ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 22-517

Can learning be measured by phone? Evidence from Kenya

Daniel Rodriguez-Segura
University of Virginia

Beth E. Schueler
University of Virginia

School closures induced by COVID-19 placed heightened emphasis on alternative ways to measure student learning besides in-person exams. We leverage the administration of phone-based assessments (PBAs) measuring numeracy and literacy for primary school children in Kenya, along with in-person standardized tests administered to the same students prior to school shutdowns, to assess the validity of PBAs. Compared to repeated in-person assessments, PBAs did not severely misclassify students' relative performance, but PBA scores did tend to be further from baseline in-person scores than repeated in-person assessments from each other. As such, PBAs performed well at measuring aggregate but not individual learning levels. Administrators can therefore use these tools for aggregate measurement, such as in the context of impact evaluation, but be wary of PBAs for individual-level tracking or high-stakes decisions. Results also reveal the importance of making deliberate efforts to reach a representative sample and selecting items that provide discriminating power.

VERSION: January 2022

Suggested citation: Rodriguez-Segura, Daniel, and Beth E. Schueler. (2022). Can learning be measured by phone? Evidence from Kenya. (EdWorkingPaper: 22-517). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/gc6v-qv41

Can learning be measured by phone? Evidence from Kenya

Daniel Rodriguez-Segura
dan.rodriguez@virginia.edu
University of Virginia, School of Education and Human Development

Beth E. Schueler
beth_schueler@virginia.edu
University of Virginia, School of Education and Human Development

Draft as of January, 2022

Abstract: School closures induced by COVID-19 placed heightened emphasis on alternative ways to measure student learning besides in-person exams. We leverage the administration of phone-based assessments (PBAs) measuring numeracy and literacy for primary school children in Kenya, along with in-person standardized tests administered to the same students prior to school shutdowns, to assess the validity of PBAs. Compared to repeated in-person assessments, PBAs did not severely misclassify students' relative performance, but PBA scores did tend to be further from baseline in-person scores than repeated in-person assessments from each other. As such, PBAs performed well at measuring aggregate but not individual learning levels. Administrators can therefore use these tools for aggregate measurement, such as in the context of impact evaluation, but be wary of PBAs for individual-level tracking or high-stakes decisions. Results also reveal the importance of making deliberate efforts to reach a representative sample and selecting items that provide discriminating power.

Keywords: phone-based assessments, remote learning, school closures, education in developing countries

Author's note: Daniel Rodriguez-Segura is the corresponding author. We are grateful for feedback received from Steven Glazerman, Carolyn Pelnik, Diego Luna-Bazaldua, Betheny Gross, Robin Lake, Jim Soland, Daphna Bassok, Vivian Wong and Isaac Mbiti, and our colleagues in the Center on Education Policy and Workforce Competitiveness (EdPolicyWorks). We also thank Tim Sullivan, Veronica Kimani, Sean Geraghty, and others at NewGlobe for partnering with us on this project, Lee Crawfurd for generously sharing the assessments used in Crawfurd et al. (2021), and Shannon Kontalonis for excellent administrative support. This work has been supported by Innovations for Poverty Action (IPA) through the Research Methods Initiative (grant code NWU0004-X1), and the Center for Reinventing Public Education (CRPE). The authors received IRB approval the University of Virginia, protocol number 3751. This trial was pre-registered at the AEA RCT Registry (number AEARCTR-0006913) after the data collection was completed but before analysis and transfer of data. Declarations of interest: Daniel Rodriguez-Segura has accepted a job at NewGlobe, the research partner for this project. Daniel's new role started after the analysis of the data and the writing of this paper.

I. Introduction

The 2020 COVID-19 outbreak induced almost universal school closures around the world, placing heightened emphasis on methods for monitoring student learning at a distance. Even prior to these shutdowns, tracking academic outcomes has historically been challenging for out-of-school children, children living in remote areas or who are mobile, and those experiencing humanitarian or natural disasters. These monitoring difficulties add a layer of complexity to the process of designing policy, especially in developing countries for students experiencing learning disruption. This is particularly worrying given that even prior to the pandemic the World Bank estimated that over half of all children in low- and middle-income countries (LMIC) could not read a short passage with minimal comprehension by the end of primary school—often described as the "learning crisis." Some project that this share could increase to almost two thirds post-pandemic (World Bank, 2020a; Azevedo, 2020). Still, the unprecedented nature and magnitude of these school closures, along with the diverse set of governmental responses, add a layer of uncertainty about the state of learning outcomes during and after the pandemic, overall and for different subgroups of students. As such, the measurement of academic skills for hard-to-reach populations is a key first step towards the design of interventions that minimize the long-term consequences of COVID-19 and other educational disruptions on educational equity in the developing world.

Given the high degree of market penetration of cell phones even in some of the world's poorest areas, phone-based assessments (PBAs) emerge as a potential tool to assess learning levels for hard-to-reach groups. Researchers including Angrist et al. (2020a), Crawfurd et al. (2021), and Schueler and Rodriguez-Segura (2021) have already used these assessments as outcomes in the context of impact evaluations, and both Angrist et al. (2020b) and Luna-Bazaldua et al. (2021) provide logistical and empirical recommendations for researchers and policymakers assessing children by phone. In spite of this rise in the use of PBAs, to the best of our knowledge, there has not yet been much examination of the validity of these assessments. How representative are PBA samples? How well do PBAs appear to discriminate between students of different ability levels? How well do they assess the constructs they are intended to measure? Do they work better for assessing certain subjects than others? These questions are the focus of our paper.

A careful examination of phone-based assessment validity is valuable given the potential advantages and disadvantages associated with PBAs relative to in-person assessments. The clearest advantage of PBAs is that they allow assessors to reach students remotely through a fairly ubiquitous technology. As a result, PBAs allow assessors to be centrally located without the need for travel time between assessments, significantly reducing costs. For example, the total cost of the data collection process for the PBAs that we study here was USD 2.1 per completed assessment. Researchers have found that the

costs of phone surveys in Sub-Saharan Africa may decrease total costs by one or two orders of magnitude per completed assessment (Zezza, et al., 2021; Crawfurd et al., 2021). This feature is particularly enticing for increasing the potential reach of large-scale formative assessments growing in popularity in LMICs. In the most extreme cases, such as during school closures, remote assessments like PBAs, may be the only practical way to reach certain students safely.

On the other hand, PBAs also have potential disadvantages that could threaten the extent to which the information obtained through them is useful for policy design. First, cell phone use – although high – is not yet universal in the developing world. Phone numbers may more often change for families experiencing other forms of disadvantage. As such, PBAs might disproportionately exclude the most disadvantaged students who, in turn, might be those that policymakers are most interested in assessing. Relatedly, there may be serious logistical challenges to reaching parents at a convenient time, when their children are nearby, and when children are in a physical and mental space conducive to being accurately assessed, as documented for a PBA conducted in Sierra Leone (Sam-Kpakra, 2021). Another potential disadvantage is that PBAs may lend themselves to cheating, or to friends and relatives solving the exercises for the children, also documented in Sierra Leone. In fact, this worry might grow if the perceived importance of and stakes attached to PBAs also increase. Finally, PBAs may not allow for the measure of all skills and subjects equally. For example, assessing some skills may require visual prompts not available during a PBA or requiring distribution. Therefore, in spite of the promise of PBAs, logistical challenges and empirical questions remain that should be investigated further before PBAs are used more broadly and to interpret results from studies that have already used PBAs as outcome measures.

In this paper, we address these gaps in the literature by examining the validity of a phone-based assessment measuring foundational literacy and numeracy (FLN) for primary school students in Kenya. To do so, we leverage a unique longitudinal, student-level dataset that includes, for the same students, data from two PBAs—one of math and one of literacy skills—along with baseline test scores obtained via in-person testing in math, English, and reading fluency on an exam for which significant validity evidence has already been collected. Importantly, we are able to link standardized in-person and PBA scores for a large sample of the same third, fifth, and sixth graders. As such, this paper's contribution to the literature is to provide empirical evidence on the validity, strengths, and limitations of PBAs as tools to accurately measure FLN across multiple subjects.

We find evidence that deliberate attention to prioritizing groups that are less likely to be reached via phone is needed to achieve a representative sample of students assessed with PBAs. We also find that PBAs, while on aggregate positively correlated with in-person performance, are significantly noisier than repeated in-person assessments. Despite this, the

numeracy PBA, but to a lesser extent the literacy PBA, is generally as rank-preserving as a repeated in-class assessment. In other words, the PBAs did not systematically misclassify students on the relative distribution of performance. As a result, we conclude that PBAs appear to be appropriate tools to track aggregate measures of numeracy and literacy, such as in the context of impact evaluation research but may not be reliable tools to track individual-level student performance or for the purpose of individual-level high-stakes decision-making without further development. Additionally, PBAs do not appear equally effective at measuring all sub-skills. For example, we find evidence that more research is needed before phone-based fluency assessments are adopted widely, if at all.

II. Why measure foundational literacy and numeracy?

While there is no technical, agreed-upon definition for the concept of foundational literacy and numeracy (FLN), it is usually understood as the set of basic skills children must master to learn higher order concepts at school, and which are intended to be developed in the first years of primary education (Evans and Hares, 2021). Given their "foundational" nature and the fact that 4 in every 10 children in LMIC do not go to any sort of pre-primary institution (World Bank, 2020b), the concept of FLN typically overlaps substantially with the primary school curriculum in the early grades. This fact has been used to operationalize assessments measuring FLN, as FLN do not represent a single construct, but rather a set of "sub-skills" within each area which constitute the basis for their "mastery". In numeracy, this usually covers counting, logical operations and inequalities, and basic operations with a few digits. For literacy, this usually includes alphabetic knowledge, word reading and recognition (i.e., "decoding"), oral language, and reading fluency, which jointly constitute the basis for reading comprehension, or the ability to make sense of a text.

The concept of foundational literacy and numeracy (FLN) has gained prominence in the past decade in research and policy circles (Evans and Hares, 2021). This increasing visibility has arisen partly because of poor learning outcomes documented for children across several LMIC in even the most basic building blocks of FLN. Given that many students in LMIC are no longer enrolled in school by lower secondary—as many as 17% of adolescents or 60 million in these countries (UNICEF, 2019)—the development literature has placed great emphasis on ensuring that if these students do leave the system, they do so having at least mastered FLN. The concept of FLN has also gained popularity through its close connection to policy targets set forth by donor agencies like the World Bank, through the concept of "learning poverty" (Azevedo, 2021). As such, these metrics are used as policy goals not only because of their intrinsic value as to reflect overall FLN levels, but also because they are understood to be summary statistics of broader systematic features relating to learning in other subjects, enrollment, and equity. Therefore, it is essential that leaders

have tools for accurately assessing and tracking FLN, and for measuring the impact of policies and programs designed to improve these important gateway skills.

To that end, we focus on measuring foundational literacy and numeracy achievement. We also explore the measurement of reading fluency as a literacy sub-skill by itself, as oral reading fluency has been gaining traction among governments as a metric to track system-wide literacy achievement, such as in the case of Kenya and Liberia. Moreover, fluency is also an interesting and important research metric in its own right, being the closest literacy sub-skill representing the more colloquial meaning of "being literate" (Rodriguez-Segura et al., 2021). In fact, fluency is strongly correlated with reading comprehension, and also with other more basic literacy sub-skills (Jiménez et al., 2014), making fluency a good proxy for broader literacy[1]. Similarly, reading fluency is typically quantified in "natural" units of correct words per minute (cwpm), allowing for direct sample comparisons in the original units of measurement across contexts and testing rounds. Therefore, it emerges as a potential sub-skill which policymakers would be interested in isolating during the administration of phone-based assessments. Fortunately, we have baseline data to examine this construct from the aggregate language score by itself.

III. Study context

This study was conducted in Kenya, where the initial round of COVID-19 induced school closures lasted for a total of 37 weeks, with some variation by grade level (UNESCO, 2021). The research occurred in partnership with the organization NewGlobe which operates Bridge Kenya, a network of low-cost private schools. NewGlobe also operates or supports several other programs, including both public schools operated as part of public-private partnerships and low-cost private schools in several LMIC. Our study covered students across 105 low-cost private schools in Kenya. This sample includes 29 of the 47 Kenyan counties, and all eight of the areas previously considered "provinces." There is a wide range in the socioeconomic characteristics of the locations covered. The local multidimensional poverty rate (at a 5 km radius from the school) ranges from 7.8% at the 10th percentile in our sample, to 59.2% at the 90th percentile. Although students in these schools and their families are relatively disadvantaged on a global scale, they are likely more socioeconomically advantaged, on average, than typical families enrolled in Kenya's public schools. For example, nationally, 27 percent of families report the mother having no formal education while this is true for only one percent of our sample.

[1] For example, the correlation between the [English-language] reading fluency and comprehension subtasks on the Stanford Achievement Test was 0.91. Comparable correlations between fluency and comprehension have been found for assessments conducted in numerous other languages. (Bulat et al., 2014; LaTowsky, Cummiskey, & Collins, 2013; Management Systems International, 2014; Pouezevara, Costello, & Banda, 2012, etc.)

IV. Data

Collection of phone-based assessment data

We begin by describing the data collection process and then the assessment design. The collection of PBA data was carried out in two waves, first for numeracy and then literacy. Although the goal of both waves was to collect learning outcomes for primary school students by phone, the logistical details for the two waves differed in some ways, partly due to explicit design choices and partly to the shifting educational landscape during this period of unprecedented disruption. We present all differences between the waves in Table 1 and discuss below factors that might affect the interpretation of results across waves.

The first wave of data collection, measuring numeracy, happened over 18 days in December of 2020 while face-to-face learning was on hold, after 9 months of school closures during which no student in our sample had attended school in person. The numeracy data from this PBA also served as a set of outcomes for an impact evaluation of a remote instructional program via mobile phones that happened alongside our methodological study (Schueler and Rodriguez-Segura, 2021). The sample includes students in grades 3, 5, and 6 across all 105 schools. Due to budget constraints and response rate projections based on an earlier pilot, we selected a simple random sub-sample of students to be assessed by phone from baseline performance blocks of students from all schools (6,295 students out of 8,319 in the impact evaluation sample). Each of the 6,295 students was randomly assigned to an assessor at the individual level and the compliance rate to this assignment was 98.8% for the numeracy wave. Of the 6,295 students on call lists, 2,644 were ultimately reached and assessed. Unfortunately, we cannot distinguish between those who were called but not reached, called and reached but declined to answer, and those that assessors were not able to call due to our hard stop date. These facts should give readers pause before interpreting 2,644 divided by 6,295 (42%) as an expected typical response rate for PBA.

