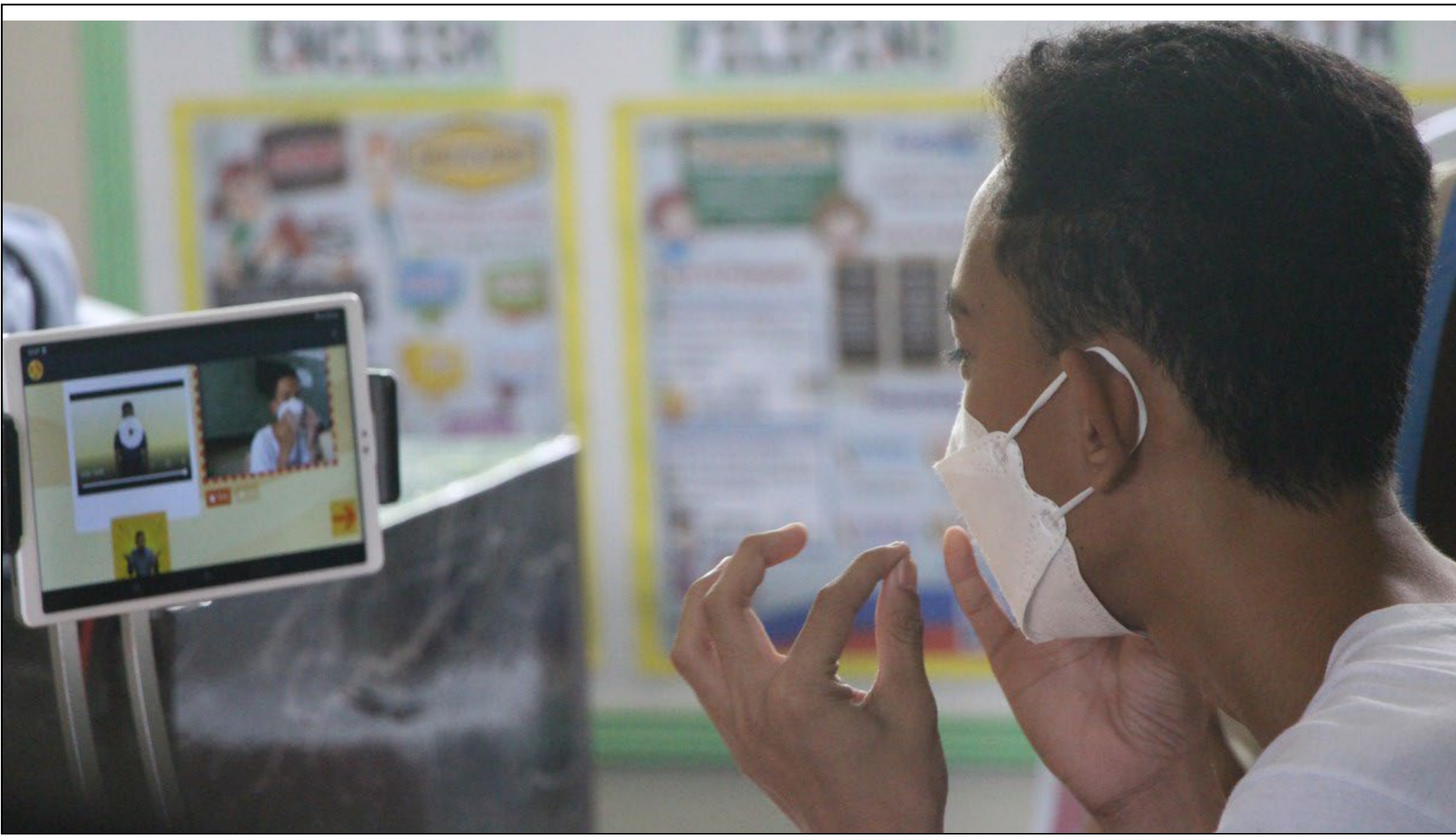
REMOTE EGRA FOR LEARNERS



WHO ARE DEAF OR HARD-OF-HEARING

FINAL REPORT

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

ACR-Asia	All Children Reading Asia
ASL	American Sign Language
COVID-19	coronavirus disease 2019
DepEd	Philippines Department of Education
EGRA	Early Grade Reading Assessment
FGD	focus group discussion
FSL	Filipino Sign Language
INGO	international non-governmental organization
IRR	interrater reliability
IT	information technology
KII	key informant interview
RBI	Resources for the Blind, Inc.
RTI	RTI International
STS	School-to-School International
TAAPA	total assessment average percent aggregate
USAID	United States Agency for International Development

EXECUTIVE SUMMARY

BACKGROUND AND PURPOSE

Early Grade Reading Assessments (EGRAs) measure students' progress in reading through individual administration of an oral survey of foundational reading skills. Administration is generally conducted on-site by teams of trainer assessors, face to face with students in a one-on-one capacity. While EGRAs are administered internationally, students who are deaf or hard of hearing are often left at a disadvantage by prevailing reading assessments.

To adapt EGRAs to fit the needs of students who are deaf or hard of hearing, USAID has supported the development of EGRAs specifically for students who are deaf or hard of hearing in Kenya, Morocco, Nepal, and the Philippines, among other countries. In the Philippines, these assessments have improved the understanding of and capability in inclusive education programming, including the development and pilot implementation of the Filipino Sign Language (FSL) curriculum and training and mentoring of teachers in FSL.

By design, these EGRAs adapted for students who are deaf or hard of hearing are administered in person and require both an assessor and an enumerator. The assessor sits in front of the student with a device or paper stimulus that displays images, letters, words, or sentences used in the assessment. The enumerator uses another device to record the student's responses. The administration of EGRAs is also time-sensitive, often scheduled at the end of a set of interventions, usually at or near the end of a school year.

However, there are a growing number of challenges that impact the ability to conduct EGRAs for students who are deaf or hard of hearing. The coronavirus disease 2019 (COVID-19) pandemic and ensuing restrictions against in-person contact with students have prevented further administration of the EGRA in its original design. Other adverse weather and geological situations like typhoons, flooding, volcanic eruptions, and earthquakes, often affect the ability to conduct activities on-site and in person. Changes and differences in the school calendar also affect the results and, potentially, the validity of assessments.

As there is no information on existing models of remotely administered EGRAs, the purpose of this activity was to prototype—design, develop, and test for proof of concept and acceptability—an early grade reading assessment that is administered asynchronously with assessors and enumerators who are not on-site, for students who are deaf or hard of hearing. Such a model can be deployed in outbreaks and emergencies that affect the ability to administer EGRAs in person and at a specified period and specifically adapted for students who are deaf or hard of hearing.

RESEARCH QUESTIONS

Four research questions guided this activity:

1. Which subtasks from existing EGRAs for students who are deaf or hard of hearing (like the USAID Gabay EGRA) allow for asynchronous administration? Which, if any, subtasks that are not part of the existing EGRAs could be considered?
2. What type of asynchronous administration is operationally feasible, technically rigorous, and suited to the context of Deaf education in the Philippines?
3. What are appropriate protocols for asynchronously administered subtasks? How do these diverge from protocols of the in-person administration of these subtasks? Protocols should consider preferred media platforms, suitable locations, length of testing, assess-ee identity, and data privacy, among other things.
4. Which factors are the most determinant drivers of the cost? Which factors impact the efficiency and effectiveness of asynchronous administration? Is the design scalable within the Philippines beyond the proof of-concept?

SUMMARY CONCLUSIONS AND RECOMMENDATIONS

In summary, the learnings from the pre-test, alpha test, and beta test of the asynchronous administrated EGRA can be summarized in the following points:

1. Non-FSL-fluent proctors are effective and scalable.
2. Stronger protocols for scoring expressive tasks are needed—both in definition of scorable responses and in the process of how scorers review responses.
3. Receptive tasks can reduce the scoring challenges.
4. The length of the assessment between receptive and expressive subtasks is on average equivalent, but as the FSL level of the learner increases, the time of the assessment decreases.
5. Assessment delivery through tablets and Tangerine:Learn is user friendly and scalable, but students could use additional exposure to technology.

