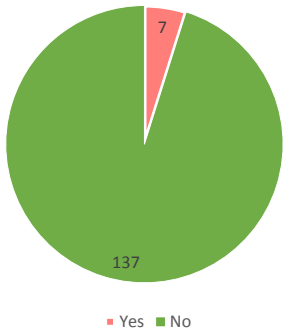

Figure 12 - Pie Chart Displaying Whether the Word Reading Test Was Scored Properly by Assessors

2.8 Are you confident the assessor did not influence the outcome of the assessment in any way?

Finally, we commented on whether we felt that, overall, the assessor could have influenced the outcome of the assessments in any way. As Figure 8 shows, out of 144 pupils, we felt that the assessor did not influence the outcome of the assessments for 125 pupils, which is 87%. For 19 pupils (13%), however, we were slightly concerned about the practice of the assessor.

These 19 pupils were broadly split into two categories: the first was where we felt that the assessor could have done more to explain to the pupil what was required of them and/or give them more time to answer and; the second was where the assessor slightly helped the pupil will displaying actions for sounds and/or with descriptions. Nevertheless, once again, we felt that these influences were minor, relating to 1 or 2 sounds or words, and overall we believe that the results are reliable enough to show an accurate representation of the abilities of pupils.

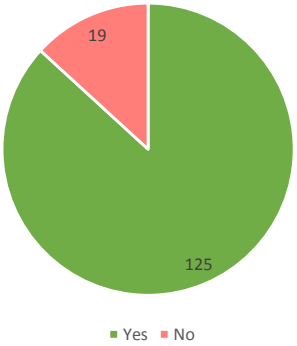

Figure 13 - Pie Chart Displaying Whether We Were Confident that the Assessors Did Not Influence the Outcome of the Assessments

Our overall impressions of the assessment process

Overall, despite some minor issues discussed above, we both felt confident that the assessment process was carried out to a high enough standard to ensure that the data collected provides a reliable representation of the results of pupils in these schools. There did not seem to be a biased towards either Jolly Phonics schools or non-Jolly Phonics schools in our samples, with the same minor errors being observed across both groups.