The second wave of data collection, measuring literacy outcomes, lasted approximately 5 weeks between April and May 2021 after schools had reopened for in-person learning. This round of data collection is different from the numeracy wave in three additional ways. First, this round targeted only grade 3 students, as opposed to grades 3, 5, and 6 for numeracy. This was because measuring reading fluency outcomes required students to have access to standardized and relatively long passages (>100 words) of printed text, and only third graders in our partner's network had access to a homework book with text that met these requirements. With only one grade we did not need to select a sub-sample to stay within budget, which allowed for all 2,218 third graders to be included on the literacy assessment call list. Ultimately, assessors reached 1,082 students, for a response

rate of 49% — though for reasons stated above, we interpret this as a lower bound estimate of possible PBA response rates.

The second difference arises in how we prioritized reaching students on the call list. For the numeracy assessment, once students were selected into the sample, they all had the same level of priority to be called. Although the selected sub-sample (n=6,295) was representative of the universe of students (n=8,319), the sample of students that were actually reached by phone and assessed in numeracy (n=2,644) did not end up being representative of the whole universe, as we show in Appendix Table 1 and discuss in greater detail in our results section below. In light of this, our calling prioritization for the literacy assessment was more deliberate in the hope of reaching a more representative sub-sample. Specifically, we gave calling preference to students from backgrounds that were under-represented during the numeracy assessment by using calling blocks with different shares of students based on observable demographic characteristics. We began with the full roster of 6,295 students for the math assessment and ran a logistic regression with a binary indicator for whether a student was reached as the dependent variable, and with grade-, school-, assigned assessor-, and decile of individual baseline performance-fixed effects as the independent variables. Using this model, we predicted each child's propensity to be reached for the literacy assessment and aggregated these into average propensity scores for grade 3 classes. We separated classes into quintiles of "propensity to be reached", where quintile 1 represented students in classes least propense to be reached (i.e., most under-represented in the numeracy round), and quintile 5 included students who were most over-represented in the numeracy round. Finally, we then created five "calling blocks." Assessors were instructed to complete a full block before moving on to the next block. The first blocks were composed of more students who were less likely to be reached[2].

The final key difference between the two waves is that while the assessors for the numeracy round were all Bridge teachers, our partner hired 11 external assessors to conduct the literacy assessments. This was partly due to the fact that in-person classes had resumed by the time the literacy PBA were conducted, so our partner was reluctant to have teachers conduct this task on top of teaching duties. The assessors were all bilingual in English and Kiswahili, had at least a post-secondary education, and were trained by Bridge employees in the same way that numeracy assessors were trained. We achieved a high compliance rate of the random assignment of assessors to students of 90.3% for the literacy wave. We lay out all differences between the numeracy and literacy data collection waves because they could complicate our ability to draw direct comparisons between numeracy and literacy PBA on the basis of subject matter alone. We address this possibility in more detail below.

[2] Specifically, the first calling block had 30% of students from quintile 1, 25% from quintile 2, 20% from quintile 3, 15% from quintile 4, and 10% from quintile 5. For the second block, the lean towards less propense students was less pronounced, starting at 25% from quintile 1, and 15% from quintile 5, with a gradient for the other quintiles in between. Block 3 assigned equal shares for all students, and blocks 4 and 5 gave preference to more propense students.

Phone-based assessment design

The PBAs were designed to assess the basic numeracy and literacy constructs that the development literature has emphasized through the use of major assessments like ASER/Uwezo, the Early Grade Reading Assessment (EGRA), and the Early Grade Mathematics Assessment (EGMA) (Evans and Hares, 2021; Mbiti et al, 2019; RTI International, 2015). The structure for some items was borrowed from these assessments, and other adaptations by teams such as Crawfurd et al. (2021). Ultimately, the numeracy PBA consisted of 14 questions. We intentionally designed the assessment ex ante to have two distinct sections: "core numeracy" questions, which included 9 items that were shared by all grades, and "curriculum-aligned" questions, which included 5 items more in line with what students would have been learning in class. The curriculum-aligned questions were the same for grades 5 and 6, but different for grade 3. The literacy PBA was designed around sub-skills that would be easier to measure over the phone and that would be age-appropriate for grade 3 students. Specifically, we focused on spelling, vocabulary, oral comprehension, and fluency. The literacy PBA consisted of 12 questions, although two of these consisted of longer fluency questions, making the expected duration of the assessment comparable to the numeracy assessments. We include the actual instruments in the Appendix. Additionally, at the end of both phone-based assessments, we included a short student and parent survey, with one question for students, along with few questions for parents on at-home study habits, COVID-related shocks, and their educational attainment.

Our original estimate for the duration of a completed assessment call was of ~20-25 minutes, informed by our partner and previous studies using phone-based assessments. Given families' time and our budget limitations, we implemented a "stop rule" where students would not be asked further questions if they got three questions wrong in a row. The average assessor successfully assessed approximately 13 individual students per day. If assessors worked an 8-hour day, which we believe is a slight underestimate, that comes to approximately 37 minutes per successfully assessed child (including breaks, unsuccessful calls, etc.). Ultimately, the number of reached students fell below our intended target, which we suspect happened for at least two reasons. First, qualitative data recorded by teachers and assessors suggests that it may have been hard to get the students' and/or parents' full attention for this call and therefore took longer than we expected to get through the questions. Second, the students that were assessed generally performed well on the questions, which limited the extent to which the stop rule was used.

For numeracy, we graded assessments by grade level, both in the aggregate and by section (i.e., core numeracy and curriculum-aligned), in two ways: first, the number of correct responses as a share of the total number of questions in the test, and second, using a basic 2-parameter IRT model, which is our preferred outcome because of the more appropriate weighting of items by difficulty and discrimination that this method provides.

However, the correlation between these two scores is 0.98, and we confirm that it makes little empirical difference which approach we choose. We standardize both types of scores within grade. The grading of the literacy assessment was different, given that the outcomes of the vocabulary and fluency sections were continuous scores that do not lend themselves to IRT models, and as such we simply standardized these raw outcomes by section. We graded the spelling and oral comprehension sections using an IRT model, which were then separately translated into standardized scores. Finally, all four standardized scores are averaged into a single literacy score. To conduct further robustness tests, we also create an average score which does not include the fluency scores.

In-person assessments and additional data

One of the main strengths of our research design is the availability of a rich set of consistent and standardized baseline measures of performance from in-person assessments that can be linked at the student-level to PBA results. Our partner's instructional model is highly centralized, and technology-led, which has two key implications for this study. First, while classes are in session, all students in the same grade across all schools within one of our partner's programs (in this case, Bridge Kenya) take the same assessment at the same time. Second, the students' achievement data is uploaded into a centralized system with unique student and school IDs which track student performance over time. This technology-led approach allows us to have extraordinary access, especially for a LMIC context, to in-person achievement data at baseline.

NewGlobe goes to great lengths to ensure these assessments are relevant for the context, capture skills appropriate to grade-level, and provide enough variance to distinguish between different learning trajectories. The in-person written exams are a mix of "market exams" designed by teachers within organizations in Kenya approved by the National Research and Development Centre (NRDC) and which make use of grade-appropriate curricula provided by the Kenya Institute of Curriculum Development, international assessments (e.g., DIBELS fluency subtask) in the case of reading fluency, and exams that are internally written by the specialized NewGlobe instructional team in Kenya. The market exams and reading fluency items were initially piloted to ensure that they were appropriate for the context. Although within a testing round all classes of a given grade across all schools in Kenya take the same version of exam, the exact version that is administered is either repeated every year with small tweaks to the questions, or alternates between two versions every-other year. The outcome data that emerge from these assessments are constantly monitored and analyzed to ensure that the administration was successful, and that the tests do yield meaningful variation in the outcomes. When issues have come up in the past, this information has been shared with the instructional team in Kenya to make the proper adjustments, and as such, these exams have been refined over

time. Bridge Kenya had been using a similar type assessment for over 10 years at the time of this study, although the centralized collection of data began a few years after the start of the assessments, and the mix of market-developed and internally developed tests has varied over time. In all, these data quality standards give us confidence in the appropriateness of the in-person assessments as a valid and reliable benchmark of student-level achievement in numeracy and literacy. At the very least, they represent a policy-relevant measure of student achievement.

For these in-person exams, we have access to results from the previous two rounds of assessments administered before each wave of PBAs. In practice, this means that for numeracy, we have access to math scores from February, 2020 (10 months before the PBA) and October, 2019 (14 months before the PBA). For literacy, including reading fluency, we have English and fluency scores from March, 2021 (1 month before the PBA), and February, 2021 (2 months before the PBA). For each subject, we use the closest in-person assessment in time to the PBA as the "main" baseline assessment, and the other score as the "repeated assessment", which are used as a benchmark to understand how similarly repeated, in-person scores behave. The difference in timing between assessments for literacy and numeracy, particularly the longer periods between each round for numeracy, is another key distinction to contextualize the results. In other words, if longer time windows between assessments decrease the correlation between these assessments, the relationships between numeracy assessments should be understood as weaker than they would have been had they happened closer together in the relative timeline (similar to the literacy assessments). For all subjects, we standardize scores within grade and testing round.

Additionally, upon returning to in-person school in January, 2021, students were given an in-person recreation of the numeracy PBA, which we call the "in-person PBA". In other words, this in-person assessment had the same number of questions, and the same skills tested in the same order as the PBA. The items were identical to the questions asked in the PBA, except for the digits in each question. In other words, while question 4 in the PBA was 62+14, question 4 in the in-person recreation may have been 60+16. Unfortunately, our partner had some technical issues that hindered our ability to collect data for grade 5 students, and to further link most observations from the in-person recreation of the PBA to most students' PBA. However, we are still able to use these data to provide suggestive evidence to disentangle the simultaneous changes of the PBA relative to baseline in-person assessments (e.g., the difference in number of questions included), and the mode of assessment.

Finally, in terms of covariates, we know for each pupil their grade and school, gender, age, average attendance rate for their class, and years attending schools within our partner's system. For each numeracy assessor, we know the grade and school where they teach, their age, gender, average attendance rate, and lesson completion rates. For each

school, we know its location, the surrounding population count, adult female literacy rate, and poverty rate at a 5 km radius, the average pupil-teacher ratio, the female-to-male student ratio, enrollment, and school measures of relative performance.

V. Research questions and empirical approaches

To explore the properties and best-uses of phone-based assessments, we are interested in understanding whether phone-based assessments are a valid tool for measuring learning outcomes remotely. Below we outline our methodological plan for answering each of our research questions.

Did the PBAs generate a representative sample of assessed students?

We begin by examining the extent to which the assessed sample was representative of the full roster of students who assessors were asked to call by conducting balance tests, separately for each subject, along individual and community-level covariates between those who did end up being assessed and those who did not. We know that the call list rosters were representative of the full sample of students included in the field experiment but here we compare those rosters to the lists of students who were ultimately reached and for which we therefore have PBA results, separately for the numeracy and literacy PBA. We also include grade- and school-level fixed effects to check the robustness of these differences. From these balance tests, we draw inferences on the broader categories along which the two samples do or do not display systematic differences. Finally, we compare the representativeness of the PBA samples when we did and did not employ efforts to oversample students with characteristics most likely to be underrepresented in a phone-based data collection.

Did the PBAs have discriminating power and internal consistency?

Next, we leverage the individual item-level data to understand the test properties of our PBAs. Our goal is not to validate this exact version of the assessment or items, but we still consider these test properties informative for the design of future assessments. To understand the discriminating power and difficulty of each item, we use item characteristic curves (ICCs), one of the main tools of item response theory (IRT), and their corresponding item information function (IIFs) plots to visually inspect whether items presented relatively large differences in discriminating power and difficulty. ICCs display the probability of answering a question correctly at varying levels of the latent construct "student ability". We therefore examine the shape and position of these curves. Items with a "steeper" middle section of the curve have a higher "item discriminating parameter", as they can more cleanly distinguish student ability. The horizontal position of the curve speaks to the difficulty of items, as the further to the right a curve is located, the higher the "item difficulty parameter",

as it takes a higher level of "ability" to answer the question correctly. Similarly, the corresponding IIF simply plots the "steepness" of the corresponding item characteristic curve at every level of ability, that is, the first-order derivative of the ICC. If the IIF is fairly flat, there is no sharp increase in the ICC at any point, and hence, the discriminating power of the item is low. If instead, an IIF has a large "peak", then this item has high discriminating power – at least relatively speaking. The visual inspection of the item characteristic curves and item information functions will allow us to better understand not only the performance of specific items, but the type of sub-skills that provided the most and least discriminating power in our sample. To understand the internal consistency of the assessment, we calculate Cronbach's alpha for the assessments as a whole, but also by grade and sub-section (e.g., "core numeracy" or "spelling")[3] to determine how tightly correlated the individual items on the assessment are with one another.

To what extent do PBAs accurately reflect in-person performance?

We also explore the extent to which PBA scores were correlated with baseline student performance on in-person assessments to see whether there is evidence for the "construct validity" of the PBAs. In other words, we seek to examine whether PBAs appear to be measuring what they are intended to measure (Strauss and Smith, 2009). To do so, we regress standardized in-person baseline scores by subject on standardized PBA scores at the individual student level. Importantly, we would not expect PBA scores to be perfectly correlated with in-person scores, even if PBAs accurately reflect in-person performance, for at least two reasons. First, the in-person and phone-based assessments did not test students on exactly the same skills. The baseline assessments were in-class assessments of curriculum-aligned content for each subject. Instead, the PBAs tested both curriculum-aligned subjects and core numeracy. Even within the curriculum-aligned section of the PBAs, students were not asked exactly the same items, as there were differences in what was "curriculum-aligned" at the time of assessment. Second, the number of items during the phone-based assessment was less than half the number of items for in-person assessments, potentially reducing the scope for discrimination between different levels of performance. Therefore, we seek to benchmark the magnitude of the relationship between the main in-person assessment and the PBA by also understanding the magnitude of the correlation between the main in-person assessment and a repeated in-person assessment. We therefore repeat the regression exercise but replacing the PBAs from the independent variable with the repeated in-person assessment such that we are correlating two in-person assessment scores with each other. This is helpful given that there was also a difference in actual content between the repeated in-person assessment and the main assessment, which

[3] We calculate Cronbach's alpha by sub-section with the caveat that, mechanically, this metric is higher for assessments with a larger number of items, all else being equal.

allows us to benchmark the PBA against an assessment that also did not test exactly the same skills as the main in-person assessment. We also conduct robustness checks by running the same analyses, but only using the curriculum-aligned portion of the PBAs.

We also explore whether the PBA appear to measure their intended constructs for some portions of the baseline performance distribution better than others. To do so, we create "gap" measures between each student's percentiles (by grade and subject) in the main in-person assessment vs. the PBA and the main in-person assessment and the repeated in-person assessment. We then calculate three separate gap measures $A \in$ {PBA, Repeated in-person assessment} by taking the difference:

$$\text{Gap } A_i = (\text{Percentile on main in-person assessment})_i - (\text{Percentile on assessment A})_i$$

From these measures, we can interpret that a positive gap means that student i's relative position in A is lower than their relative position in the main in-person assessment. In other words, a positive gap means that, roughly speaking, assessment A is "underestimating" the relative position of student i. The inverse is also true, as a negative gap indicates that assessment A is overestimating student i's performance.