BACKGROUND AND PURPOSE

EGRAs measure students' progress in reading through the individual administration of an oral survey of foundational reading skills. Administration is generally conducted on-site by teams of trained assessors face to face with students in a one-on-one capacity. Because progress is often gauged against intervention activities, international nongovernmental organizations (INGOs) often administer EGRAs at or near the end of an academic year.

However, adverse weather and geological situations—like typhoons, flooding, volcanic eruptions, and earthquakes—have long affected INGOs' ability to conduct assessments on-site and in person. Changes and differences in the school calendar also affect results and, potentially, the validity of benchmarking assessments like the EGRA. These challenges to timely, in-person diagnostics have been exacerbated—and made truly global—by the COVID-19 pandemic and ensuing restrictions against in-person contact with students.

The need exists for models of remotely or asynchronously administered EGRAs.[1] Such models could be deployed in health outbreaks and other emergencies that affect the ability to train assessors and administer an EGRA as designed. However, little to no information is currently available on such models. Similarly, little is known about adaptations of EGRAs that are inclusive of students who are deaf or hard of hearing.

The purpose of this activity is to prototype—that is, to design, develop and test for proof of concept and acceptability—an early grade reading assessment for early students who are deaf and hard of hearing to be administered asynchronously with assessors and enumerators who are not on-site with the students being assessed.

Under funding from the USAID All Children Reading Asia (ACR-Asia) task order, and in collaboration with and under guidance provided by RTI International (RTI), School-to-School International (STS) has been the principal technical assistance and implementing partner to design a “proof of concept” research activity to pilot potential ways of asynchronously conducting EGRAs for students who are deaf or hard of hearing in the Philippines. RTI and STS have built on best practices and lessons learned from previous EGRAs for students who are deaf or hard of hearing. The design has been done in consultation and collaboration with the USAID Gabay project implemented by Resources for the Blind, Inc. (RBI), the Philippines Department of Education (DepEd), and the USAID/Philippines Mission. In consultation with RTI, STS has coordinated with RBI on in-country logistics to coordinate key stakeholders on the ground, obtain necessary government approvals, and implement field testing.

The proof of concept research activity took place in three phases: a pre-test, alpha test, and beta test.

In May 2022, RTI, STS, and RBI conducted a pre-test with students to focus on user experience of the adapted Tangerine: Learn application. The students tested Tangerine:Learn’s video capture functionality. Student responses to user experience questions, in addition to observations from RBI and USAID, provided critical feedback to the application’s functionality and informed the research design.

In May and June 2022, RTI, STS, and RBI conducted the alpha testing with 28 primary-grade students in three schools in the Metro Manila and Visayas regions of the Philippines. The alpha test explored three scenarios for administering an assessment via Tangerine:Learn.[2] RTI, STS, and RBI developed the three scenarios and their various aspects based on a landscape review and consultative meetings. The primary goal was to experiment with the scenarios in a controlled environment. By controlling the environment, RTI, STS, and RBI identified and corrected major problems in the three scenarios before testing the remote EGRA in situations that will likely mirror real-life implementations.

[1] Asynchronous administration refers to an as assessment that does not take place with a student in real-time and, instead, can be conducted virtually or through other modes.

[2] Tangerine:Learn is part of RTI’s open-source Tangerine® software, with student-facing interface developed specifically for young learners and capacity to show and capture video in asynchronous learning and assessment scenarios.

In September 2022, RTI, STS, and RBI conducted the beta test in National Capital Region, Calabarzon, and Central Visayas regions. The asynchronous EGRA was administered to 177 students across 18 schools. The beta test explored the assessment parameters that support or inhibit the remote EGRA's scalability for students who are deaf or hard of hearing. Building on findings from the alpha test, the beta test focused on in-person proctor support and tested two assessment formats. Improvements were also made to Tangerine:Learn, FSL instruction videos, proctor training, and scoring protocols. To vary the contexts of the testing environment, both rural and urban schools were included in the sample.

This report summarizes the methodology, findings, and recommendations for future explorations of a remote or asynchronously administered EGRA for students who are deaf or hard of hearing.

RESEARCH QUESTIONS

To guide this project, RTI and STS established the following research questions:

1. Which subtasks from existing EGRAs for students who are deaf or hard of hearing (like the USAID Gabay EGRA) allow for asynchronous administration? Which, if any, subtasks that are not part of the existing EGRAs could be considered?
2. What type of asynchronous administration is operationally feasible, technically rigorous, and suited to the context of Deaf education in the Philippines?
3. What are appropriate protocols for asynchronously administered subtasks? How do these diverge from protocols of the in-person administration of these subtasks? Protocols should consider preferred media platforms, suitable locations, length of testing, assess-ee identity, and data privacy, among other things.
4. Which factors are the most determinant drivers of the cost? Which factors impact the efficiency and effectiveness of asynchronous administration? Is the design scalable within the Philippines beyond the proof of concept?

DESIGN AND METHODOLOGY

CONSULTATIVE PROCESS

In December 2021 and January 2022, STS consulted with stakeholders in the Philippines and the global Deaf community to provide insights into the proof of concept design and implementation of a remote or asynchronous assessment. STS utilized a snowball approach, in which consulted individuals were asked to recommend additional persons who could contribute knowledge or recommendations.

The consultations included individuals and representatives of groups that have expertise in (1) Deaf education in the Philippines, (2) administration of assessments for student who are deaf or hard of hearing, and (3) administration of remote or asynchronous learning assessments. These consultations were coordinated with input and participation from RTI, USAID/Philippines, USAID/Washington, and RBI, as appropriate.

STS engaged individuals through focus group discussions (FGDs) or key informant interviews (KIIs) to explore the following key questions, as applicable:

1. What assessments and assessment modalities (technologies) are being used for students who are deaf or hard of hearing globally? In the Philippines?
2. What is the lived experience for children who are deaf in the Philippines? What technologies are used in the Philippines by people who are deaf, especially young people?
3. What are the appropriate terms that should be used within the Filipino context when referring to students who are deaf or hard of hearing, different technologies, educational adaptations, and assessments, etc.?
4. What different types of skills should be measured to assess foundational reading skills of learners who are deaf or hard of hearing?
5. What are potential challenges to administering reading assessment remotely in the Philippines?
6. What are technological limitations to remote administration of a reading assessment in the Philippines?
7. What sorts of considerations for protocols (i.e., assessment rules) should we be aware of in the Philippines? These may include: technology exposure and access by the learners, connectivity, location of assessment, presence of others during assessment, etc.
8. What types of remote administration of a reading assessment are feasible and appropriate for learners who are deaf or hard of hearing in the Philippines? Which types may be acceptable by learners who are deaf or hard of hearing based on their lived experience?
9. Are there other projects or initiatives in the Philippines that may influence or impact the testing of these assessments?

FGDs were conducted with three groups within the Philippines—one group of government officials and members of the USAID/Philippines Mission, one group of academics and implementers focused on education for students with disabilities, and one group of teachers from the USAID Gabay project. In total, STS consulted 15 participants through these FGDs.

STS conducted five KIIs with USAID/Washington, Deaf education experts, and remote learning assessment experts.

Throughout the consultative process, STS interviewed eight participants who are deaf or hard of hearing.

LANDSCAPE REVIEW

The consultative process, together with a literature review, comprised a landscape review of existing assessments, technologies, potential challenges, and the lived experience for students who are deaf or hard of hearing—both globally and within the Philippines. Components of the consultative process and the literature review complemented and informed each other; together, they served as the basis for developing this proof of concept research design. The review found 12 separate assessments for lower-level skills in American Sign Language (ASL) in the United States, and noted challenges in developing tests in ASL, which include the need for highly trained examiners and prohibitive costs of purchasing standardized assessments.

Based on learnings from the consultative process and landscape review, RTI, STS, and RBI, in consultation with USAID/Washington and USAID/Philippines, decided on the following aspects of the asynchronous administration (**Table 1**).

