Measuring literacy and numeracy over the phone

Case Study: Phone-based assessments in Kenya

Daniel Rodriguez-Segura & Beth Schueler, University of Virginia
July 2021

Motivation

As a result of COVID-induced school closures worldwide, researchers and policymakers have placed a greater emphasis on alternative ways to monitor student learning and engagement beyond in-person assessments, but this challenge is not limited to the current pandemic. Tracking educational metrics has historically been logistically difficult when assessing children who are out-of-school, living in remote areas, or experiencing humanitarian and natural disasters. Given the high degree of market penetration of cell phones in even the world's poorest areas, phone-based assessments (PBAs) have potential to reach these groups and track their educational trajectories.

However, to the best of our knowledge, there has not yet been a rigorous examination of the validity of these assessments, where scores obtained by phone are compared to the same students' in-person achievement measures. Furthermore, beyond the validation of these assessments, other important features of PBAs, such as the role different assessors play in influencing results, have yet to be studied. We partnered with NewGlobe, and through funding provided by IPA's Research Methods Initiative, we collected and analyzed PBA and administrative data from primary school students in Kenya to inform future design, use, and implementation of this type of assessment. Overall, we find:

- PBA did a reasonable job measuring aggregate numeracy skills, literacy skills to a lesser extent, and did a poor job measuring reading fluency.
- PBA did a better job measuring aggregate than individual achievement in all subjects.
- There were large and systematic differences in the extent to which assessors influenced the scores recorded through PBAs.
- Assessing a representative sample by phone was challenging and required intentionality.

Project

NewGlobe completed 3,690 phone-based assessments of 3rd, 5th, and 6th graders enrolled in 105 low-cost private schools across Kenya. NewGlobe routinely collects standardized in-person test scores from all schools and grades, allowing us to capitalize on two rounds of comparable, baseline scores in math ("numeracy"), English ("literacy"), and oral reading fluency as a literacy sub-skill of particular interest to our partner ("fluency," measured in units of number of words correctly read aloud per minute in a grade-level passage), against which we could compare PBAs. Conveniently, NewGlobe maintains a database of student phone numbers, aiding the PBA data collection process. Similarly, all NewGlobe students within grades have access to the same homework book at home, allowing us to use the same reading passages across all students for literacy and fluency PBAs.

Data collection was completed in two stages. During the first stage (December, 2020), we assessed numeracy skills, using both core numeracy questions and curriculum-aligned questions, among 3rd, 5th, and 6th graders who were engaged in remote learning due to Covid-19 school shutdowns. NewGlobe completed 2,644 total assessments. During the second stage (April, 2021), we focused on measuring vocabulary, oral comprehension, and spelling and fluency (aggregated into a single "literacy" score), among 3rd graders. Importantly, we randomly assigned students to assessors at the student level to explore whether assessors record similar achievement levels with different scores.

The two stages were different on two additional dimensions. First, assessors in the first stage were teachers from our partner's schools. Since classes were back in session during the second stage, NewGlobe hired and trained external assessors, all of which had at least a post-secondary diploma, and were fluent in English and Kiswahili. The second difference relates to sampling: during the first round, we selected a sub-sample of students from all classes to be assessed, using simple random sampling. The order in which assessors were expected to call students was also assigned using simple randomized methods. During the second round, the order was also randomized but we gave priority to students from less advantaged backgrounds to attain a more representative sample of assessed students, compared to the first stage.

Results

First, PBA results have a positive though modest correlation with in-person scores, as shown in the first column of the table below, suggesting that there is some - but not perfect - overlap in what these two assessments are measuring. To benchmark this correlation, we also show the correlation of repeated in-person scores in the second column. The third column shows that, for numeracy and literacy, the correlation between PBA results and baseline in-person scores for the same student was about one-fourth the size of the correlation between repeated in-person scores. This was not the case for reading fluency, where the correlation between baseline in-person scores and PBA scores was almost non-existent. The difference in the magnitude of the correlation between in-person scores and the PBA mechanically comes from a flattening of the distribution at both ends: PBA underestimated achievement for high-performing students and overestimated performance for lower-performing ones. Besides comparing assessments from two different modes, the PBA were also shorter than the in-person assessments and measured fewer constructs than the repeated in-person assessments, which likely also contributes to the lower correlations.

Descriptively, we found assessment features that could improve the ability to discriminate individual performance via future PBAs. For instance, some numeracy items exhibited a wider range in the extent to which they help distinguish between different achievement levels ("item discrimination") than others. Questions that were too hard or too easy did not add significant information to overall scores, while items whose difficulty was more carefully tailored to the grade level (e.g., two-digit multiplication in grade 3 math) offered stronger discriminating power. We also found that even though the reliability of PBA measurement of fluency on its own was weak, its inclusion into the scoring of the overall literacy exam strengthened the correlation between the PBA score and the baseline in-person score.