Did PBAs misclassify student performance at a higher rate than in-person assessments?

Another reason we do not want to rely on the correlations between PBAs and in-person results as the sole evidence of validity is that the PBAs and in-person assessments were not administered concurrently. As a reminder, there were ten months between the numeracy PBA administration and the most recent in-person baseline math achievement and one month between the literacy in-person baseline assessment and the literacy PBA. This may depress the correlations between PBAs and in-person assessments, particularly if other factors affected student achievement in between the two assessments, even if the PBAs are measuring what they are intended to measure. This is particularly likely in the case of the numeracy PBA as the COVID-19 outbreak and associated school shutdowns occurred in time between the in-person assessment administration and the PBA data collection. Therefore, we also study whether PBAs severely misclassify the relative performance group for a student, or the extent to which the PBAs displays "rank-preserving properties". In other words, are PBAs significantly more likely to (erroneously) show that a low-performing student is a high-performing student (or vice-versa) than repeated in-class assessments? The idea here is that, even if the two assessments are not perfectly correlated, if the PBA preserve the relative ranking of students, it suggests that they are still measuring what we think they are measuring. To investigate this, we create a measure of "misclassification", where for student i, assessment A, and misclassification threshold X:

$$\text{Misclassified at } X_i = 1(|\text{Gap } A_i| > X)$$

In other words, we code a student i as being "misclassified" by an assessment if their relative position on this assessment is further than X percentiles (in either direction) from their relative position in the main in-person assessment. For example, if a student is at the 30th percentile in the main in-person assessment, 10th percentile in repeated in-person assessment, and 35th in the PBA, their (absolute) gap for the repeated in-person assessment is 20, and their gap for the PBA assessment is 5. Under a misclassification threshold of 10, this student would have been coded as misclassified by the repeated in-person assessment but not by the PBA. Under a threshold of 3, they would have been misclassified by both assessments. In a sense, the misclassification threshold can be thought of as a level of "tolerance" for relative misclassification of an assessment, with respect to the main baseline assessment.

One threat to the validity of our approach to examining misclassification as evidence of validity would be if the factors affecting student achievement in the intervening time between the in-person assessment and the PBA differentially affected different students in such a way that the rank order was not preserved (even if PBA measure what they are intended to measure). The math assessments would be more susceptible to this phenomenon for two reasons. First, there was more time that passed between the in-person test and PBA for math (10 months) than literacy (one month) and second, the in-person math assessments were administered prior to COVID-19-induced school shutdowns and it is quite possible that COVID-19 influenced student achievement in a variety of ways. Again, this is not a problem if COVID-19 affected all students similarly but could disrupt rank-ordering if it differentially affected students. It is also not a problem if COVID-19 differentially negatively affected already-low performers in such a way as to widen performance gaps while preserving relative ranks. It would only be a problem if it affected students in such a way as to shift relative ranks. For example, if average-performing students came from families with a greater likelihood of economic disruption than low-performing students because they had more to lose in the resulting recession, this could cause them to fall in the relative rankings. To address this possibility as an additional robustness check, we conduct our analyses for the numeracy sample separately for students from families who do and do not report COVID-19-related disruptions as of December, 2020.

VI. Results

The assessed PBA sample was not representative without intentional oversampling

We begin by describing how representative the sub-samples of students who were successfully reached and assessed by PBAs were of the broader universe of students. After

the first wave of data collection for the numeracy PBA, we find that the sample of students that enumerators ended up assessing was not fully representative of the universe of students from which they were drawn, even though the call lists of students that assessors received during the numeracy wave was indeed representative. More specifically, assessed students live in more populated areas (0.09 SD), with more literate female adults (0.07 SD), and less poverty (-0.09 SD) than students from the call list who were not assessed, as we show in Appendix Table 1. Similarly, the attendance rates of teachers and students in assessed students' schools is higher (0.08-0.10 SD), and these students also attend larger schools (0.06 SD). In terms of baseline performance, assessed students scored 0.10 SD higher in English than non-assessed students, and 0.05 SD higher in math and Kiswahili, though the differences in math and Kiswahili are imprecisely estimated. In short, the numeracy PBA sample represented a relatively more advantaged subset of students.

In light of these results, we used a prioritization mechanism during the literacy PBA administration to determine whether a more representative sample could be generated when using more intentional sampling techniques (described in the Data section above). Broadly speaking, we find that this effort was largely successful. As we show in Appendix Table 1, we observe no statistically significant differences between students from the call list who were and were not assessed in the literacy wave. The one exception is that reached students attended schools that were 26-48 meters closer to a cell phone tower (shown as a 0.06 SD smaller distance from cellphone tower in Appendix Table 1). In short, with an intentional sampling technique, we were able to generate a representative sample for the literacy PBA on all dimensions except, perhaps intuitively, distance from a cellphone tower.

PBA had discriminating power and internal consistency

Next, we turn to basic descriptive patterns in the PBA scores. We find that the numeracy assessment behaved relatively well in terms of expected patterns of correct response rates, internal validity, and item discrimination for most items. The average percent correct score in the numeracy PBA as a whole was 72% (SD=0.21). Although scores on the overall assessment were relatively high, collectively there was a range of item difficulty. The average score on the word problem having the minimum score of any item (25%) and the counting problem having the maximum (96%). As we would expect to see if our instrument was measuring what we intended it to measure, scores were higher on average for the more foundational "core numeracy" questions (79%) than for the more advanced curriculum-aligned questions (58%). Similarly, scores generally increased with grade level for all core numeracy questions and most curriculum-aligned questions, with exceptions between grade 3 and 5 (plausible given the content changed between these two particular grades in this section).

The total numeracy scores had strong internal consistency with a Cronbach's alpha of 0.81. The core numeracy section had a higher alpha than the curriculum-aligned questions even when examined separately by grade level. This could reflect the fact that the core numeracy section was almost twice as long as the curriculum-aligned section. Once disaggregated by section and grade, the maximum Cronbach's alpha was for the curriculum-aligned section for grade 3 (0.79), and the minimum was the curriculum-aligned section for grade 6 (0.64). In contrast, the performance of the literacy PBA was much more variable depending on the sub-skill, as we show in Table 2b. The spelling section, although it displayed a relatively high mean performance of 70%, had Cronbach's alpha 0.51, an unreliable coefficient based on the measurement literature. On the other hand, the fluency exercises displayed a very high Cronbach's alpha of 0.96 and a mean of 61 correct words per minute (SD=58), while oral comprehension questions displayed a mean of 0.43 (SD=0.33), with a lower alpha of 0.66, on par with the lowest sub-group for the numeracy assessment. Finally, on average students were able to name 7 words for the vocabulary question (SD=3). Given that the literacy PBA was only given to grade 3 students, we cannot compare performance across grades as we do for numeracy.

In terms of specific item performance, we see a wide range in the difficulty and discrimination parameters of each item. We visually show in Appendix Tables 2a, b and Appendix Figures 1a, b, c, the wide variance in the discriminating power of items within subjects, and the fact that while some items behaved well in terms of this feature (e.g., spelling "fight" or one digit multiplication for grade 3 students), some items displayed close to no discriminating power. For example, the word problem was too difficult for grade 3 students, counting being too easy for grade 6 students, or spelling the word "children" which was too easy even for grade 3 students. These findings suggest that the quality of PBAs lies not only on the logistics of the assessment, but also on the assessment instrument per se, and how well suited it is to successfully measure learning outcomes for the population of interest.

PBA had greater construct validity at the aggregate than individual level

We now turn to describing the validity of the PBAs, or in other words, how well they reflect previous in-person performance and therefore appear to measure what they are intended to measure. First, we show in the top row of Table 3 that for numeracy, a 1 SD change in PBA score is predictive of a 0.14 SD change in the main baseline in-person assessment, with a high degree of statistical precision. Instead, a 1 SD change in the October, 2019 in-person assessment (the "repeated in-person" assessment) predicts a 0.59 SD change in the main baseline in-person assessment, with a high degree of statistical precision. In other words, the numeracy PBA was about a quarter as predictive of in-person

performance as repeated assessments[4]. This fraction is very similar for the literacy PBA score (22%), but not so for the fluency assessment, as the PBA fluency measurement is basically uncorrelated with in-person assessments of fluency[5]. In spite of the lower predictive power of PBA scores relative to repeated in-person scores, the PBA scores in numeracy and literacy do predict in-class performance with a high degree of statistical precision at the $p<0.000$ significance level. We take this as initial evidence that these PBAs may possess some validity to measure learning in aggregate in the absence of in-person assessments. Still, given the significant lower correlations, they do not emerge as perfect replacements of in-person assessments, especially for the tracking individual learning trajectories.

Ideally, we would be able to attribute the lower correlations between PBAs and in-person scores fully to the mode of assessments. However, the differences in the content and length of the PBA vs. in-person assessments, make it difficult for us to tease apart the extent to which this lower predictive power of PBAs comes from the mode of assessment, and how much comes from these other differences. To isolate the mode of assessment, our partner gave children an in-person exam which had the same items as the numeracy PBA, with slightly different digits in each item, right after students came back to school in 2021. Due to technical issues in the data collection process, we can only link these scores for 106 students[6]. When we correlate the PBA with in-person PBA scores, we obtain a coefficient of 0.63, significant at the 0.01 level, meaning that a 1 SD deviation change in the "in-person PBA" is correlated with a change of 0.63 SD in the main in-person assessment. This is much higher than the coefficient of 0.14 obtained using the PBA as the explanatory variable, and even larger than the coefficient of 0.59 obtained using the repeated in-person assessment. Despite the limited sample, this is suggestive evidence that the mode of assessment, as opposed to the other differences between the in-person assessment and the PBA, was the feature that drove the different magnitudes in the correlations. This builds on top of work by Crawfurd et al. (2021), which provides more conclusive evidence for the large role that the mode of assessment plays in determining individual-level performance.

[4] Note that this relationship is the same if the dependent and independent variables are included as z-scores or as percentiles.

[5] This is the case regardless of whether we use the standardized version of these variables, or whether we use this variable in its "natural" units of correct words per minute. In fact, the difference between the average in correct words per minute during the in-person assessment and the PBA is over 60 words – which we take to be a clear disconnect between the constructs measured by both assessments. Beyond the lack of reliability of PBAs, another possibility is that the baseline in-person fluency data is also fundamentally wrong. Unfortunately, without additional on-site auditing of this data collection process, we were unable to say more about this, especially since the distribution of both the PBA scores and the main in-person assessment scores are technically plausible under typical reading fluency norms for grade 3 students. However, the very high correlation between the two in-person fluency assessments (coefficient of 0.75) re-assures us that, at least in relative terms, the in-person assessments have "accuracy" in the sense of test accuracy, and that PBAs do not accurately measure the same construct that the in-person assessments do.

[6] This sample is, for the most part, representative of the broader call lists for numeracy, with the important exception of baseline scores, where the students for which we can link their PBA scores with their in-person recreation of the PBA perform significantly weaker (~0.3-0.4 SD) than their counterparts.

Given PBA scores are significantly less correlated with the main in-person assessment results than repeated in-person assessment results are to each other, the next question of interest is the extent to which PBA result in differences in students' relative positions in the performance distribution compared to in-person assessments, and for what portions of the baseline performance distribution this mismatch is most pronounced. To do so, we leverage the "gap" metric that we describe in the previous section, which measures for each child the distance in percentiles between a given assessment and the main in-person assessment. As a starting point, we visualize the distribution of these gaps for the PBA and the repeated in-person assessment in Figure 1. We find that while the distributions of these gaps are mostly centered around the 0 percentile, the gaps for the PBAs display a much wider variance than the gaps for the repeated in-person assessments. The wider variance of these gaps for the PBA, and the fact that they are roughly centered around 0, support the previous finding that on average, the PBA does benchmark learning *in aggregate* similarly yet in a noisier manner for specific students than repeated in-class assessments.

To locate the performance sub-groups for which these gaps tend to happen, Figure 2a shows the average gap per baseline percentile in the main in-person assessment. Note that, by default of how the plot is constructed, some positive trend is expected[7]. We create a benchmark for this plot using the gap between the main in-person and the repeated in-person assessments. In general, the line shows gaps in the PBA is steeper than the benchmark: it is lower than the benchmark for the low portions of the distribution, and higher for the high portions of the distribution. In other words, PBAs are overestimating relative performance for low performers, and underestimating relative performance for high performance. To further reinforce this point, in Figure 2b, we estimate the gap in the gaps displayed in the top panel (i.e., the difference between the two curves, at the individual level). This panel also displays a slight positive trend, which highlights that there is some overestimation for low performers at baseline, and some underestimation for high performers at baseline. Therefore, this is evidence that in this case, PBAs forced a more "equal" distribution of achievement levels than in-person assessments, and hence had less discriminating power as a whole, relative to repeated in-person scores.

PBA preserved relative performance rankings

Given our findings that PBAs are significantly less predictive of baseline performance than repeated in-class assessments, and that PBAs underestimate high performers and overestimate low performers, we are also interested in learning whether PBAs systematically misclassify students' relative performance. Using the definition of relative misclassification

[7] In other words, the range for the baseline performance is 1-100, but this was also the first term in the subtraction with which the gap was calculated. Using the first percentile in February, 2020 as an example, the gap was at most 0, as the lowest that the second term could have been was 1. This fact, coupled with some expected reversion to the mean, makes us indeed expect some positive relationship for any assessment A.

that we describe in the previous section, we present the misclassification rates for numeracy, literacy, and reading fluency, for thresholds between 0 and 50 in Figure 3. Overall, the misclassification rates for numeracy are very similar, even being the same at a threshold of 28. As we show in Appendix Figure 2, this finding remains constant if we repeat this analysis separately for those students who report a household shock caused by COVID-19 during school closures, and those who do not. This is an encouraging finding as one could hypothesize that external shocks to these households might alter students' relative position and hence threaten the extent to which our baseline distribution holds – which we do not find to be the case here. The misclassification rates for literacy are similar, but not as similar as the numeracy rates. This is partly driven by the better performance of repeated in-person assessments for literacy, relative to the repeated in-person assessments for numeracy[8]. This is evidence that, even though PBAs do tend to be noisier predictors of in-class performance compared to repeated in-class assessments, the extent to which they misclassify students' relative performance is similar to repeated in-class assessments, and hence can be a valuable tool for relative (low-stakes) relative classification of students' performance.