Table 1. Aspects of the Asynchronous EGRA to Be Tested

Aspects	Decision Points
Synchrony	**Asynchronous administration**
	Rationale: Synchronous assessments were dismissed as too challenging due to the need for a reliable and consistent Internet connection. In this asynchronous administration, the student will receive instructions through pre-recorded videos in FSL. Student responses will be video captured and will also be asynchronously scored after the student has completed the assessment.
Technology	**Tablet**
	Rationale: Tablets are the most universally available device that still accommodates a screen large enough for viewing signs.
Software	**Tangerine for assessment administration** **Zoom (or another cloud-based video communications app) for remote FSL support**
	Rationale: Tangerine is an open-source software with a student-facing interface and capacity to show and capture video in asynchronous scenarios. Cloud-based applications, like Zoom, were identified in the landscape review as the most accessible form of video communications.
Student qualifications	**Enrolled in formal education (special education or mainstream class) in target grade**[3] **Basic FSL**[4]
	Rationale: Students' qualifications will be determined based on availability and feasibility at sampled schools. The qualifications will not vary between scenarios.
Testing site	**School**
	Rationale: Schools will allow for a more controlled environment to identify challenges during the alpha test.

[3] Depending on final determination of the grade(s) to be included in the prototype testing, schools and teachers will provide a list of eligible students.

[4] The determination of the grade(s) to be included in the prototype testing will consider that studengartents should have basic FSL skills. For example, it may not be appropriate to test with kinder or Grade 1 students due to their FSL skills. The prototype testing will include FSL subtasks to better understand students' language skills in addition to their reading skills. Teachers at assessment sites will be asked to recommend students for testing who have FSL comprehension skills and will be able to understand the assessment instructions. These may be students from Grade 2 through Grade 5.

Aspects	Decision Points
Instructions for EGRA	**Video on tablet**
	Rationale: Students may not have sufficient reading comprehension, therefore instructions will be provided through pre-recorded videos is FSL, which will allow the assessment to be administered asynchronously.
(In-person) Proctor during EGRA	**Teacher**
	Rationale: Teachers are familiar with students and can help keep them on task. Teachers selected to proctor will receive training on their role and responsibilities, including how to remain neutral.
Scorer qualifications	**Fluent in FSL** **Previous experience with EGRA and FSL assessments** **Trained in the model's administration**
	Rationale: These qualifications are similar to those of the USAID Gabay EGRA scorers. Qualifications and scorers should be constant across scenarios.
Scoring[5]	**Offsite** **Following individual completion of the assessment**
	Rationale: Because the assessment will be administered asynchronously in the three scenarios, data will be uploaded after the student has completed the assessment. Data will be available through Tangerine's secure data storage platform.
Subtasks	**Standard list**
	Rationale: Please see Table 2 *for details.*

INSTRUMENTS

The EGRA tool that will be used for the proof of concept was developed by USAID Gabay, with support from STS. For the prototype testing, STS modified the EGRA for asynchronous administration and developed additional tools in conjunction with project stakeholders, including RTI, USAID/Philippines, RBI, and the Deaf community.

Adaptation of USAID Gabay EGRA for Students Who Are Deaf or Hard of Hearing

STS, in consultation with RTI and RBI, reviewed each subtask from the USAID Gabay EGRA and updated administration protocols to account for asynchronous administration (**Table 2**). Assessment length was an important consideration in the adaptation discussions. To reduce length of the assessment and potential for assessment fatigue, RTI and STS adapted six out of eight subtasks from the USAID Gabay EGRA. The two subtasks not adapted were Sign Language Comprehension (Level 2) and Fingerspelling. The number of items in each subtask was reduced by half. Sign Language Comprehension (level 1) and Sentence Reading Comprehension remained at five items.

STS updated instructions to reflect these modifications to administration protocols. RBI provided interpretation instructions translated into FSL.

[5] Although the main purpose of the prototype testing is not to measure student performance, a scoring committee will be engaged to review student responses for correctness. Student scores will be assessed to provide feedback on the feasibility of scoring response videos taken by tablet, not to report on student performance.

Table 2. USAID Gabay EGRA Adapted for Asynchronous Administration

Subtask	Description	Number of Items	Modified Protocol for Asynchronous Administration
Receptive Vocabulary	Measures students' receptive comprehension of common vocabulary words	10	• Instructions delivered through videos • Demonstration video of student doing the subtask • Additional practice item (beta only) • Items delivered through video—an assessor signs the word twice • Student selects response on tablet
Expressive Vocabulary	Measures students' ability to produce the sign for common vocabulary words	10	• Instruction delivered through videos • Demonstration video of student doing the subtask • Additional practice item (beta only) • Image individually shown in application • Student signs response into camera
Sign Language Comprehension (level 1)	Measures students' ability to understand FSL grammar and comprehend sentences	1 story of 5 sentences; 1 comprehension question per sentence	• Instruction delivered through videos • Story and comprehension questions delivered through video • Student signs response into camera
Letter Name Identification	Measures students' written alphabet knowledge and knowledge of the correspondence between English letters and FSL	13	• Instruction delivered through videos • Demonstration video of student doing the subtask • Additional practice item (beta only) • Each item individually shown in application* • Student signs response into camera*
Familiar Word Reading	Measures students' word recognition and decoding skills and knowledge of the correspondence between common words and signs	7	• Instruction delivered through videos • Demonstration video of student doing the subtask • Additional practice item (beta only) • Each item individually shown in application* • Student signs response into camera*

Subtask	Description	Number of Items	Modified Protocol for Asynchronous Administration
Sentence Reading Comprehension	Measures students' ability to read and comprehend connected text	1 story of 5 sentences; 1 comprehension question per sentence	• Instruction delivered through videos • Story sentences individually shown in application • Comprehension questions delivered through video • Student signs response into camera

*Note: Letter Name Identification and Familiar Word Reading protocols varied between the two assessment forms for beta test. This will be discussed later in this report.

In collaboration with RTI and RBI, STS developed the observer checklist, student feedback survey, and proctor feedback survey to track assessment administration for each student and provide more general observations and feedback on the process. The tools contained a mix of closed and open-ended questions, including Likert scales measuring the level of agreement and the frequency of various behaviors demonstrated during the assessment. Additional data came from assessment scoring of response videos, scorer feedback, and debriefs with participants. Tools are attached in **Annex C, D, G and H** for further reference.

Observer Checklist

During the administration of each assessment, an observer would complete the observer checklist. The checklist captured observations on student ease using the tablet and software, proctors' adherence to roles and responsibilities, technological challenges, and other qualitative observations.

Student Feedback Survey

An FSL-English interpreter[6] facilitated the student feedback survey at the end of the assessment. The survey examined student perceptions of the assessment, their ability to understand the instructions, and whether they asked for help during the assessment.

Proctor Feedback Survey

The observer administered the proctor feedback survey after each assessment. The survey captured the frequency and extent of support the proctor provided to the student, the proctors' perceptions of student engagement with the assessment and technology, and the proctors' perception of the student's FSL skills.

Assessment Scoring and Scoring Feedback Form

Scorers reviewed student response videos during the scoring exercise. This exercise aimed to understand the feasibility of scoring an asynchronous assessment. Each video was reviewed to examine if the recordings were scorable. First, the scorers considered whether they could understand the learners' response. The scorers specifically examined whether there were any issues with the video file with regard to their ability to see the learner and decipher the student's response. No rubric was provided for this process. The criteria were simplified into one question

[6] The FSL interpreters were fluent in both their native language—either Tagalog or Cebuano—and English. FSL interpretation fluctuated between Tagalog or Cebuano and English.

"Is this video scorable?" to which the scorers either answered "yes" or "no." Scorers provided qualitative feedback on reasons why videos were not scorable. The scorers then reviewed the response, comparing the response to a provided answer guide. Scorers used the scoring feedback form to provide qualitative feedback on the overall feasibility of scoring assessments and any challenges encountered.

PRE-TEST

A key component of the asynchronous administration was video functionality in the assessment application. As this function was newly developed in Tangerine:Learn, RTI, STS, and RBI pre-tested the application to collect critical feedback on the design and development of the application, insights into the user experience of both thestudents and proctors, and critical data on efficacy. The pre-test was conducted with five students from a school in Metro Manila.

As the students navigated through the assessment on the tablet, RBI asked the students a list of questions, as provided in a user-testing protocol. These questions investigated how the students interacted with the technology; their navigation of the software interface; and the performance of the software to accurately present the testing items and transitions, and to accurately capture results.