Table: Correlations between assessments, by subject

Subject	Correlation between PBA and in-person assessments	Correlation between repeated in-person assessments	Share of correlations for PBA, compared to repeated in-person assessments
Numeracy	0.15***	0.59***	25%
Literacy	0.17***	0.79***	22%
Fluency	0.01	0.75***	1%

Notes: each correlation represents the coefficient resulting from regressing the most recent in-person assessment on either the previous in-person score, or on the phone-based assessment score. The scores for all assessments have been normalized. Standard errors were clustered at the school-level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Importantly, in numeracy, the lower correlation between the PBA and in-person assessments (relative to repeated in-person tests) did not translate into an assessment that misclassifies students' *relative* performance. In fact, the numeracy PBA performed very similarly to repeated in-person assessments in terms of assigning students a relative position in the overall achievement distribution. This was also true, though to a lesser extent, for the literacy PBA but much less so for fluency. To illustrate these points, we calculate

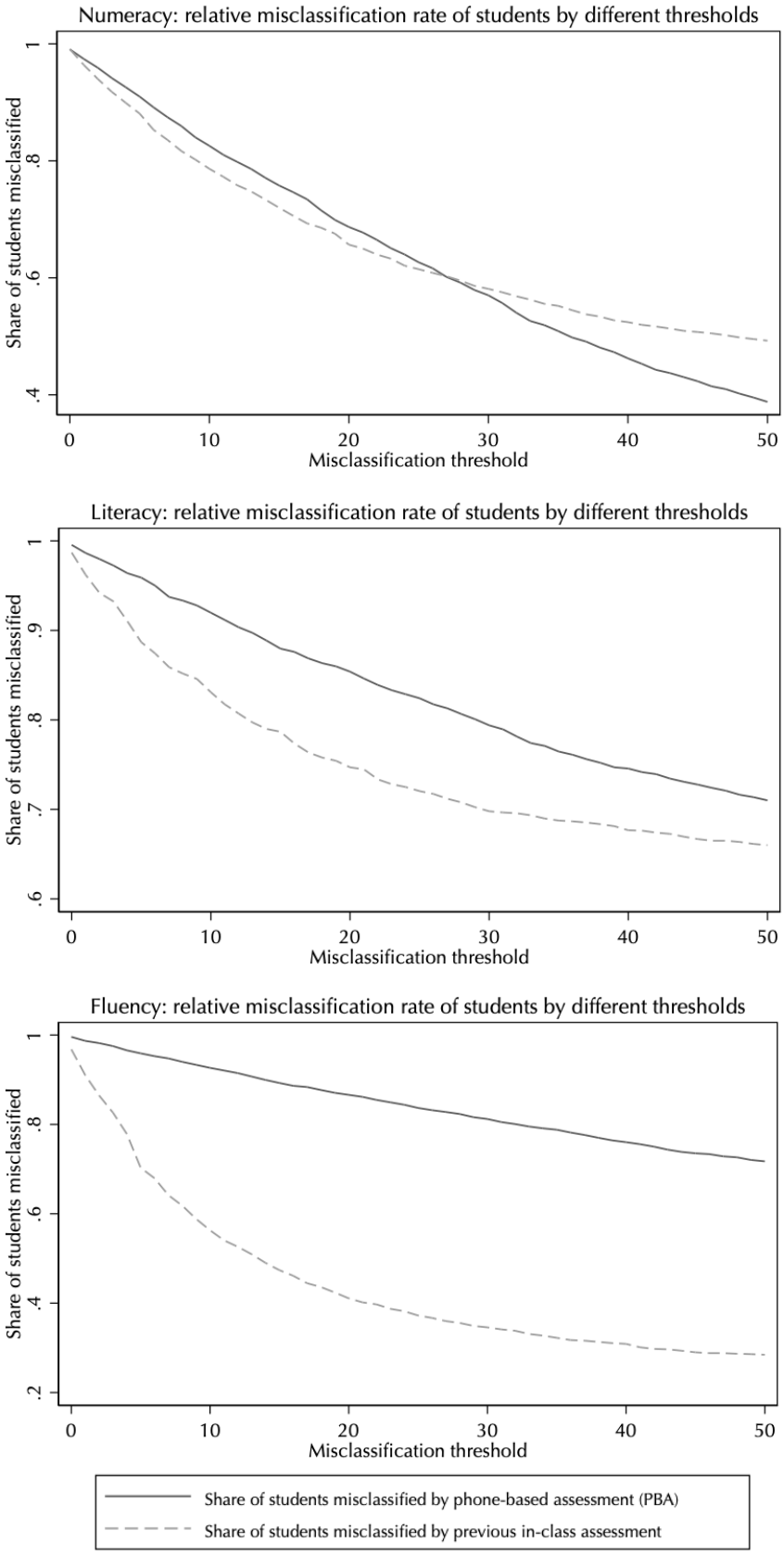
each student's percentile in the relative distribution of the PBA and in the relative distribution of in-person assessment scores, and compare these two percentiles to see how far off these two relative positions are. Comparing two repeated in-person assessments, the median student is between zero to three percentiles off, depending on the subject. When comparing the PBA to the most recent pre-PBA in-person assessment, the median student is similarly between one to three percentiles off. In other words, the numeracy and literacy PBA ranked the median student very similarly to the in-person assessment, and in a similar fashion to repeated in-person assessments.