VII. Discussion

The growing need to track learning outcomes remotely for globally disadvantaged populations during school closures and ongoing pandemic-induced disruptions to in-person learning has brought significant attention to phone-based assessments. Yet, to scale PBAs more broadly to inform policy, researchers need to understand better certain key features of these assessments, such as whether they can accurately measure learning and for whom. In this paper, we shed light on these issues by leveraging a unique set of baseline, in-person test scores and a set of PBA data on learning outcomes for the same students. Our findings are a first step towards understanding for what purposes PBAs are best suited, and what can be done to improve these assessments.

The implication of our first finding is that weak sampling adjustments may not be enough to reach and assess a representative sub-sample of students when administering PBAs. To achieve this goal, we had to explicitly design a calling system during our second wave of data collection that strongly prioritized students who were not likely to be reached. Indeed, this comes as confirmation that the high penetration rates of cell phones in LMIC do not guarantee that all children will be equally likely to be reached during a PBA. Instead, there are still individual- and community-level factors that drive the likelihood each student is successfully assessed. In our case, perhaps expectedly, the likelihood to be assessed during the first wave of data collection was correlated with higher socioeconomic status and baseline in-person performance, and as such, our results regarding the validity of the

[8] Unfortunately, the misclassification rates for fluency are much higher than for numeracy or literacy. This is driven by both a better performance of repeated in-person assessments and a worse performance of the PBA.

numeracy PBA only hold in a somewhat non-representative sample. Future research should examine the validity of numeracy PBA among students least likely to be assessed by phone. Although we managed to correct this by the second wave of data collection, this should be a warning for policymakers and practitioners that even one of the most ubiquitous technologies in the developing world may exclude the most disadvantaged populations if explicit efforts to correct this fact are not put in place. Luckily, our findings suggest that basic attempts to deliberately oversample underrepresented groups can successfully address this problem.

We also find that the discriminating power and internal reliability of phone-based assessments varied substantially by item and test sub-section. In general, the different rates of discriminating power by item, and internal validity by section suggest a need to thoughtfully consider the contents of items included in phone-based assessments. While all assessments, phone-based and otherwise, are subject to differences in the discriminating power of their items, the logistics of PBAs need to be especially responsive to this, particularly due to considerations around the financial and opportunity costs of PBAs. The financial costs stem from the fact that PBAs are one-on-one assessments, and as such, enumerator time has to be invested on each item over as many students each enumerator assesses. Furthermore, airtime for phone-based interventions, including assessments, is typically charged by the minute or the second, and as such, there is a very obvious marginal cost to each item. In addition to these costs, the opportunity cost of each item stems from the fact that PBAs tend to have a tighter time constraint than in-person surveys or assessments. In our case, we aimed for 20-minute calls with an upper limit of 25 minutes. For every construct included, there was another element which was likely excluded from the assessment.

In hindsight, our counting item proved too easy even for grade 3 students, our word problem proved too hard for even grade 6 students, and it is unclear whether we learned much from the vocabulary question. These items did not provide enough discriminating power and likely crowded out other items from sub-skills, such as two-digit addition or oral comprehension for grade 3 students, that provided significant discriminating power. Even when the goal is to create a vertically-aligned exam to compare across grades, there is likely a set of sub-skills that provide more variation than the extremely easy or extremely hard items that we included on both ends of our numeracy assessment. To address this issue, piloting the assessment and adapting it to the local context seems particularly relevant to select items that display the smallest floor or ceiling effects, allowing for higher variance for the assessment as a whole, and hence for policymakers and researchers to learn more from these assessments.

Encouragingly for those interested in using PBAs to track aggregate performance, we find that PBAs, especially in numeracy, do not severely misclassify students' relative

performance compared to repeated in-person assessments. However, they are not as predictive of students' precise percentile performance as repeated in-person assessments. In other words, they do not systematically misclassify students on average, but do rank individual students relative to their baseline in-person ranks with more "noise" than repeated in-person assessments. This finding has clear implications for the future use of PBAs. For one, this type of assessment does not seem conducive to accurately track individual performance, much less if this is attached to high-stakes results, given the increased risk of individual performance "misclassification" relative to repeated in-person assessments. On the other hand, PBAs do seem to maintain relative performance rankings at similar rates than repeated in-person assessments. Therefore, to the extent that researchers and policymakers trusted in-person assessments in the past to yield aggregate results, especially in terms of relative performance such as in the case of an impact evaluation, PBAs appear to perform well in this respect.

Although phone-based assessments do not emerge as an ideal replacement for in-person assessments entirely, our study suggests that they do can serve as appropriate substitutes for in-person assessments, when designed and implemented properly, when in-person assessments are not logistically feasible, and when assessing aggregate rather than individual performance. For a successful collection of learning data from phone-based assessments, researchers and policymaking agencies must ensure that they carry the proper sampling strategy, assessment design, and content and item selection, so that the findings that emerge from these assessments may inform robust policy design for educational recovery worldwide.

VIII. References

Angrist, N., Bergman, P., Matsheng, M. (2020a). School's Out: Experimental Evidence on Limiting Learning Loss Using 'Low-Tech' in a Pandemic. Working Paper. https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=3735967

Angrist, N., Bergman, P., Evans, D. K., Hares, S., Jukes, M. C. H., & Letsomo, T. (2020b). Practical lessons for phone-based assessments of learning. BMJ Global Health, 5(7), e003030. https://doi.org/10.1136/bmjgh-2020-003030

Azevedo, J. P. (2020). Learning Poverty: Measures and Simulations. World Bank Policy Research Working Paper. No. 9588

Azevedo, J.P., Goldemberg, D., Montoya, S., Nayar, R., Rogers, H., Saavedra, J., Stacy, B.W., William. (2021). Will Every Child Be Able to Read by 2030? Defining Learning Poverty and Mapping the Dimensions of the Challenge. World Bank Policy Research Working Paper. No. 9588

Bulat, J., Brombacher, A., Slade, T., Iriondo-Perez, J., Kelly, M., & Edwards, S. (2014). Projet d'Amélioration de la Qualité de l'Education (PAQUED): 2014. Endline report of Early Grade Reading Assessment (EGRA) and Early Grade Mathematics Assessment (EGMA). Prepared for USAID under Contract No. AID-623-A-09-00010. Washington, DC: Education Development Center and RTI International.

Chronbach, L. & Meehl, P. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281-302.

Cohen, J., & Goldhaber, D. (2016). Building a More Complete Understanding of Teacher Evaluation Using Classroom Observations. Educational Researcher, 45(6), 378–387. https://doi.org/10.3102/0013189X16659442

Crawfurd, L., Evans, D. K., Hares, S., & Sandefur, J. (2021). Teaching and testing by phone in a pandemic. Working paper.

Durrant, G.B., Groves, R.M., Staetsky, L., Steele, F. (2010). Effects of interviewer attitudes and behaviors on refusal in household surveys. Public Opinion Quarterly. 74: 1–36

Evans, D., & Yuan, F. (2020). How big are effect sizes in international education studies? CGD Working Paper 545.

Evans, D., & Hares, S. (2021). Should Governments and Donors Prioritize Investments in Foundational Literacy and Numeracy? CGD Working Paper 579.

Jiménez, J., Gove, A., Crouch, L., Rodríguez, C. (2014). "Internal structure and standardized scores of the Spanish adaptation of the EGRA (Early Grade Reading Assessment) for early reading assessment." Psicothema. 26(4):531-7. https://doi.org/10.7334/psicothema2014.93.

LaTowsky, R.J., Cummiskey, C., & Collins, P. (2013). Egypt grade 3 Early Grade Reading Assessment baseline. Draft for review and comment. Prepared for USAID under the Education Data for Decision Making (EdData II) project, Data for Education Programming in Asia and the Middle East (DEP-AME) task order, Contract No. AID-278-BC-00019. Research Triangle Park, NC: RTI International.

Luna-Bazaldua, D., Jiberman, J., Levin, V. (2021). Assessing outside of the "classroom box" while schools are closed: the potential of phone-based formative assessments to support learning continuity. Education for Global Development. World Bank Blogs.

Management Systems International (MSI). (2014). Early Grade Reading Assessment baseline report. Balochistan province. Prepared for USAID under the Monitoring and Evaluation Program (MEP), Contract No. AID-391-C-13-00005. Washington, DC: MSI. http://pdf.usaid.gov/pdf_docs/PA00KB9N.pdf

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (2019). Inputs, incentives, and complementarities in education: experimental evidence from Tanzania. The Quarterly Journal of Economics, 134(3), 1627-1673. https://doi.org/10.1093/qje/qjz010

Olson, K. P. A. (2007). Effect of interviewer experience on interview pace and interviewer attitudes. Public Opinion Quarterly. 71: 273–86

Pouezevara, S., Costello, M., & Banda, O. (2012). Malawi National Early Grade Reading Assessment survey. Final assessment – November 2012.Prepared for USAID under the Malawi Teacher Professional Development Support (MTPDS) program, Contract No. EDH-I-00-05-00026- 02; Task Order No. EDH-I-04-05-00026-00. Washington, DC: Creative Associates International, RTI International, and Seward, Inc. http://pdf.usaid.gov/pdf_docs/PA00JB9R.pdf

RTI International. (2015) Early Grade Reading Assessment (EGRA) Toolkit, Second Edition. Washington, DC: United States Agency for International Development. https://pdf.usaid.gov/pdf_docs/PA00M4TN.pdf.

Sam-Kpakra, R. (2021). Sierra Leone – Distance learning during COVID-19: Qualitative Research – Final Report. Center for Global Development. https://www.cgdev.org/sites/default/files/CGD-background-paper-sam-kpakra-sierra-leone-distance-learning.pdf

Schueler, B. E., Rodriguez-Segura, D. (2021). A Cautionary Tale of Tutoring Hard-to-Reach Students in Kenya. EdWorkingPaper: 21-432. Retrieved from Annenberg Institute at Brown University: https://- doi.org/10.26300/43qs-cg37

Strauss, M. E., & Smith, G. T. (2009). Construct Validity: Advances in Theory and Methodology. Annual Review of Clinical Psychology, 5(1), 1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

United Nations Educational, Scientific and Cultural Organization Institute for Statistics. (2019). New Methodology Shows that 258 Million Children, Adolescents and Youth Are Out of School (UIS/2019/ED/FS/56; Fact Sheet). http://uis.unesco.org/sites/default/files/documents/new-methodology-shows-258-million-children-adolescents-and-youth-are-out-school.pdf

von Hippel, P. T., & Bellows, L. (2018). How much does teacher quality vary across teacher preparation programs? Reanalyses from six states. Economics of Education Review, 64, 298–312. https://doi.org/10.1016/j.econedurev.2018.01.005

von Hippel, P. T., Bellows, L., Osborne, C., Lincove, J. A., & Mills, N. (2016). Teacher quality differences between teacher preparation programs: How big? How reliable? Which programs are different? Economics of Education Review, 53, 31–45. https://doi.org/10.1016/j.econedurev.2016.05.002

World Bank. (2020a). Learning poverty in the time of COVID-19: a crisis within a crisis. Policy Brief. https://openknowledge.worldbank.org/bitstream/handle/10986/34850/Learning-Poverty-in-the-Time-of-COVID-19-A-Crisis-Within-a-Crisis.pdf

World Bank, World Development Indicators. (2020b). School enrollment, preprimary (% gross) [Data file]. Retrieved from https://data.worldbank.org/indicator/SE.PRE.ENRR

Zezza, A., Martuscelli, A., Wollburg, P., Gourlay, S., & Kilic, T. (2021). Viewpoint: High-frequency phone surveys on COVID-19: Good practices, open questions. Food Policy, 105, 102153. https://doi.org/10.1016/j.foodpol.2021.102153

IX. Tables and figures

Table 1: comparison of phone-based assessment characteristics for wave 1 and wave 2

Category	Feature of assessment	Wave 1 (Numeracy)	Wave 2 (Literacy)
Target population	Grades	3, 5, 6	3
	Sampling of students to be called	Simple random selection of sub-sample. Total number of students in roster was an arbitrary guess of how many assessors might reach.	All students in focal grade at baseline were included in the call lists.
	Number of students in assessors' initial roster	6,295 across all the grades	2,218
	Sample size of completed assessments	2,644 across all three grades (42% of initial roster)	1082 (49% of initial roster)
Constructs measured	Specific sub-skills	Counting, inequalities, addition, subtraction, multiplication, division, other grade-appropriate skills.	Spelling, vocabulary, oral comprehension, reading fluency
	Format for questions	All questions recorded as multiple-choice for correct/incorrect/no answer.	Questions for spelling and oral comprehension recorded as multiple-choice for correct/incorrect/no answer. Questions for vocabulary and reading fluency recorded as number of correct answers provided.
	Number of content questions	14	12 (including two longer fluency questions)
	Number of survey questions	1 for students, 5 for parents	1 for students, 3 for parents
	Scoring method	IRT model (2 parameter) by grade. Standardized to create a single numeracy score.	Separate IRT scores for sub-skills in the format of multiple-choice questions, and then standardized by subject. Other two sub-skills are just standardized. The standardized score for all four sub-skills is then averaged to create a single literacy score.
Logistics	Dates administered	December, 2020	April/May, 2021
	As an outcome in an impact evaluation	Yes	No
Enumerators	Number of enumerators	20	11
	Enumerators	Bridge teachers	Outside assessors, with high educational achievement
	Compliance rate to student assignment	98.8%	90.3%
	Assessor assignment	Randomly at the student-level. Order to be reached also randomly assigned.	Randomly at the student-level. Order to be reached also randomly assigned, but with prioritization of students from backgrounds that were under-represented in the numeracy assessment.
Baseline in-person scores	Dates	Math in-person scores from 10 and 14 months before assessment	ELA and reading fluency scores from 1 and 2 months before assessment
	Actual subjects for in-person scores	Mathematics	English Language Activities for literacy as a whole; reading fluency in units of correct words per minute

Table 2a: summary statistics of numeracy PBA, by section and grade

		Sample			
		All grades	Grade 3	Grade 5	Grade 6
Assessment	Full assessment	0.72	0.65	0.73	0.77
		(0.21)	(0.25)	(0.18)	(0.18)
		{0.81}	{0.85}	{0.77}	{0.76}
		N=2644	n=985	n=866	n=793
	Core numeracy questions	0.79	0.69	0.83	0.86
		(0.21)	(0.24)	(0.16)	(0.16)
		{0.76}	{0.77}	{0.68}	{0.68}
		N=2644	n=985	n=866	n=793
	Curriculum-aligned questions	0.58	0.58	0.55	0.61
		(0.32)	(0.35)	(0.31)	(0.29)
		{0.69}	{0.78}	{0.69}	{0.64}
		N=2644	n=985	n=866	n=793

Notes: each cell shows the average score obtained, the standard deviation in parentheses, Cronbach's alpha in curly brackets, and the number of observations for each subgroup.