From the pre-test, RTI, STS, and RBI collected data from student responses as well as observations from RBI and representatives from USAID/Philippines and RTI. Notable results from the pre-test are as follows:

- Increase in size of the response option buttons, as students struggled to see the images.
- Modification of "don't know" button, as students seemed to not understand the icon.
- Improvement of Tangerine:Learn's video recording, as observations reported that the resolution was low.
- Emphasis in proctor training to ensure student is captured on camera and can been seen while signing.
- Addition of specific guidance to proctor training on how to guide students during assessment.

ALPHA TEST

Building on learnings from the pre-test, the alpha test attempted three possible scenarios of asynchronous administration as identified from the landscape review and consultative process. The primary goal of the alpha test was to experiment with the scenarios in a controlled environment, such as a school or classroom context, and examine how proctor fluency in FSL and the presence of the online help desk influenced assessment feasibility. By controlling the environment, STS and RTI identified—and corrected—major problems before testing the scenarios in less-controlled environments, such as students' homes. The three scenarios are outlined in **Table 3**.

Table 3. Alpha Test Scenarios

Scenario Aspects	Scenario 1: Proctor	Scenario 2: Proctor + Remote FSL Support	Scenario 3: Proctor + On-Site FSL Support
(In-person) Proctor FSL fluency	**Non-fluent**	**Non-fluent**	**Fluent**
	Rationale: Teachers' levels of FSL fluency vary widely. By testing with FSL-fluent and non-fluent teachers, the team will be able to understand the appropriateness of the assessment in real-life contexts. Teacher FSL fluency will be evaluated by their FSL training, self-assessment of their level of FSL, number of years teaching in FSL, and other criteria, which will be determined prior to alpha testing.		
(Remote) FSL support or helpdesk	**Not present**	**Online real-time**	**Not present**
	Rationale: Iterating the presence of online language support can help the team understand the type of FSL support, in person or online, that is most appropriate for administration.		

The alpha test provided an opportunity to understand in detail whether in-person or online support was most appropriate for administration. Additionally, because teachers' levels of FSL fluency vary widely in the Philippines, it was essential to know if student outcomes differed based on the language ability of the person fielding questions during the assessment.

Sample

Alpha test assessments were conducted on May 31 and June 1, 2022, with 28 students in the Biliran, Antipolo, and Metro Manila divisions. The alpha test sample comprised 18 girls (64.3 percent) and 10 boys (35.7 percent). Student ages ranged from 10 years old to 21 years old, with an average age of 13 years. Most students were in Grade 2 (17.9 percent) or Grade 3 (39.3 percent). The remaining 42.9 percent of the sampled students were in Grades 4, 5, or 6. STS initially sought a sample of 30 learners for the alpha test, but an accident in Biliran precluded completing the last two assessments for that location.

RBI selected Naval Central Special Education Center, Bagong Nayon IV Central School, and Philippine School for the Deaf as alpha testing sites, as these schools were not included in the pilot or baseline administration of the USAID Gabay EGRA. It was necessary to minimize student exposure to the assessment and assessment items to protect the integrity of monitoring and evaluation plans for both this alpha test and the USAID Gabay project.

Procedure

In all scenarios, assessments were conducted with one student at a time in a dedicated classroom, allowing students to work at their own pace. Two proctors were present at each testing site. The two proctors alternated proctor duties throughout the day to avoid fatigue. During the assessment, the proctor sat with the student and was responsible for orienting them to the tablet and software; that proctor also provided support in navigating the tablet as needed. Additionally, two observers monitored each assessment and provided general observations of the assessment, the student's engagement, and proctor interactions with the student. Finally, an

FSL-English interpreter[7] was also present in each assessment to facilitate the student feedback survey following the assessment's conclusion and provide FSL-English interpretation as needed. The two observers and the interpreter sat at a distance from the learner and proctors to allow the learner to focus on the assessment.

In Scenario 2 only, online help desk support was present via Zoom video on a tablet located in the student's line of vision next to the assessment tablet. The individual providing online help desk support answered the student's questions as directed to the online help desk. They also provided qualitative feedback on the student's FSL fluency and comprehension following the assessment.

Student response videos—produced in all subtasks except Receptive Vocabulary—were reviewed and scored following the conclusion of all assessments. One scorer was assigned for each school, based on their fluency with the applicable regional FSL. Two of the three scorers were also scorers during the Gabay baseline. All three scorers were teachers of students who are deaf. All three scorers were highly proficient in FSL; although their FSL fluency was not formally assessed, they were recommended by schools and the Deaf community.

To review the response videos, scorers accessed the Tangerine:Learn web portal, logged into Tangerine, and accessed the uploaded results data with referenced links to all response videos for each student. Scorers marked the start and end time of review for each student record, and indicated whether each response was correct, incorrect, or not scorable using a scoring guide. After completing their review of all records, each scorer filled out a feedback form to provide comments on the process and any recommendations for future phases.

Training

Training for the alpha test took place on May 26, 2022, from 8:00–16:30 at RBI offices in Metro Manila. The training objective was to orient participants to the purpose of the alpha test, assessment administration, roles and responsibilities, and the Tangerine:Learn application. Sessions also discussed child safeguarding, research ethics, working with vulnerable populations, assessment set-up, and logistics. Time was also allocated for small group practice and role-play.

The training emphasized the roles and responsibilities unique to observers, proctors, sign language interpreters, and online help desk support personnel. It provided time for practice and discussion so that each participant understood their part in the alpha test. Participants practiced working through Tangerine:Learn to familiarize themselves with all subtasks and navigation.

Following the alpha test, STS and RBI conducted scoring training with three participants on June 3. Together, the scorers reviewed the purpose of the alpha test and oriented themselves to the assessment. They then reviewed the scoring guide, instructions, and feedback form before practicing scoring assessments.

BETA TEST

Following findings and recommendations from alpha test—which will be discussed in the following section—RTI and STS discontinued the testing of the alpha test scenarios. This

[7] The FSL interpreters were fluent in both their local native language—either Tagalog or Cebuano—as well as English. FSL interpretation fluctuated between Tagalog or Cebuano and English.

allowed for testing of the assessment with proctors with a wider range of FSL abilities, recognizing, in the Philippine context, teachers of students who are deaf or hard of hearing have varying language abilities. In eliminating testing of the scenarios, this led to further understanding whether this asynchronous modality could work, even with proctors with low levels of FSL, and elucidation of related challenges and opportunities. These future learnings could be applied further to similar contexts where there are also limited educators with local sign language proficiency.

Instead of varying the scenarios, the beta test varied the assessment forms to address concerns around assessment length. Form 1 featured the same assessment used during the alpha test, with five of the subtasks using expressive response formats—Expressive Vocabulary, Sign Language Comprehension, Letter Name Identification, Familiar Word Reading, and Sentence Reading Comprehension—and one subtask using receptive response format—Receptive Vocabulary. Form 2 maintained expressive response formats in the Expressive Vocabulary, Sign Language Comprehension, and Sentence Reading Comprehension subtasks. However, Letter Identification and Familiar Word Reading subtasks used receptive response formats in Form 2 rather than expressive response formats. **Table 4** provides a comparison of Form 1 and Form 2.

Table 4. Beta Test Assessment Forms

Subtask	Form 1	Form 2
Receptive Vocabulary	Receptive	Receptive
Expressive Vocabulary	Expressive	Expressive
Sign Language Comprehension	Expressive	Expressive
Letter Name Identification	**Expressive**	**Receptive**
Familiar Word Reading	**Expressive**	**Receptive**
Sentence Reading Comprehension	Expressive	Expressive

In their receptive formats, the asynchronous administration protocols for Letter Name Identification and Familiar Word Reading were conducted as follows: the instructions were delivered by video, a demonstration video of a student doing the subtask was shown, individual items were delivered through video where an assessor signs the letter or word twice, and the student responded to the item by selecting their answer from among four letters or words provided.

Testing receptive and expressive assessment forms allowed the project to investigate critical lines of inquiry:

1. Allowed for investigation into whether receptive subtasks could also reduce the assessment duration, as learners will not need to record their response through video.
2. Allowed for the investigation into whether receptive subtasks could reduce time spent on scoring with less items to score manually.
3. Allowed for the investigation on whether receptive subtasks might require less external support for the learner.

4. Allowed for comparison between the forms and the opportunity to examine the relationships between subtasks within each form.