Interestingly, we find PBAs "misclassify" students' relative position differently by subject, with the lowest rates of misclassification for numeracy, then literacy, and fluency with the highest rates. The figures below show the relative "misclassification" rate (y-axis) of different assessments by different thresholds defining "misclassification" (x-axis). Solid lines show the misclassification rate of PBAs and dotted lines show the misclassification rate of repeated in-person assessments to help benchmark PBAs. The lower a line is, the more consistent the relative classification of students in this assessment is. Similarly, the closer the two lines are to each other within a given plot, the more similarly PBAs perform compared to repeated in-person assessments. For instance, a threshold of 30 on these figures means that a student is misclassified by an assessment if they are 30 or more percentiles off from the latest in-person assessment. Using this threshold, the numeracy PBA assessment appropriately classifies about 42% of students, very similar to the rate of 43% in repeated in-person assessments. For literacy, 21% of students are properly classified by the PBA, and 30% by the repeated in-class measure. For the fluency assessment, the worst performer of the three, the PBA only properly classifies 19% of students, while the repeated assessment properly classifies 65% of students. Though we see differences across subjects, we cannot entirely rule out the possibility that these differences were due to changes in the assessment implementation between the two stages (e.g., teachers vs. hired assessors) rather than exclusively the subject being assessed.

Our next key finding is that the simple random selection of students for the first round of data collection yielded a more "advantaged" and therefore non-representative subsample of students for the numeracy PBA. Specifically, assessed students came from communities with lower poverty rates, larger urban populations, and higher baseline scores. However, explicitly oversampling from those groups less likely to be reached in round one yielded a significantly more representative sample for the literacy PBA.

Finally, different assessors tended to record similar levels of performance differently. Because we randomly assigned assessors to students, on average, the average scores recorded by different assessors should be relatively similar, particularly after controlling for school, grade, and baseline scores of students assigned to each assessor. While we do not find specific assessor characteristics or features of student-assessor match that strongly and consistently predict differential scores, we do find strong evidence for "enumerator effects", or assessors recording different scores in the PBA for students who would have been expected to score very similarly. We also see some suggestive evidence that assessors get more lenient over time, particularly for students with lower levels of baseline achievement.

Figure: relative misclassification rate of students by different thresholds, and subjects



Recommendations

Use PBA for aggregate measures of numeracy and literacy learning, not tracking individual trajectories. PBA measured relative performance in numeracy, and to a lesser extent literacy, fairly well, on-par with repeated in-class assessments. PBAs are therefore well-suited to aggregate monitoring of groups (e.g., areas in remote areas, or classrooms aggregated at the treatment level within an RCT), although not for fluency. Having said this, the lower correlations between in-person and PBA scores compared to repeated in-person scores suggests PBA scores were imperfect proxies of each individual student's performance compared to in-person tests. PBAs, at least of the length and type we administered, therefore do not seem particularly conducive individualized tracking of learning outcomes particularly for high-stakes decisions without further testing and development.

Study reading fluency PBA further before using them to track aggregate or individual trajectories. Results from the fluency assessment we administered were not correlated with in-person measures of fluency and had higher rates of misclassification than in-person assessments. These differences between PBA and in-person assessments were larger for fluency than the other two subjects we assessed, suggesting fluency PBA tools are not yet ready for "prime time." However, we cannot definitively rule out the possibility that fluency could be measured by PBA. The differences by subject that we see could be due to other factors that varied from the first to the second stage and researchers therefore have more work to do to assess the validity of such measures prior to their adoption.

Test multiple different sub-skills and properly target item difficulty to strengthen PBAs. We recommend the inclusion of multiple different numeracy or literacy "sub-skills" within each assessment. Measuring single skills, like reading fluency, might be too noisy or too reliant on the validity of measures for a single construct. However, the aggregation of several sub-skills into a single assessment seems to better reflect previous achievement levels. Similarly, given the high opportunity cost in terms of airtime and assessor time attached to the less informative items in PBA, we recommend that future PBA designs give serious consideration to the extent to which questions are properly suited for the focal grades.

Randomly assign students to assessors, as well as the order they are assessed. Individual-level randomization of assessors to students allows systematic assessor biases to be spread throughout the whole sample in a manner not correlated with student characteristics, avoiding unwanted concentrations of assessors at certain units of aggregation (e.g., schools), which could potentially yield spurious results. We also recommend that the order assessors call each student be randomized to ensure systematic differences in how assessors record student performance over time does not affect average results. Both of these recommendations allow researchers to include controls (i.e., assessor- and day of assessment- fixed-effects) into their regression models to account for systematic differences in enumerating patterns across assessors and over time. This is logistically much easier with PBAs than field surveys or in-person assessments.

Prioritize assessing sub-groups less likely to be reached by phone to achieve a representative sample. Perhaps expectedly, our first stage of data collection systematically reached a group of students that was more "advantaged" on average than the overall student population. In the second stage of data collection, we systematically gave calling priority to students from schools that were most underrepresented during the first round of data collection, and this indeed yielded a more representative subsample of reached students. Therefore, researchers could benefit from prioritizing subsamples that may be less likely to answer and complete the assessment to end up with a more representative overall sample.