Table 2b: summary statistics of literacy PBA

		Grade 3
Assessment	Oral comprehension	0.43
		-0.33
		{0.66}
		N=1082
	Spelling	0.70
		(0.28)
		{0.51}
		N=1082
	Vocabulary	7.15
		(3.06)
		.
		N=1081
	Fluency	57.91
		(23.97)
		{0.96}
		N=1047

Notes: each cell shows the average score obtained, the standard deviation in parentheses, Cronbach's alpha in curly brackets, and the number of observations for each subgroup.

Table 3: Correlation strengths between in-person and phone-based assessments

		Dependent variable: score on main in-person			
		Numeracy	**Literacy**	**Literacy (no fluency)**	**Fluency**
	Score on PBA	0.14***	0.17***	0.09**	0.02
		(0.02)	(0.05)	(0.04)	(0.03)
		[1956]	828	866	[825]
		{0.02}	0.02	{0.01}	{0.00}
	Score on repeated in-person	0.59***	0.78***		0.78***
		(0.04)	(0.04)		(0.03)
		[1387]	[767]		[765]
		{0.35}	{0.62}		{0.59}
Relative strength of PBA to repeated score		$\frac{0.14}{0.59}$= 24%	$\frac{0.17}{0.79}$= 22%		$\frac{0.02}{0.78}$= 4%

Notes: Coefficients resulting from regressing each combination of dependent and independent variable. Sample includes all students in the initial call lists, that is, the largest possible subset of students for which we have PBA scores. However, the results are robust to running the same analyses on constant samples to rule out that these differences are due to compositional changes across data collection rounds. In other words, the results are robust to re-analyzing the data subsetting only to students for whom we have all three observations within a subject (PBA, main in-person, and repeated in-person), as shown in Appendix Table 3a, and also subsetting to only students for whom we have PBA scores in literacy and numeracy, as shown in Appendix Table 3b. All assessments are normalized. Standard deviation in parenthesis. Regression observations in squared parenthesis. R^2 of each specification in curly parenthesis. * p<0.10, ** p<0.05. *** p<0.01.

Figure 1: Distribution of gaps in relative achievement between main assessment and other assessments

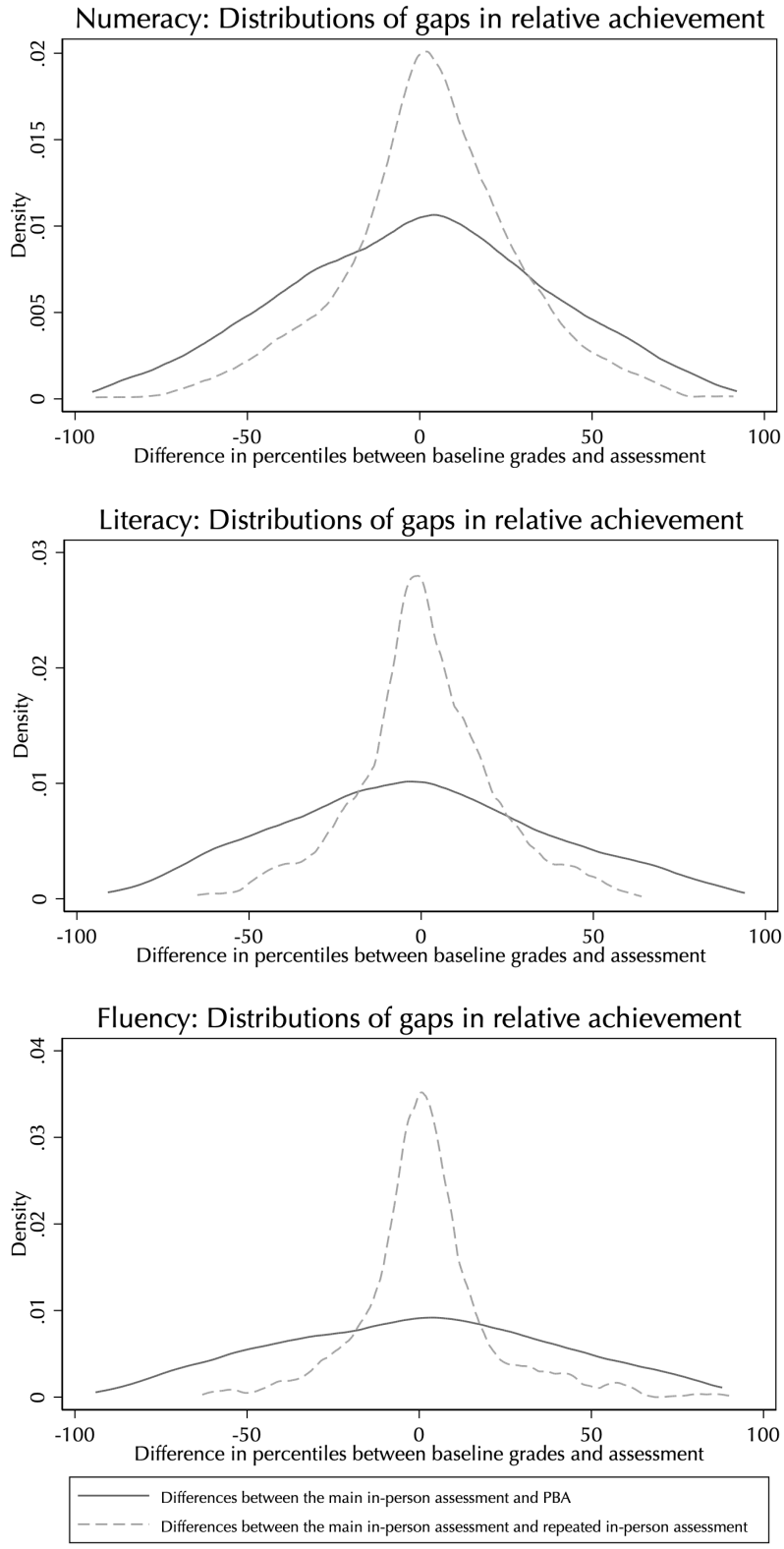
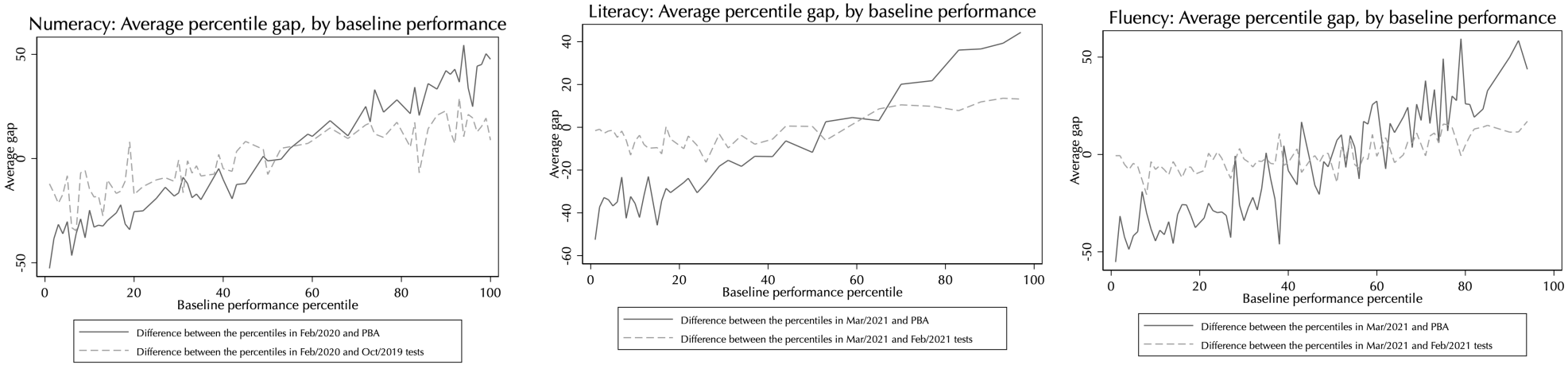
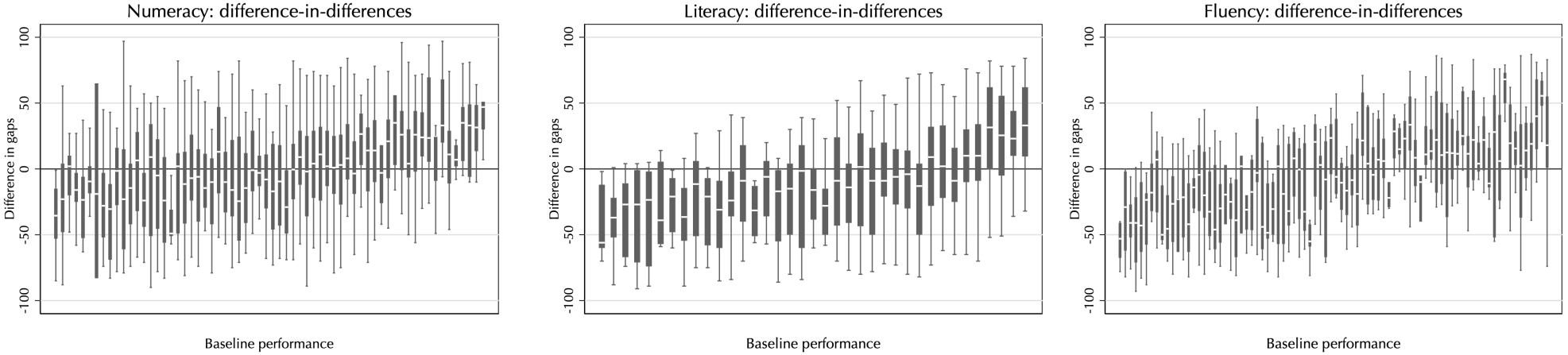


Figure 2a: Average percentile gap, by baseline performance and assessment



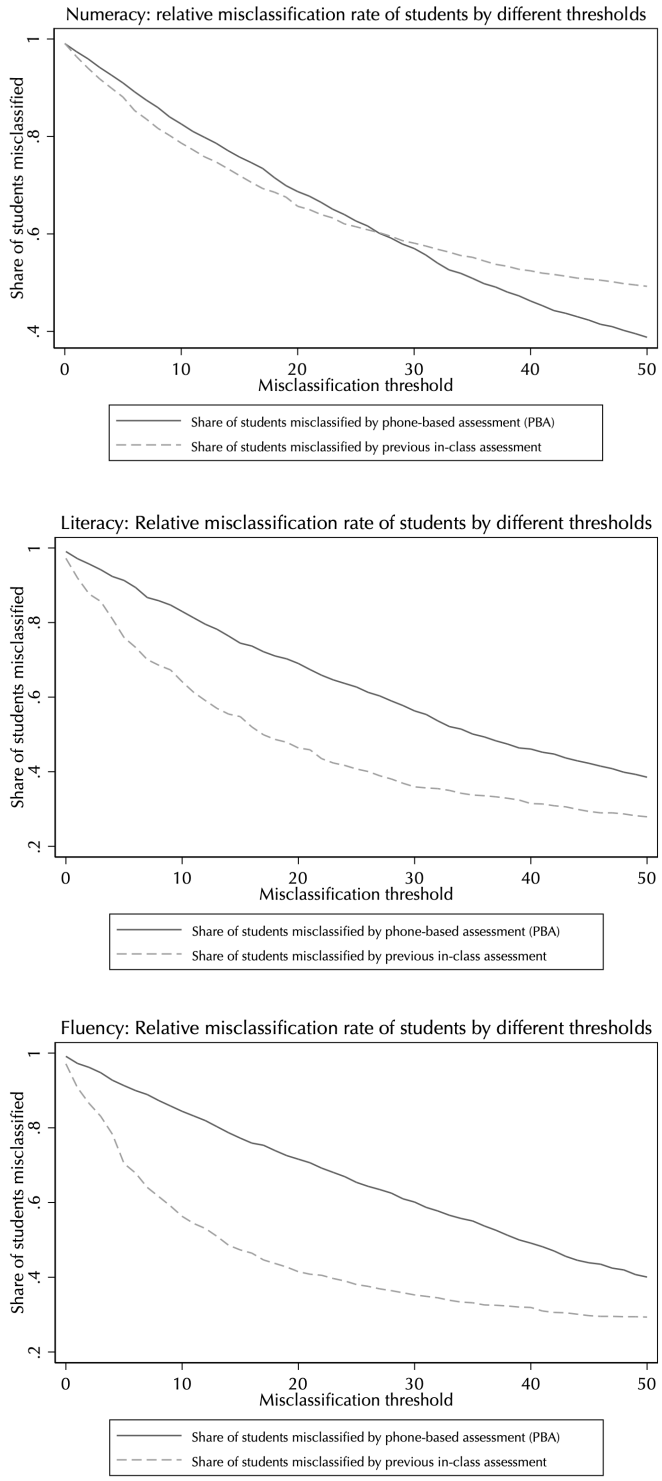
Notes: For each child, their percentiles in the PBA, main in-person assessment, and repeated in-person assessments is calculated. The solid line shows the difference between their percentile in the main in-person assessment and the PBA. The dashed line shows the difference their percentile in the main in-person assessment and the repeated in-person assessment.

Figure 2b: Average percentile gap, by baseline performance and assessment



Notes: For each child, we compute the difference between the gap in their percentile in the main in-person assessment and the PBA, and the difference between the gap in their percentile in the main in-person assessment and the repeated in-person assessment. For each baseline xtile feasible, the distribution of the difference in gaps is displayed.

Figure 3: Relative misclassification rate of repeated in-class assessment and PBA, by different thresholds and subjects



Notes: A child is recorded as misclassified if the absolute difference between their percentile in the February/2020 assessment, and their percentile in the other assessment is above each threshold. The horizontal axis displays the thresholds used to compute misclassification rates in each assessment.