Sample

Beta test assessments were conducted September 22–30, 2022, with 177 learners in 18 schools in the National Capital, Calabarzon, and Central Visayas regions.

RBI selected these 18 schools as these schools were not included in the pilot or baseline administration of the USAID Gabay EGRA. Again, it was necessary to minimize student exposure to the assessment and assessment items to protect the integrity of monitoring and evaluation plans for both the beta test and the USAID Gabay project.

The beta test sample comprised 50.8 percent girls and 49.2 percent boys. Student ages ranged from 7 to 29 years old. Most students were in Grade 4 (22.0 percent), Grade 6 (18.6 percent), or Grade 1 (18.1 percent). STS initially sought a sample of 180 students for the beta test, but absences and discrepancies in enrollment data impacted the sample size.

Further sample characteristics are provided in **Figures 1–5** below.

Figure 1. Beta Test Sample by Division

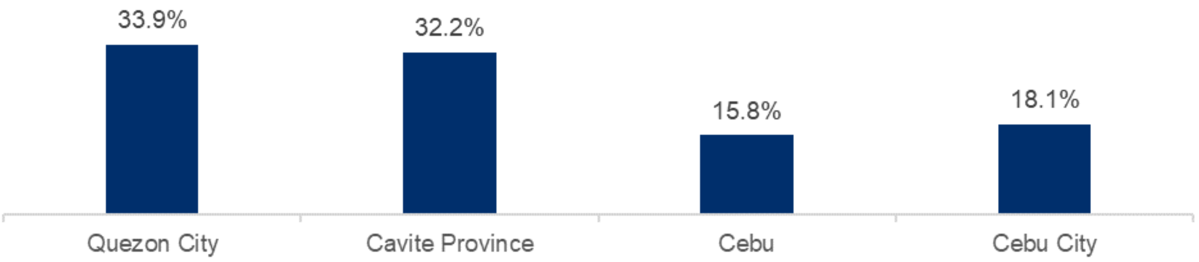


Figure 2. Beta Test Sample by Urbanicity

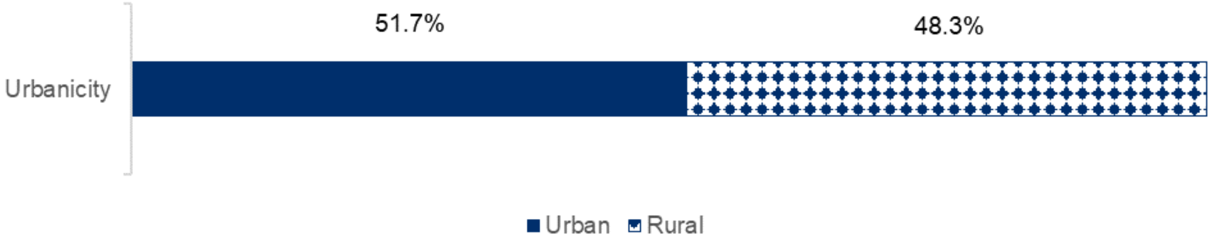


Figure 3. Beta Test Sample by Age

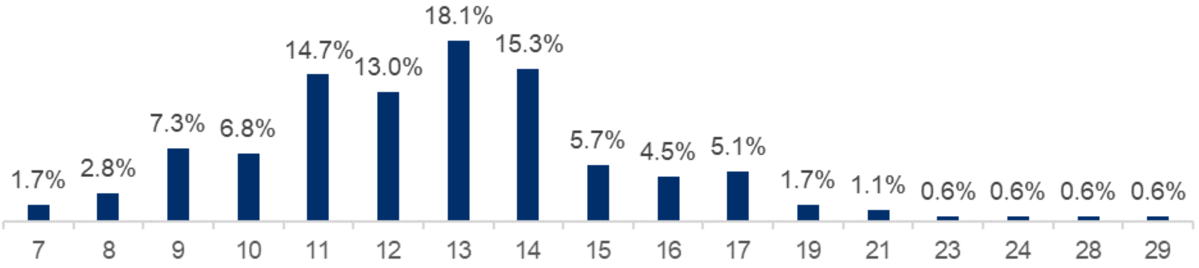


Figure 4. Beta Test Sample by Grade

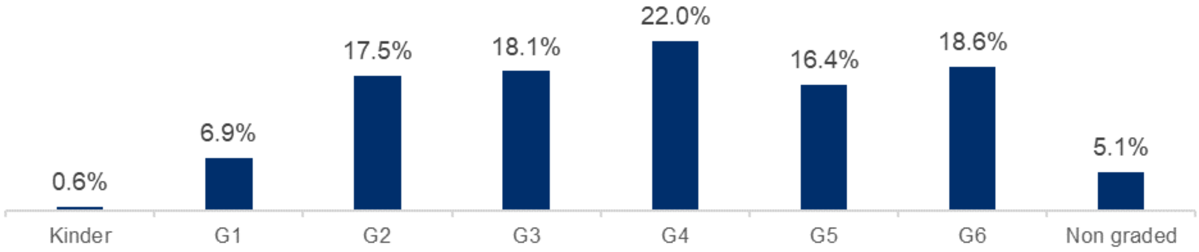
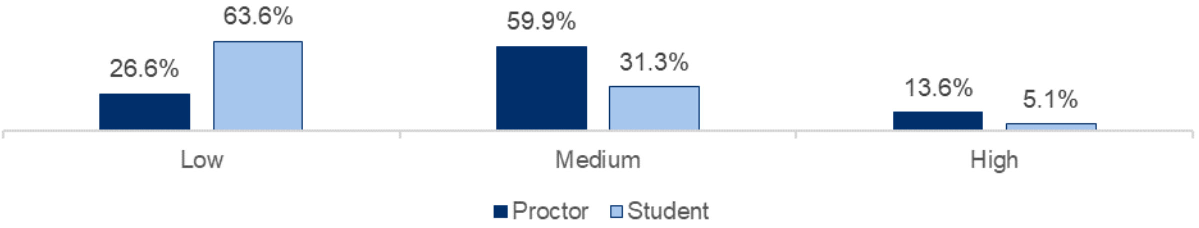


Figure 5. Beta Test Proctor and Student FSL Ability as Rated by Teacher*



*Note: Teachers provided the FSL abilities of their students as well as self-assessed their own FSL abilities. Because there is no commonly used standardized assessment of FSL in the Philippines, teachers provided these ratings based on their own understandings of FSL and their abilities to communicate in FSL.

Procedure

On the day of the school visit, participants—observers, interpreters, proctors, and other RBI staff—set up the testing areas. Two proctors were present at each testing site—one designated to proctor for students taking Form 1, and another designated to proctor Form 2. Each proctor had their own dedicated area, either a separate classroom or enough space between testing areas as to not cause disruptions. The participants ensured the testing site met the following requirements:

- A desk could be used by the student and proctor; the height of the desk was appropriate and allowed the child to comfortably sit.
- The tablet was at eye-level height; adjustments for taller students were made by placing a book or box under the tablet stand.
- The lighting of the room or testing area was sufficient for the student to see the tablet, but did not impede the student's response videos by causing the student to be backlit on camera.
- Observers and interpreters were seated far enough away from the student to not cause disruption, but close enough that they could see the student and their interactions with the proctor and the tablet.

Additional orientation on the tablet and assessment was provided to the proctor, prior to the start of the assessments.

As with the alpha test, assessments were conducted with one student at a time in a dedicated classroom, allowing students to work at their own pace.

All students who were deaf and hard of hearing and present on the day of the school visit were included in the beta test. In the sampled schools, this ranged from 5 to 14 students. Students were assigned to either Form 1 or 2, with the criteria to ensure equal distribution of grades across the two forms.

During the assessment, the proctor sat with the student and was responsible for orienting them to the tablet and software; that proctor also provided support in navigating the tablet as needed. The two observers monitored each assessment and provided general observations of the assessment, the student's engagement, and proctor interactions with the student. Finally, an FSL-English interpreter was also present in each assessment to facilitate the student feedback survey following the assessment's conclusion and provide FSL-English interpretation as needed.[8]

Student response videos—produced for expressive subtasks—were reviewed and scored following the conclusion of all assessments. Twelve scorers in total reviewed student responses. Two scorers were assigned to the same assessments to provide data on interrater reliability.