X. Appendix

a. Appendix tables and figures

Appendix Table 1: differences between students reached and assessed through PBA, and students not assessed by enumerators

	Numeracy		Literacy	
	[1]	[2]	[1]	[2]
Population in surrounding community	0.09**	0.1**	-0.03	-0.04
	(0.03)	(0.03)	(0.04)	(0.05)
Adult female literacy in surrounding community	0.07**	0.06**	-0.04	-0.04
	(0.03)	(0.03)	(0.04)	(0.05)
Poverty rate in surrounding community	-0.09**	-0.09**	-0.01	-0.01
	(0.03)	(0.03)	(0.04)	(0.05)
Distance from school to cellphone tower	0.00	0.01	-0.06**	-0.11*
	(0.01)	(0.02)	(0.03)	(0.06)
School principal attendance rate	0.03	0.02	0.04	0.00
	(0.02)	(0.02)	(0.05)	(0.06)
Average teacher attendance rate at school of origin	0.08***	0.08***	-0.06	-0.09*
	(0.02)	(0.03)	(0.04)	(0.05)
Average pupil attendance rate at school of origin	0.10***	0.1***	-0.02	-0.05
	(0.03)	(0.03)	(0.04)	(0.05)
School enrollment	0.06**	0.06**	0.02	0.01
	(0.02)	(0.02)	(0.04)	(0.04)
Share of female students at school	0.00	0.00	0.02	0.00
	(0.03)	(0.03)	(0.04)	(0.05)
Enrollment in grades 3, 5, and 6	0.04	0.04*	0.05	0.05
	(0.02)	(0.02)	(0.04)	(0.05)
Share of females in grades 3, 5, and 6	0.00	0.00	0.02	0.02
	(0.03)	(0.03)	(0.04)	(0.05)
Mean pupil-teacher ratio at school	0.04*	0.04*	0.04	0.04
	(0.02)	(0.02)	(0.04)	(0.05)
Baseline Math score from February, 2020	0.05	0.05	-0.01	0.01
	(0.03)	(0.03)	(0.05)	(0.06)
Baseline Kiswahili score from February, 2020	0.05	0.04	0.01	0.04
	(0.03)	(0.03)	(0.05)	(0.06)
Baseline English score from February, 2020	0.1***	0.09***	-0.03	0.00
	(0.03)	(0.03)	(0.05)	(0.06)
Observations	6290	6290	2208	2192
Fixed-effects	None	Assessor, grade	None	Assessor

Notes: Coefficients come from regressing each characteristic on an indicator variable for whether student was successfully assessed through PBA or not. Sample includes all students present in February, 2020. * p<0.10, ** p<0.05. *** p<0.01.

Appendix Table 2a: correct response rates by item, grade, and sub-skill measured for numeracy

Section	Skill measured		Items		Sample			
	G3-G6				Full sample		G3	G5	G6
Core numeracy questions	Counting		Q1: Can you count from 20-30?		0.96		0.95	0.97	0.97
	Inequalities		Q2: Which is greater? 64 or 38?		0.87		0.82	0.89	0.91
	Two-digit addition		Q3: What is 62+18?		0.88		0.78	0.92	0.94
			Q4: What is 33+49?		0.85		0.74	0.90	0.93
	Two-digit subtraction		Q5: What is 43-20?		0.89		0.81	0.93	0.94
			Q6: What is 81-43?		0.70		0.53	0.79	0.82
	One-digit multiplication		Q7: What is 3x4?		0.88		0.78	0.93	0.93
	One-digit division		Q8: What is the result of 8 divided by 2?		0.85		0.75	0.90	0.93
	Word problem (multiplication)		Q9: Oil is 200 shillings per liter and rice is 100 shillings a kilogram. How much should I pay for 3 liters of oil and 4 kilograms of rice?		0.25		0.09	0.27	0.41
Section	**G3**	**G5/6**	**G3**	**G5/6**	**Full sample**		**G3**	**G5**	**G6**
Curriculum-aligned questions	Skip counting/patterns	Four-digit addition	Q10: Complete the following number pattern: 13, 19, ___, 31	Q10: What is 3487+2325?	0.42	0.79	0.42	0.76	0.82
	Three-digit addition	Four-digit subtraction	Q11: What is 145+213?	Q11: What is 4756-2149?	0.69	0.67	0.69	0.64	0.70
	Three-digit subtraction	Two-digit multiplication	Q12: What is 278-124?	Q12: What is 42x26?	0.61	0.49	0.61	0.46	0.53
	One-digit multiplication	Two-digit division	Q13: What is 8x5?	Q13: What is the result of 96 divided by 12?	0.65	0.69	0.65	0.67	0.71
	Two-digit division	Fraction addition	Q14:What is the result of 35 divided by 7?	Q14: What is the result of three sevenths (3/7) + one fourth (1/4)	0.54	0.26	0.54	0.21	0.31

Notes: responses were coded as a 1 if a child gave the right answer to a question, and as a 0 otherwise, such as in the case of a wrong answer, non-response, or if a question was skipped because of the skip rule.

Appendix Table 2b: average scores by item, grade, and sub-skill measured for literacy

Fluency	Fluency passage 1	59.08
	Fluency passage 2	56.58
Vocabulary	Name some foods that can be bought from the market. Try to name as many things as you can think of.	7.15
Spelling	Please spell the word "sick"	0.77
	Please spell the word "when"	0.70
	Please spell the word "flight"	0.50
	Please spell the word "children"	0.83
Oral comprehension	What did the pupils wear to the sports day?	0.65
	Why was Nuru getting better at football?	0.43
	What football skills did Nuru have?	0.23
	Why were the other pupils proud of Nuru?	0.41

Notes: responses for spelling and oral comprehension were coded as a 1 if a child gave the right answer to a question, and as a 0 otherwise, such as in the case of a wrong answer, non-response, or if a question was skipped because of the skip rule. The responses for the fluency passages and the vocabulary section were recorded as continuous variables in units of number of correct words.

Appendix Table 3a: Correlation strengths between in-person and phone-based assessments, subsetting to students who have all three observations (PBA, main in-person, repeated in-person) within each subject

		Dependent variable: score on main in-person			
		Numeracy	**Literacy**	**Literacy (no fluency)**	**Fluency**
	Score on PBA	0.14***	0.17***	0.11**	0.02
		(0.03)	(0.05)	(0.04)	(0.03)
		[1387]	767	767	[766]
		{0.02}	0.02	{0.01}	{0.00}
	Score on repeated in-person	0.59***	0.78***		0.78***
		(0.04)	(0.04)		(0.03)
		[1387]	[767]		[766]
		{0.35}	{0.64}		{0.59}
Relative strength of PBA to repeated score		$\frac{0.14}{0.59}$= 24%	$\frac{0.17}{0.78}$= 22%		$\frac{0.02}{0.78}$= 3%

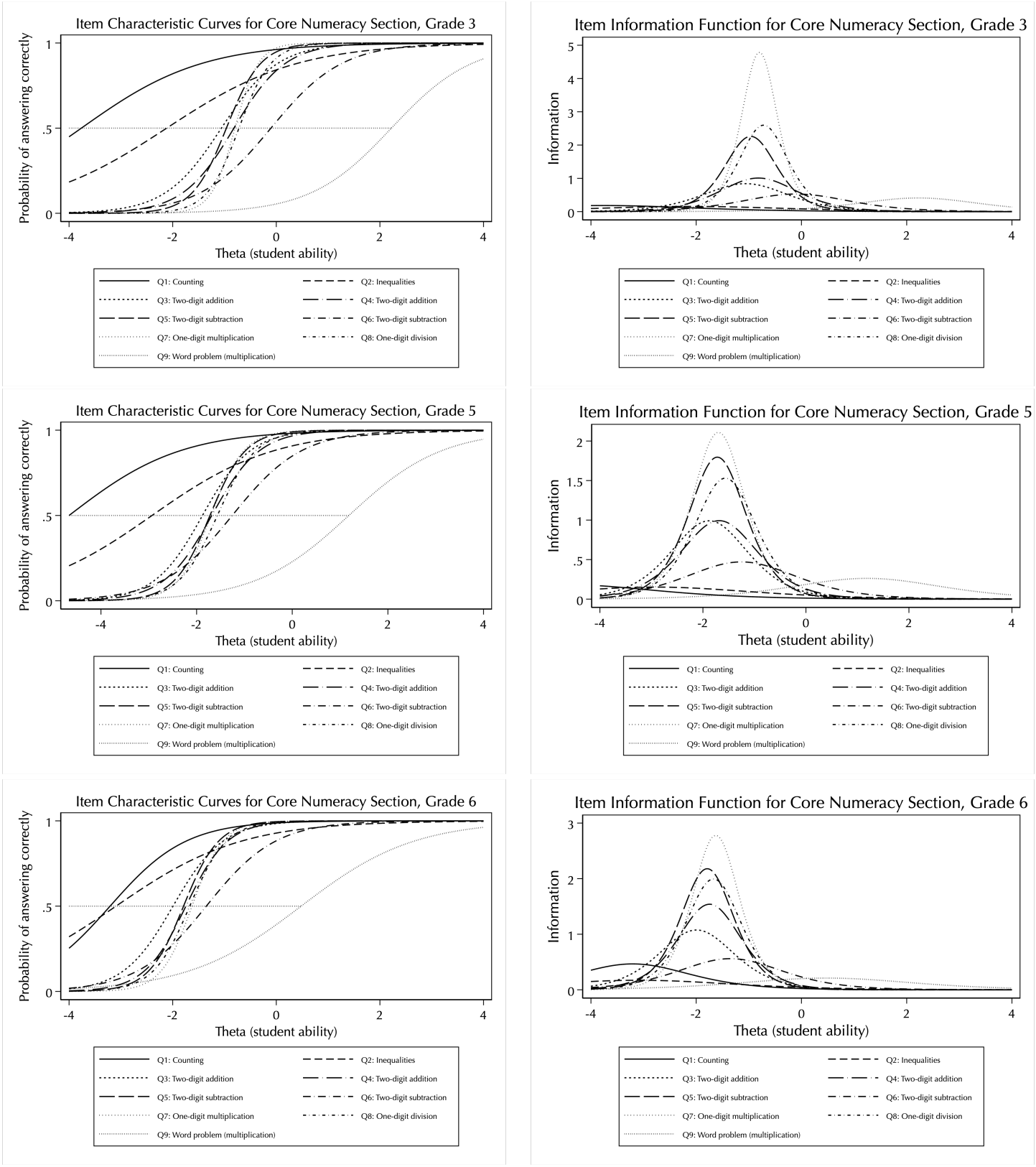
Notes: Coefficients resulting from regressing each combination of dependent and independent variable. All assessments are normalized. Standard deviation in parenthesis. Regression observations in squared parenthesis. R^2 of each specification in curly parenthesis. * p<0.10, ** p<0.05. *** p<0.01.

Appendix Table 3b: Correlation strengths between in-person and phone-based assessments, subsetting to students who have PBA scores for both numeracy and literacy

		Dependent variable: score on main in-person			
		Numeracy	**Literacy**	**Literacy (no fluency)**	**Fluency**
	Score on PBA	0.13***	0.24***	0.13**	0.05
		(0.05)	(0.08)	(0.01)	(0.06)
		[308]	312	312	[311]
		{0.02}	0.03	{0.01}	{0.00}
	Score on repeated in-person	0.50***	0.80***		0.79***
		(0.08)	(0.06)		(0.03)
		[214]	[290]		[289]
		{0.30}	{0.67}		{0.62}
Relative strength of PBA to repeated score		$\frac{0.13}{0.50}$= 26%	$\frac{0.24}{0.80}$= 30%		$\frac{0.05}{0.79}$= 6%

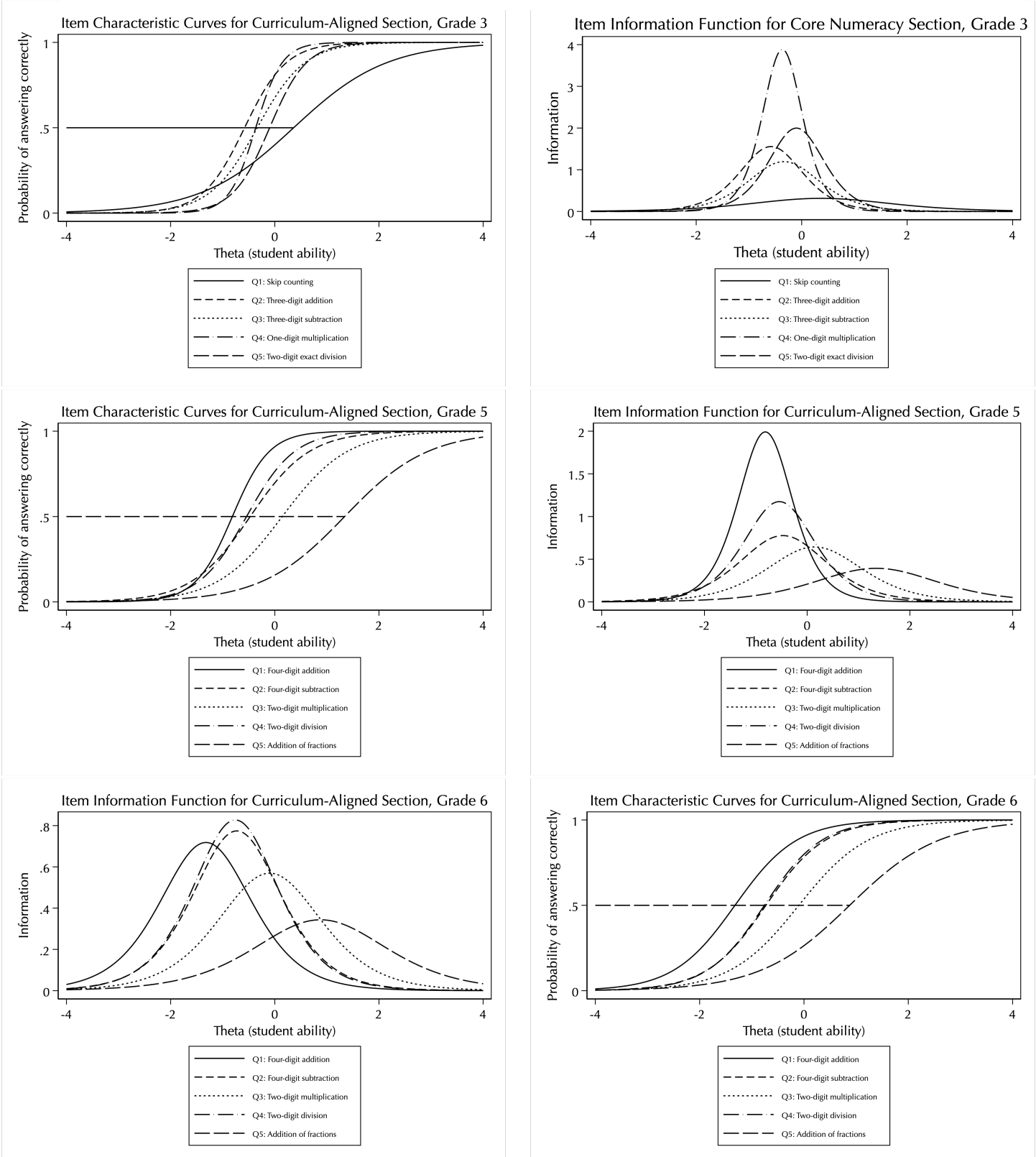
Notes: Coefficients resulting from regressing each combination of dependent and independent variable. All assessments are normalized. Standard deviation in parenthesis. Regression observations in squared parenthesis. R^2 of each specification in curly parenthesis. * p<0.10, ** p<0.05. *** p<0.01.

Appendix Figure 3a: Item characteristic curves, and Item information functions for core numeracy, by grade



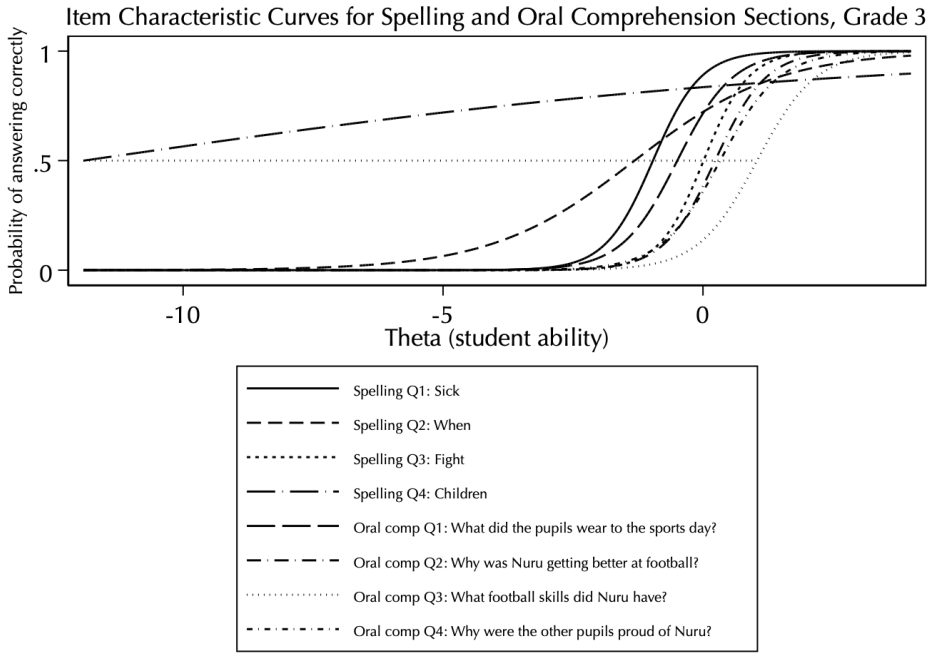
Notes: graphs estimated using a two-parameter IRT model

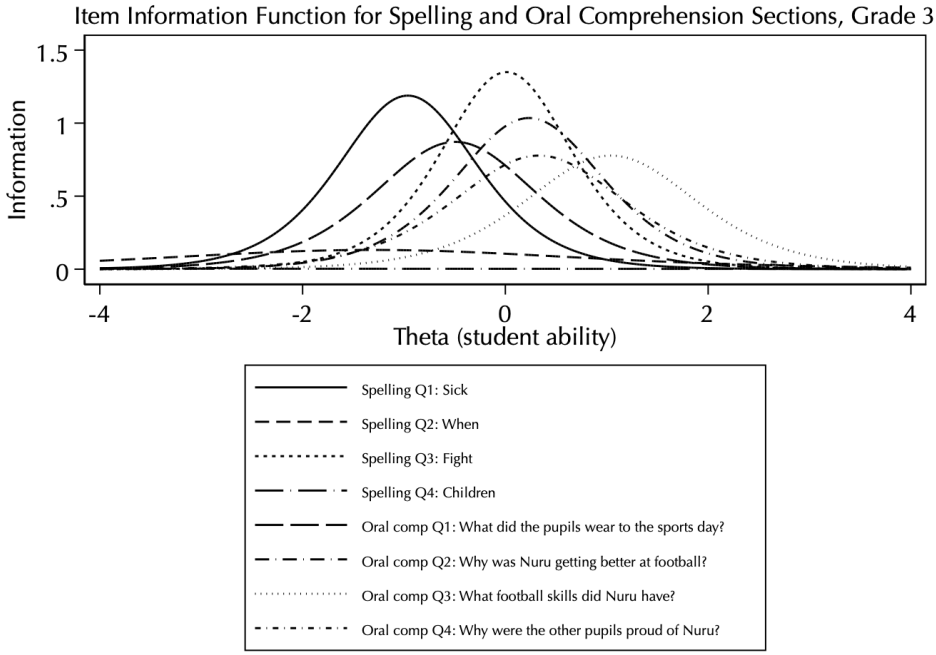
Appendix Figure 3b: Item characteristic curves, and Item information functions for curriculum-aligned questions, by grade for numeracy



Notes: graphs estimated using a two-parameter IRT model

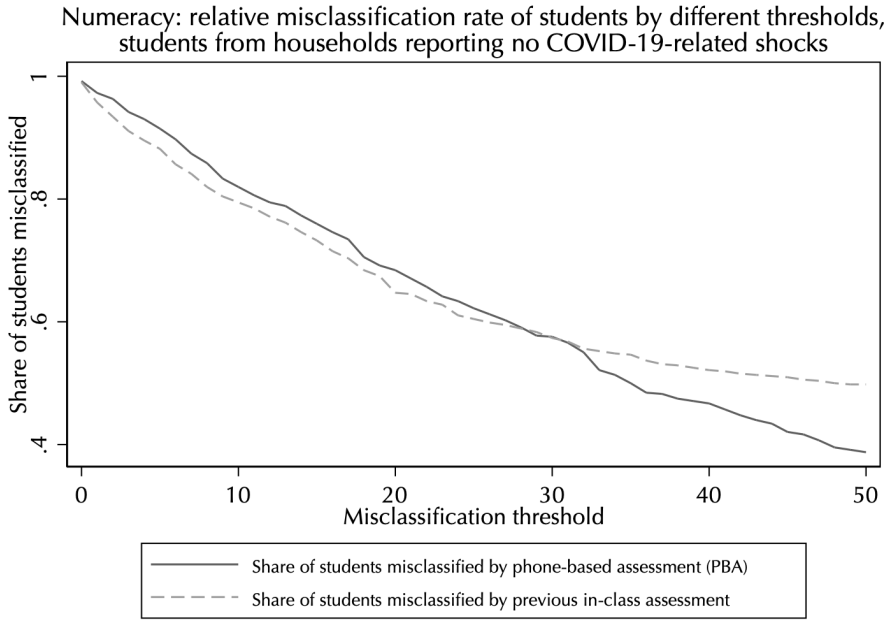
Appendix Figure 3c: Item characteristic curves, and Item information functions for literacy questions with binary outcomes, for grade 3 students

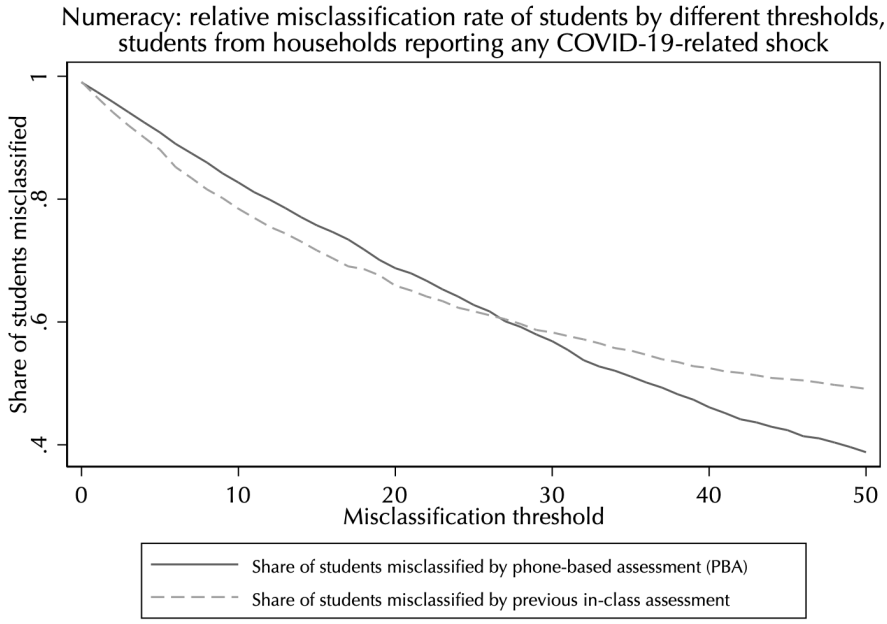




Notes: graphs estimated using a two-parameter IRT model

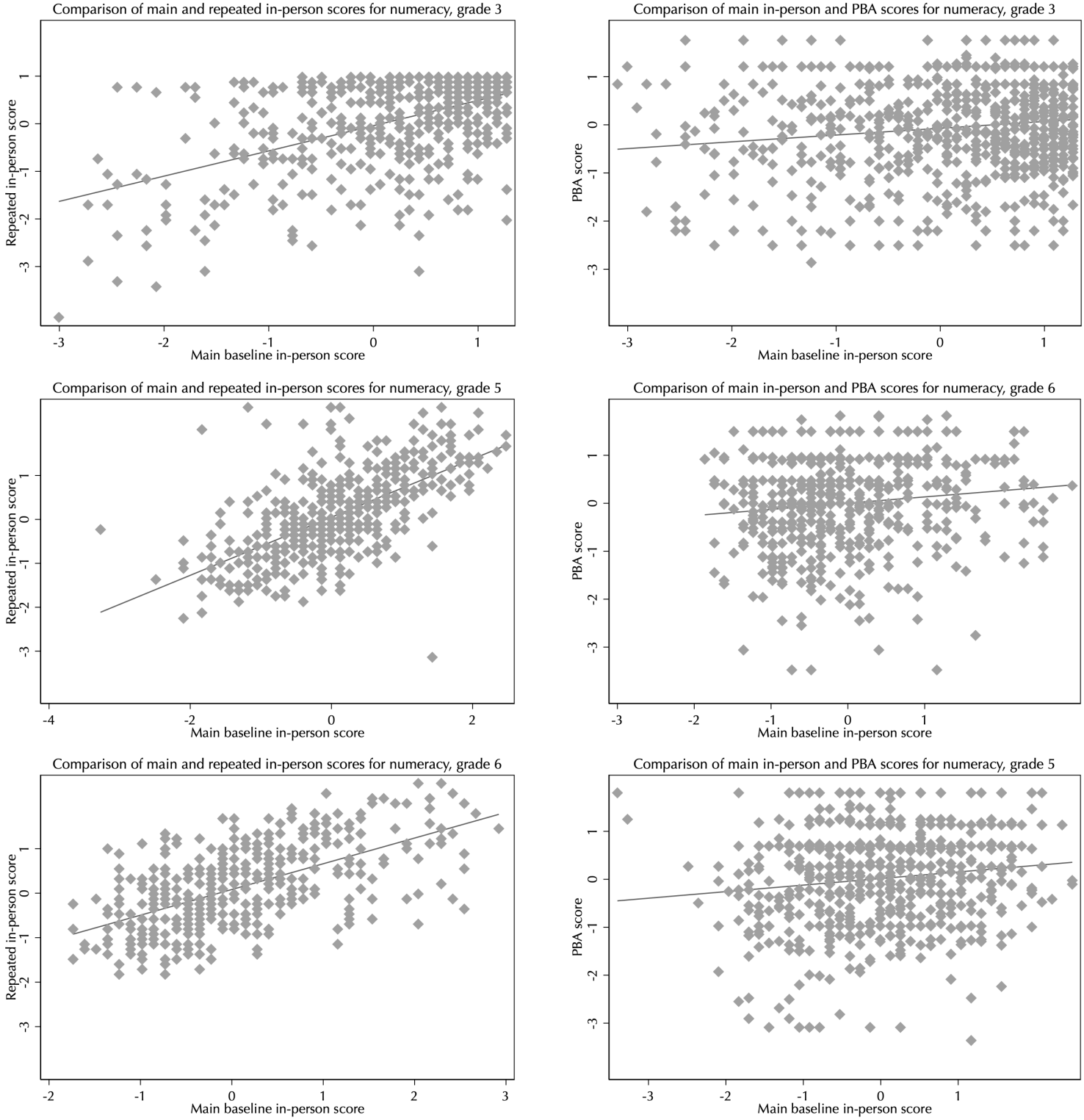
Appendix Figure 2: Relative misclassification rate of repeated in-class assessment and PBA, by different thresholds and subjects, separately for students from households reporting COVID-19-related shocks and for students who do not



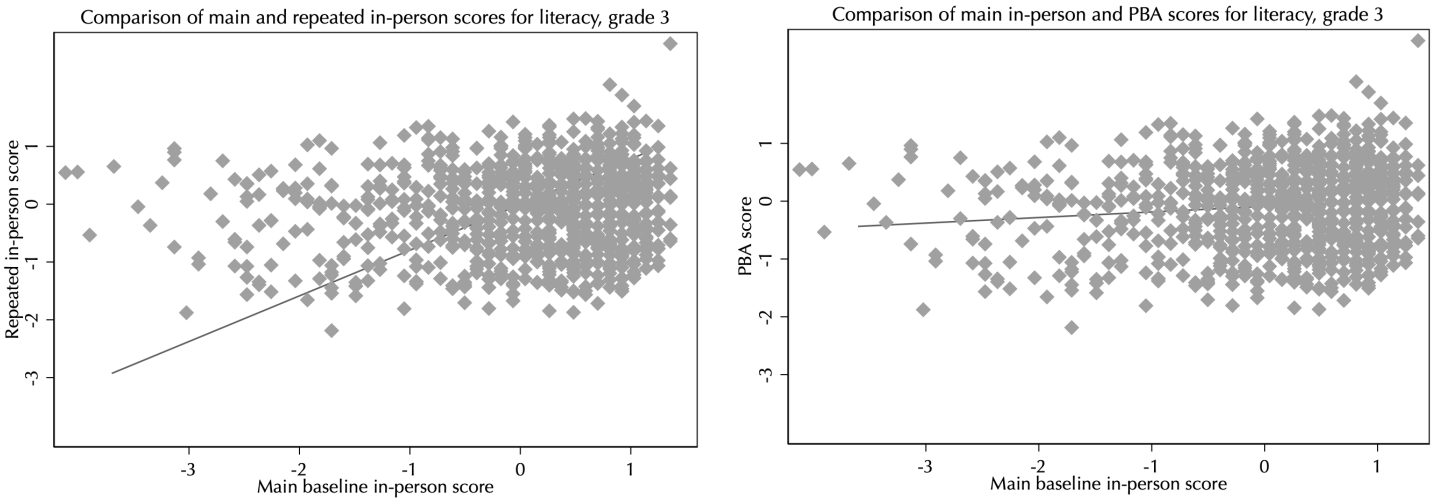


Notes: top panel subsets to students who do not report any COVID-19 related shock (health, income, or housing). Bottom panel subsets to students who report at least one of these shocks.