All scorers were teachers of students who are deaf or hard of hearing, although they were not teachers in the schools where the beta test was conducted. All scorers were highly proficient in FSL; although their FSL fluency was not formally assessed, they were recommended by schools and the deaf community. Scorers were assigned to score assessments based on their familiarity between the regional variations of FSL.

As one of the updates to the beta test, a scoring dashboard was created on the Tangerine:Learn web portal. To review the response videos, scorers accessed the Tangerine:Learn web portal, logged into Tangerine, and accessed individual student records. All responses can be accessed within the web portal, allowing a scorer to review all response

[8] The interpreter did not provide any interpretation support between the proctor and the student. The interpreter only provided this support to the observers to facilitate their understanding of any interactions between the proctor and the student.

videos for a specific student on one page. As with the alpha test, scorers marked the start and end time of review for each learner record, and indicated whether each response was correct, incorrect, or not scorable using a scoring guide. After completing their review of all records, each scorer filled out a feedback form to provide comments on the process and any recommendations for future phases.

Training

The Beta Test Observer and Proctor Training was held on September 20, 2022, in Metro Manila. Participants attended both online and in person, with 4 observers attending online and 8 in person; 24 proctors attending online and 12 in person; and 1 sign language interpreter attending online and 5 in person. Training topics included safeguarding and research ethics, roles and responsibilities, and assessment navigation and review.

In response to learnings generated from the alpha test, the beta test training included sessions on set-up of the assessment space so that student videos could be scored. This included how to lock tablets in landscape mode, the importance of ensuring full and proper signing space in assessment videos, and minimizing background distractions. A training component was added to include a 30-minute to 1-hour review of tools, roles and responsibilities, and the Tangerine:Learn application between observers and proctors before the start of assessments at each school. This was especially important in the Cavite testing locations, as all 12 proctors from the region had attended online.

The scoring training was conducted on September 24 in preparation for the scoring process to begin October 3 and conclude October 10. A scoring training refresher on the web-based system was held on October 1.

ALPHA TEST FINDINGS

The alpha test stage was highly informative in exploring the parameters of a remote EGRA. By testing the three scenarios,[9], RTI and STS gained insight into student interaction with the assessment technology, the type of support needed, and the extent and necessary conditions to which data—specifically, video responses—could be captured by a tablet-based tool and asynchronously scored. Learnings from the alpha test resulted in adjustments to the beta test design: variation of expressive and receptive response models for the Letter Identification and Familiar Word subtasks; elimination of the testing of the proctor scenarios; and modifications to the application, assessment, scoring procedure, and proctor and scoring training.

Findings from Duration Analyses

Duration analysis provided insight into the feasibility and limitation of asynchronous administration through the comparison of the USAID Gabay EGRA and the alpha test assessment.

The average length of the alpha test assessment was calculated at 37.6 minutes, while the average length of the USAID Gabay EGRA was 21.0 minutes. In comparing the two modalities,

[9] The three scenarios tested were Scenario 1: non-FSL-fluent proctor; Scenario 2: non-FSL-fluent proctor + remote online help desk; Scenario 3: FSL-fluent proctor.

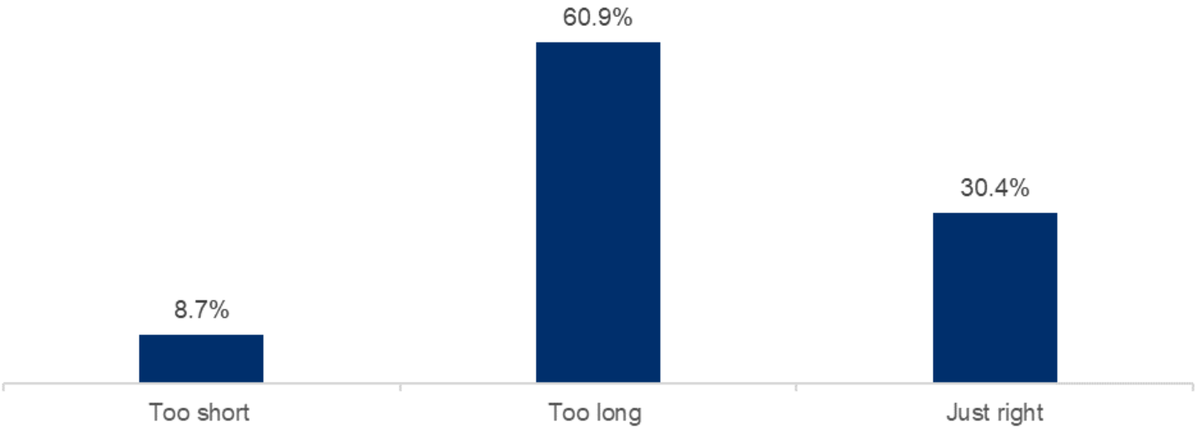
there are several distinctions between the alpha test assessment and USAID Gabay EGRA that may have contributed to this difference in average length.

As previously mentioned, the USAID Gabay EGRA has two additional subtasks as well as double the number of items for six out of eight subtasks. In theory, the USAID Gabay EGRA should have a longer average time. However, the calculated averages show the contrary.

The mechanism of the autostop trigger varies between each modality. In the in-person administration of the USAID Gabay EGRA, an in-person scorer marks the student's responses as correct, incorrect, or no response as the student responds—i.e., synchronous scoring. If the student responded incorrectly or was not able to provide a response, the scorer would mark this in the assessment. An autostop would be triggered if the student responded incorrectly or was not able to provide a response for the first five items of a subtask. For the asynchronous modality, the autostop function is dependent on the student responding to the items with the 'don't know' button—i.e., if the student selects 'don't know' for five consecutive items. As scoring is completed asynchronously, even if a student provides an incorrect response for five consecutive items, the assessment will continue with the remaining items of the subtask, prolonging the student's assessment time.

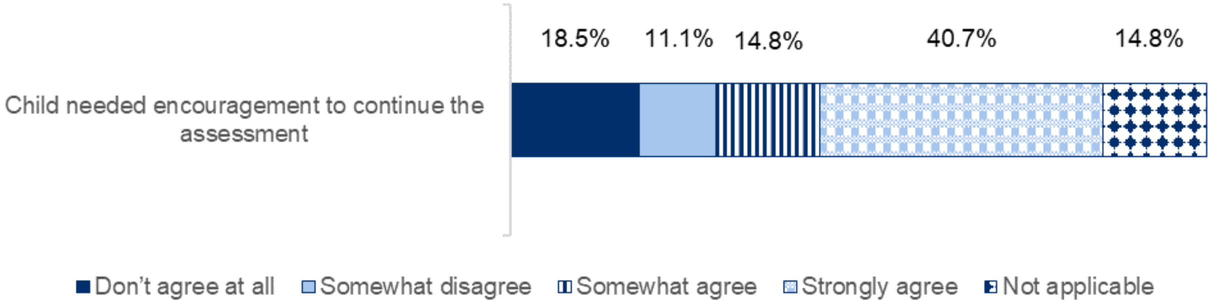
Findings from the student and observer feedback survey also suggest the length of the assessment may not be appropriate for the students in the current format. It was found that 60.9 percent of students felt the assessment was too long (**Figure 6**).

Figure 6. Student Responses: Was the Game Too Short, Too Long, or Just Right?



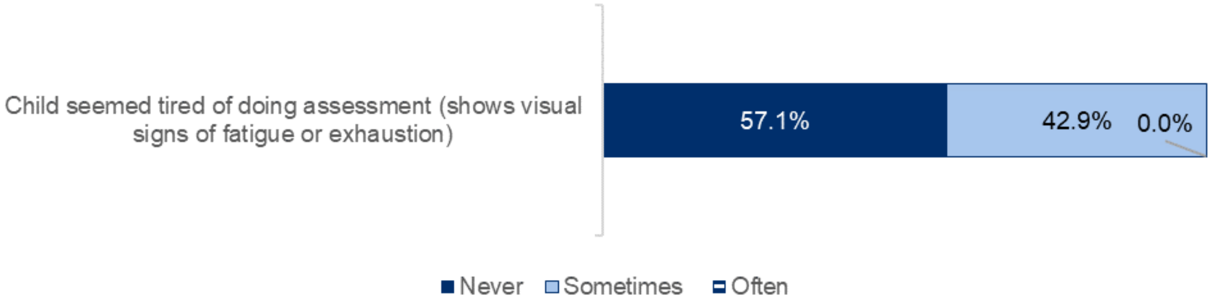
From the observer checklist, observations suggest that students may have struggled with the length of the assessment and that students needed external motivation to complete the assessment. Observers reported for 55.5 percent of assessments, they somewhat agreed or strongly agreed that the student needed encouragement to continue the assessment (**Figure 7**).

Figure 7. Observer Responses: Child Needed Encouragement to Continue the Assessment



Observers also reported that 42.9 percent of the students seemed tired of doing the assessment through visual signs of fatigue or exhaustion (**Figure 8**).

Figure 8. Observer Responses: Child Seemed Tired of Doing Assessment



To explore methods to address assessment length, RTI and STS adapted two assessment forms. Form 1 was the same assessment used during the alpha test, with five of the subtasks using expressive response formats (Expressive Vocabulary, Sign Language Comprehension, Letter Name Identification, Familiar Word Reading, and Sentence Reading Comprehension) and one subtask using receptive response format (Receptive Vocabulary). Form 2 maintained expressive response formats in the Expressive Vocabulary, Sign Language Comprehension, and Sentence Reading Comprehension subtasks. However, Letter Name Identification and Familiar Word Reading subtasks use receptive response formats in Form 2.

In the receptive format, students would not record their response through video. For Letter Name Identification and Familiar Word Reading, the receptive format would show a video with the FSL sign of the item. The student would provide a response by selecting their response from four options.

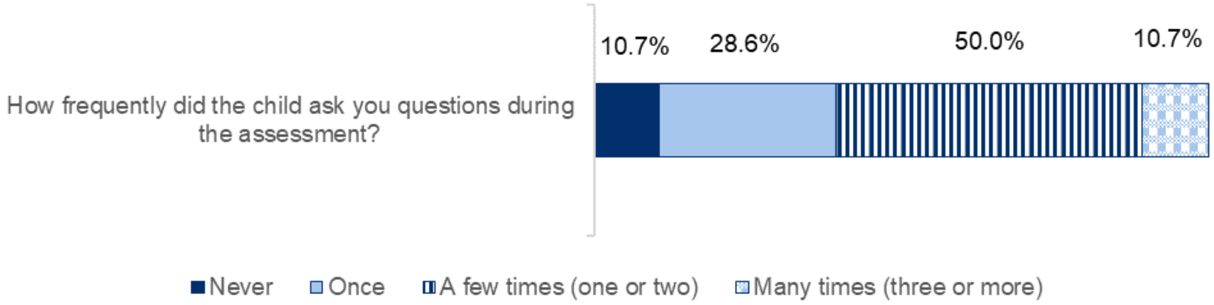
Receptive subtasks were hypothesized to reduce the assessment duration, as students would spend less time recording their responses, and further, the receptive format may require less external support for the students. Including more receptive subtasks would also reduce time spent on scoring with less items to score manually. This could also lead to automation of scoring.

Findings from Proctor Scenarios

In varying the FSL fluency of proctors, the alpha test investigated proctor fluency levels and opportunities or limitations to support students through the assessment. In both Scenario 2 and Scenario 3, a proctor who was fluent in FSL was available to the student for questions and FSL support. In Scenario 2, this FSL support was provided through an online help desk available on Zoom on a second tablet, while Scenario 3 provided in-person support through an in-person FSL-fluent proctor. Scenario 1, in contrast, included an in-person proctor who was not fluent in FSL. Testing the scenarios suggests the online help desk is not an appropriate support for students, nor is it scalable and feasible.

In 8 out of 12 observations, the online help desk personnel reported that the student never asked for help from the online help desk during the assessment. Conversely, students interacted with in-person proctors at much higher rates. In-person proctors—both FSL-fluent and non-fluent—reported that 60.7 percent of students asked questions a few times (one or two) or many times (three or more) during the assessment (**Figure 9**).

Figure 9. Proctor Responses: How Frequently Did the Child Ask You Questions During the Assessment?



During a debrief consultation at the conclusion of the alpha test, participants—observers, proctors, online help desk support, and interpreters—shared their observations that only students with higher FSL skills[10] were reported to be more able and willing to interact with the online help desk. The difference in engagement may be due to in-person proctors being able to see students' non-signed cues better than the online help desk personnel. As a result, they were able to provide more unprompted help to lower-level students in response to hesitation or confusion. The online help desk, in contrast, relied on students being able to formulate and directly pose their questions to the personnel.

An additional challenge with remote support was the Internet connection required to host the online help desk. Occasionally, the video connection for online help desk support would drop, requiring on-site information technology (IT) support to pause the assessment in order to reconnect to the Zoom meeting. This proved to be disruptive to the student.

[10] FSL skills, in this context, were rated by the participants informally. Because there is not a standardized FSL assessment widely utilized in the Philippines, the participants informally assessed the students' FSL skills by their ability to communicate with the participants during the assessment and following student feedback survey.

At a larger scale, the online help desk does not seem feasible or sustainable. Stable Internet connections may not be available at all schools or testing locations. IT personnel may also be limited across schools, requiring added travel costs for IT personnel. Future cost-effectiveness may be impacted with this modality with additional personnel labor and training—both for IT support and the online help desk support, secondary tablets to host the Zoom meeting during assessment, and potential costs for mobile Internet.

An important contextual consideration of proctor support is the limited availability and quantity of people who are fluent in FSL in the Philippines. Requiring FSL-fluent proctors would likely necessitate travel to testing locations and disengaging proctors from their employment, which may be difficult. Many potential FSL-fluent proctors are teachers of students who are deaf or hard of hearing. Engaging these teachers would leave their students at a disadvantage without their teachers for potentially long periods of time. If the assessment were conducted at a larger scale with FSL-fluent proctors, this may contradict the goals of a remote assessment that is adaptable to emergencies or other challenging contexts.

Given the results of the alpha test, RTI, RBI, and STS concluded that the beta test would focus on proctors with varying levels of FSL to provide in-person support. This also allowed for higher insight into the comparison between the receptive and expressive subtask forms.

Findings from Scoring and Scorer Feedback

To score the expressive subtasks, RTI, STS, and RBI engaged three scorers to review student responses. Scorers were provided a simplified dataset, which contained the student ID, links to the student response videos, and prompts to score the student responses. Scorers would log into the Tangerine:Learn web portal, open the link to the student response video in the web browser, review the student response, and respond to two questions. The first question asked, "Is the video scorable?" Answer options were "Yes" or "No." The second question asked, "Is the response correct, incorrect, or not scorable?" with the answer options as "correct", "incorrect", and "not scorable." The scorer completed a qualitative feedback form to provide more information on why they may have deemed responses not scorable.

In their comments, scorers reported that it took 30 to 60 minutes to score an assessment. Scorers also found the Excel file with video links challenging to navigate. All scorers reported difficulties in accessing the videos. Scorers reported non-active hyperlinks in the dataset as well as some videos simply not playing. Further investigation by RTI showed that some of the videos did not play as a result of students' pressing the record button twice. Double tapping stopped the record function before any response was captured, so that it appeared to a scorer as a video not playing.

Scorers found a high variability in quality of student response videos due to angle, lighting, or visibility of full and proper signing space.

To respond to these challenges, RTI, STS, and RBI implemented the following adjustments:

- RTI created a dashboard feature on the web platform of Tangerine:Learn. The dashboard would allow scorers to access all a student's response videos on the same page. The scorer would not have to open different links for each video.
- RTI added a feature to prevent a student from moving on to the next page if a recorded video was too short.

- STS incorporated more training for the beta test to emphasize how to show the students how to use the record button, how to monitor the students as they use the record function, and how to set up the tablet to ensure the students' full and proper signing space would be captured on the camera.
- Additional scorers were engaged in the beta test to examine interrater reliability.

Additional Findings from Field Observations

RTI, STS, and RBI made several observations around the application and assessment, which resulted in additional adjustments to the beta test.