Appendix Figure 3a: scatter plots comparing correlations between main in-person and repeated in-person assessments (left column), and between main in-person and PBA (right column), for numeracy, by grade



Appendix Figure 3a: scatter plots comparing correlations between main in-person and repeated in-person assessments (left column), and between main in-person and PBA (right column), for literacy, for grade 3



b. Survey instruments for phone-based assessments

Numeracy – Grades 3, 5, and 6

Section	Item number	Questions only for Grade 3	Questions only for Grade 5 and 6
Instructions		Assessor's name: _____ *[Anything written in this font is a note to assessor that should not be read aloud]* **Step 1: Introduction.** Hello. My name is _____ and I am calling on behalf of Bridge International Academies. I am hoping to speak first with your pupil __<insert name>___ about some math problems and then at the end to speak with you again to get a bit of information about how the term has been going for your family. Does that sound okay? **Step 2: Instructions.** First, I would like your child to work on a couple of maths problems. I ask that you put the phone on speaker or repeat out the questions to the pupil to answer. Please have your child answer the problems on their own on a scrap paper. After they are done with a problem, you or they can read out their answer to me. The answers will not count toward grades in school, so it's okay if your child does not get all of the answers correct. Is your pupil ready? **Step 3: Assessment.** ***Core Numeracy Questions*** *[Please ask students the following questions in order. If they get three questions in a row wrong, please do not ask any more of the "core numeracy questions" and move on to the "grade level questions" section below.]*	
Learning assessment – Core numeracy	1	Can you count from 20-30? *[Mark the highest number the student reached consecutively, so if they get to 27 but skipped 25, mark (e) 24]* o No answer o 23 o 27 o 20 o 24 o 28 o 21 o 25 o 29 o 22 o 26 o 30	
	2	Which is greater? 64 or 38? o Correct (64) o Incorrect o No answer	
	3	What is 62+18? o Correct (80) o Incorrect o No answer	
	4	What is 33+49? o Correct (82) o Incorrect o No answer	
	5	What is 43-20? o Correct (23) o Incorrect o No answer	
	6	What is 81-43? o Correct (38) o Incorrect o No answer	
	7	What is 3x4? o Correct (12) o Incorrect o No answer	
	8	What is the result of 8 divided by 2? o Correct (4) o Incorrect o No answer	
	9	Oil is 200 shillings per liter and rice is 100 shillings a kilogram. How much should I pay for 3 liters of oil and 4 kilograms of rice? o Correct (1000 shillings)	

		○ Incorrect ○ No answer	
Instructions		***Grade Level Questions*** *[Please ask students the following questions in order. If they get three questions in a row wrong, please do not ask any more of the "grade level questions" and move on to the survey question.]*	
Learning assessment – Curriculum-aligned items	10	Complete the following number pattern: 13, 19, ___, 31 ○ Correct (25) ○ Incorrect ○ No answer	What is 3487+2325? ○ Correct (5812) ○ Incorrect ○ No answer
	11	What is 145+213? ○ Correct (358) ○ Incorrect ○ No answer	What is 4756-2149? ○ Correct (2607) ○ Incorrect ○ No answer
	12	What is 278-124? ○ Correct (154) ○ Incorrect ○ No answer	What is 42x26? ○ Correct (1092) ○ Incorrect ○ No answer
	13	What is 8x5? ○ Correct (40) ○ Incorrect ○ No answer	What is the result of 96 divided by 12? ○ Correct (8) ○ Incorrect ○ No answer
	14	What is the result of 35 divided by 7? ○ Correct (5) ○ Incorrect ○ No answer	What is the result of 3/7 (three sevenths) + 1/4 (one fourth) ○ Correct (19/28 (nineteenth twenty-eighths) ○ Incorrect ○ No answer
Instructions		**Step 4: Student survey.** Nice work. Next I am going to ask you a general question about school. There is no right or wrong answer, please just give your best response.	
Survey – Students	15	How much do you feel your teacher cares about your learning during the remote learning period? *[Read out each answer choice]* ○ Not at all ○ A little bit ○ Some ○ Quite a bit ○ A lot	
Instructions		*[To child]:* Thank you very much. Now, I would like to ask your parent a few questions, could you put them back on? **Step 5: Parent survey.** *[To parent]:* Thank you. Now I would like to ask you a bit about your child and household during this period of school shutdowns. Your participation is totally voluntary and you are welcome to skip any questions that you do not feel comfortable answering.	
Survey – Parents	16	On average over the past week, how many hours a day has your child spent on education? *[This is in reference to the child who completed the test]*	
	17	How many times has your child's teacher called you or your child by phone in the past 7 weeks? ○ 0 ○ 4 ○ 8 ○ 1 ○ 5 ○ 9 ○ 2 ○ 6 ○ 10 or more ○ 3 ○ 7 ○	
	18	What are children in your household currently doing to learn?" *[Read out each option and mark all that apply]* ○ Educational TV programs or radio ○ Bridge@home ○ Receiving calls from child's teacher/academy manager/academy ○ Educational content on the internet ○ Books we have in the household ○ Government educational content - courses, audiobooks, or lessons ○ I/Others in my household are teaching or reading with them	

		○ I/Others encourage children to do distance learning (radio, television, phone, etc.) but do not help ourselves ○ We are paying for in-person tutoring ○ Other [*Please specify*]:_____ ○ Nothing
	19	What is the highest level of school that you or someone in your household has completed? ○ Some primary school ○ Primary school completion ○ Some secondary school ○ Secondary school completion ○ Certificate or other post-secondary ○ Some university ○ University completion ○ Post-graduate degree
	20	Finally, this has been a hard time for many families due to the coronavirus pandemic. In order to understand how this disruption has influenced children's learning outcomes, it would be helpful to know whether you have experienced any of the following since March when schools were closed: *[Read out options, pausing after each option for a yes/no, and mark all that apply]* ○ Moved to a different home ○ Had someone in your home experience health challenges ○ Had changes to your job or income
Instructions		**Step 6: Closing.** Many thanks for your help with this.

Literacy – Grade 3

Section	Item number	Questions for Grade 3
Instructions		Assessor's name: _____ *[Anything written in this font is a note to assessor that should not be read aloud]* **Step 1: Introduction.** Hello. My name is _____ and I am calling on behalf of Bridge Kenya. I am hoping to speak first with your pupil __*<insert name>*___ about some English activities. After that, I would like to speak with you again to get a bit more information about how the remote learning period went for your family. Does that sound okay? **Step 2: Instructions.** First, I would like your child to work on a couple of English activities. Please put the phone on speaker. Make sure that your child completes the activities on their own. Their answers will not count toward grades in school, so it's okay if your child does not get all of the answers correct. Is your pupil ready? **Step 3: Assessment.**
Reading fluency	1	Open your term 2 homework book to page 186 and point to the title 'Selfish Hare' at the top of the page. Are you pointing to the title 'Selfish Hare'? *Pause.* This is a short passage. I want you to read it aloud, quickly but carefully. When I say "Start," read the passage as best as you can. If you come to a word you do not know, go on to the next word. Put your finger on the first word. Ready? Start. *Start timer when pupil says the first word and follow along on the rubric as the pupil reads. Make a tick mark every time that the pupil skips a word or reads a word incorrectly.* *After 60 seconds, say STOP. Make a note of the last word that the pupil read. Thank the pupil. Fill in the results for Words Correct and Words Attempted.* ● *Words Attempted: the total number of words that the pupil attempted to read (including skipped and incorrect words)* ● *Words Correct: [Words Attempted] minus [total skipped and incorrect words]*
	2	Open your term 2 homework book to page 190 and point to the title 'Sophie and Mercy' at the top of the page. Are you pointing to the title 'Sophie and Mercy'? *Pause.* This is a short passage. I want you to read it aloud, quickly but carefully. When I say "Start," read the passage as best as you can. If you come to a word you do not know, go on to the next word. Put your finger on the first word. Ready? Start. *Start timer when pupil says the first word and follow along on the rubric as the pupil reads. Make a tick mark every time that the pupil skips a word or reads a word incorrectly.* *After 60 seconds, say STOP. Make a note of the last word that the pupil read. Thank the pupil. Fill in the results for Words Correct and Words Attempted.*
Vocabulary	3	Now let's try a word game. Imagine you are going to the market. Name some foods that can be bought from the market. Try to name as many things as you can think of and I will keep count. *[Keep time and end after 1 minute]* o Number of correct answers within 1 minute: _____

Oral comprehension	4	*[Please ask students the following questions in order]* Now I am going to tell you an interesting story. After I have told you the story I will ask you some questions. Listen carefully, okay? *[Read out the story slowly, clearly and fluently.]* *The Football Match* The school sports day was a long awaited event. All the pupils were excited for the day. During this day, all the pupils wore their track suits or shorts. In addition, there were no classes on this day. All they did during the sports day was play! Nuru had just learnt how to play football. She enjoyed the game especially with her friends and classmates. When the teams were formed, Nuru was lucky to make the school team. She wanted to play for her school team for a long time. Nuru was getting better because she practiced every day. She could kick the ball, throw and even score a goal! The first match was between Nuru's team and the neighboring school. She was excited to participate in the match. She wanted to score a goal for her school team. Nuru played hard with her teammates. She scored a goal! 'Goal!' she shouted. All the other pupils were proud of her.
	5	Now I am going to ask you some questions about the story. What did the pupils wear to the sports day? o Correct (Track suits / Shorts / Track suits and shorts / Shorts and track suits) o Incorrect o No answer
	6	Why was Nuru getting better at football? o Correct (Because she practiced every day) o Incorrect o No answer
	7	What football skills did Nuru have? o Correct (Accept one or more of the following: kick the ball / throw the ball / score a goal) o Incorrect o No answer
	8	Why were the other pupils proud of Nuru? o Correct (Because she scored a goal) o Incorrect o No answer
Spelling	9	*[Please ask students the following questions in order.]* Now, I will ask you to spell some words for me. Please spell the word "sick" o Correct (S I C K) o Incorrect o No answer
	10	Please spell the word "when" o Correct (W H E N) o Incorrect o No answer
	11	Please spell the word "fight" o Correct (F I G H T) o Incorrect o No answer
	12	Please spell the word "children" o Correct (C H I L D R E N) o Incorrect o No answer
Instructions		Nice work. Next I am going to ask you a general question, asking you to think back to the remote learning period. There is no right or wrong answer, please just give your best response.
Survey – Students	13	About how often did your parent (or another adult at home besides your teacher) help you with studying while school was closed for remote learning? *[Read out each answer choice]* o Almost never o Once or twice per month o Once per week o Twice per week o Almost every day

Instructions		*[To child]:* Thank you very much. Now, I would like to ask your parent a few questions, could you put them back on? **Step 5: Parent Survey.** *[To parent]:* Thank you. Now I would like to ask you a bit about your child and household. Your participation is totally voluntary and you are welcome to skip any questions that you do not feel comfortable answering.
Survey – Parents	14	Please think back to the period of school shutdowns. How adequate was the learning support your child was getting from his/her school during the remote learning period? *[Read out each answer choice]* o Not at all adequate o Not very adequate o Somewhat adequate o Quite adequate o Very adequate
	15	On average, in a typical week during the remote learning period, how many hours a week--if any--did you spend helping your child with learning? o Type a number of hours here:_____
	16	How confident did you feel that your child was making academic progress during the remote learning period? *[Read out each answer choice]* o Not at all confident o Not very confident o Somewhat confident o Quite confident o Very confident
Instructions		**Step 6: Closing.** Thank you for participating in this survey. Have a nice day.

Reading Fluency Passage 1: 'Selfish Hare'

Once upon a time, in Tembo Forest, there lived some animals. There was Lion, Cheetah, Elephant, Zebra, Rhino, Monkey, Hyena and Hare. They always helped each other. They looked for food together. They also drank water at the river together. Then they would take
baths and splash water on each other for fun. They were one big happy family.

Hare was however very selfish. He did not like to share his food with the others. One dry season, there was no food in Tembo Forest. Oh how the hungry the animals were! "What shall we do?" Monkey asked. He was feeling so weak and hungry. Lion said quietly, "If we do not get food soon, we shall die."

All the other animals shook their heads in sorrow. Where would they get food? One day, Hare was walking to the end of the forest. He lifted his head and saw a field full of all kinds of food! Hare was so happy. He rushed into the field. He ate so fast that he almost vomited!

"I cannot eat anymore." Hare said as he belched. "I will not tell my friends about this field." He continued. "They might finish all the food!" He ran back home and brought a basket. He filled the basket with fruits and meat. Hare walked back home with the basket. He did not want Hyena, his neighbor to see him. He tiptoed quietly to his doorstep.

Just as he was about to get into his house, Hyena got out of his door! "Oh no! Hyena saw me!" Hare thought to himself. Hyena could not believe his eyes! So much food! Where had it come from? He tried asking Hare, but Hare said nothing. Hyena immediately went looking for the other animals. He found them at Elephant's house.

They were very hungry. "I just saw Hare with a lot of food! And he refused to share!" Hyena told them. All the other animals got very angry. "Hare is very selfish." Monkey said. "I agree!" Cheetah replied. They all ran to Hare's house. They banged his door, but Hare did not answer.

Monkey decided to trick Hare. "Hare", Monkey said, "We have found some meat in the bush!" Immediately, Hare opened his door to see the meat. The other animals caught him and forced him to say where he found the food. After hours of waiting, Hare finally agreed to show them where the field of food was. They then sat Hare down and reminded him the importance of sharing. "When you share what you have with others, it means you love them." Zebra said. "It also shows that you are not selfish", Rhino added.

Hare felt very sorry for what he had done. “Forgive me please.” Hare told the other animals. “We forgive you Hare, but don’t be selfish again.” They all replied. And they lived happily ever after.

Reading Fluency Passage 2: ‘Sophie and Mercy’

Sophie and Mercy had done a good job helping Mother clean the house. They also collected the dry leaves in the compound. Mother was very happy with them that day. She sent them to the market to buy some supper. She gave them a two hundred shilling note. As a gift, Mother told them to buy themselves some fruits using the change that would remain after shopping. She also asked them to hurry and be back before it became dark.

They put on their dresses and left home in the afternoon. The field next to the market had a very big ceremony. The people of Suma Village were having an Agricultural Show. There were very many people. The farmers had brought their cows, goats, sheep and chicken to the show. There were farmers who also brought maize, beans, bananas and mangoes to the show. There was also a lot of music and some dancers were entertaining people.

Sophie and Mercy decided to go to the show for a few minutes. They got in and stayed together. They visited various stands to see what the farmers had brought. They then went to see the dancers. They were so happy that they forgot to go to the market! By the time they realized, it was already getting dark and cold. Sophie and Mercy rushed out of the show and into the market.

Sophie put her hand inside her pocket to look for the money. It was not there! They must have lost it at the show! They tried to look through the market. They hoped to find one of Mother’s friends. She would help them get some food without money. But all of them had left by that time. Sophie and Mercy ran back home. They knew Mother would be very angry with them. They had lost the money she gave them. They had gone home with no supper. They also had gone home very late.

They got home and found Mother very worried. They explained to her what had happened. Mother became angry. She however forgave them and reminded them to always be obedient. Since there was no food to eat, Mother made tea, bread and eggs. They ate it for supper and then went to bed.