The orientation and dimensions of the application caused notable disruptions for students. During the assessment, the tablet would revert from landscape to portrait mode. To correct this, the proctor had to remove the tablet from the stand, correct the orientation, and then allow the student to continue the assessment. The dimensions did not fit the screen of the tablet. Students had to scroll down to see response options. This was difficult for some students, as there was not any indicating icon or instruction in the application to signal to scroll down to see the response options.

The layout of the application was a notable area for improvement. In qualitative feedback from the field, participants reported that the screen should be more efficiently utilized. The response options were interpreted to be very small for students. Students were observed to move close to the screen and squint their eyes throughout the assessment. Participants noted that students also struggled with recognizing when their answer option had been selected. The border around the item when a selection was made was not distinguishable. Many students would select their response multiple times, not being able to see that their response had already been selected. In some cases, students interpreted this to mean that their selected response was incorrect, and therefore, would change their response.

Students were also observed to start video recording prior to watching the video prompt. Therefore, students would start recording, play the video prompt, and then restart the video recording to record their responses.

To address reported issues in the application, RTI increased the size of some components within each screen in the application. Additional training was provided to the proctors to ensure that they knew how to lock the tablet in landscape orientation. The thickness of the border for selected items was increased to signal a larger contrast for students. The record button was made unavailable to the student until the video prompt was watched, which would help guide the student in sequencing the assessment.

In addition to observations around the application, participants also made recommendations for the videos within the assessment. In demonstration videos where a student was shown modelling how to respond to a demonstration item for each subtask, an outdated version of the application was shown. The record buttons did not match with buttons in the application at the time of the alpha test. The interpreter in the video motions to his left when mentioning the response options. However, the response options are positioned below him on the screen. The alpha test participants reported that the FSL instructions may be too long and confusing for young students. During the debrief consultation, proctors reported that they often needed to provide clarification on the instructions through simple gestures and cues. In the observation checklist, this was further emphasized—in 75.0 percent of observations, observers reported that students seemed confused or stuck sometimes or often during the assessment.

The reports led to a revision and reproduction of the instruction and demonstration videos. STS and RBI revised the video scripts to make the language simpler. RBI worked with FSL artists and deaf mentors to revise the sign language interpretation in the videos, making the FSL more appropriate to young students.

BETA TEST FINDINGS

The beta test was an opportunity to further explore the assessment parameters that support or inhibit the remote EGRA's scalability for students who are deaf or hard of hearing.

Interrater Reliability Across Form 1 and Form 2

An important finding from the beta test analysis is on the scoring of assessments. For expressive subtasks, students record their responses through video. These responses are later reviewed by a scorer. For the beta test, two scorers reviewed and scored each assessment to provide insight into interrater reliability (IRR). In the analysis of this process, it was found that the scoring process increases the cost of the assessment due to the need to identify, hire, train, and support scorers. Additionally, the scoring component of expressive subtasks introduces room for error and disagreement across scorers that can impact the reliability of the results. If, however, expressive subtasks are highly desired, this error and disagreement can be mitigated by creating and implementing standard protocols for scoring videos.

For the beta test, RTI conducted IRR tests across the scorers' answers. IRR is a measure of agreement between scorers on the answer—correct or incorrect—on each question of the learning acceptive. In**Table 4** and **Table 5,** it is shown as the percentage of answers that were scored the same across both scorers for one student's answer form. Conducting IRR analyses allows us to see the level of agreement across scorers and across subtasks. It highlights subtasks that were more difficult to score, or at least had variation in scoring practices among scorers.

Reported in **Table 4** and **Table 5** are the average IRR scores and score ranges across subtasks and total disaggregated by form and location. Looking first at the total assessment average percent aggregate (TAAPA), which reflects the average percent agreement between sorcerers across all subtasks, there is notable variation. Looking within Form 1 scores, the agreement between scorers varies by testing location. Cebuano Form 1 has the highest TAAPA scores at 94 percent, which reflects the high level of agreement among scorers across all five subtasks. Similarly, the average TAAPA score for Tagalog Form 1 in Cavite schools was 92 percent. Among this sample, both Expressive Vocabulary and Language Comprehension scores fall below the 90 percent threshold (87 percent and 88 percent, respectively). However, not all scorers performed similarly on Form 1. The IRR analysis on Tagalog Form 1 scores from Quezon City schools reports much lower levels of agreement—76 percent on TAAPA. The results underscore the need for standard protocols on scoring that are accessibility to all the individuals recruited to be scorers.

Form 2 has lower average percent aggregate scores than Form 1, as shown in **Table 6**; however, this is largely because Form 2 did not contain the lower-level subtasks that generally had higher levels of agreement. Even with the lower averages, there still exists variation between locations. The majority of the aggregate percent scores by subtask fall below the 90 percent threshold. Notably, the performance on scores for Form 2 are the reverse of those seen for Form 1, with the highest level of agreement coming from scorers on Tagalog Form 2 from Quezon City (85 percent) and the lowest on Form 2 in Cebuano (75 percent). This suggests that ambiguity in scoring guidelines exists at the individual level rather at the location level.

Table 5. Beta Test IRR Scores Form 1

Assessment Form		Average Percent Aggregate					
		Expressive Vocabulary	Sign Language Comprehension	Letter Name Identification	Familiar Word Reading	Sentence Reading Comprehension	Total Assessment
Cebuano Form 1	Avg	92%	91%	99%	93%	96%	94%
	Range	60%–100%	60%–100%	92%–100%	57%–100%	80%–100%	81%–100%
Tagalog Form 1 (Cavite)	Avg	87%	88%	98%	94%	92%	92%
	Range	10%–100%	0%–100%	95%–100%	71%–100%	20%–100%	62%–100%
Tagalog Form 1 (Quezon City)	Avg	74%	68%	80%	79%	78%	76%
	Range	0%–100%	0%–100%	0%–100%	0%–100%	0%–100%	0%–100%

Table 6. Beta Test IRR Scores Form 2[11]

Assessment Form		Average Percent Aggregate			
		Expressive Vocabulary	Sign Language Comprehension	Sentence Reading Comprehension	Total Assessment
Cebuano Form 2	Avg	93%	67%	65%	75%
	Range	20%–100%	0%–100%	0%–100%	30%–100%
Tagalog Form 2 (Quezon City)	Avg	83%	81%	91%	85%
	Range	50%–100%	20%–100%	60%–100%	63%–100%

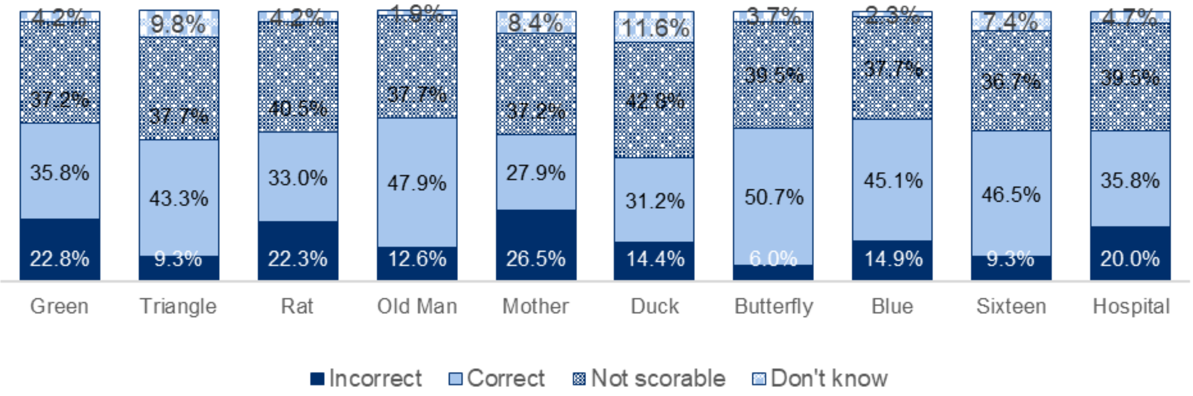
One of the reasons for the levels of disagreements across subtasks was the number of answers coded as a “not scorable” by certain scorers. During the scoring exercise, scorers answered two questions about the student response. First, the scorers considered whether they could

[11] Due to logistical challenges, two scorers did not review the Tagalog Form 2 from the schools in Cavite and therefore, IRR analysis was not conducted for these assessments.

understand the student's response. The scorers specifically examined whether there were any issues with the video file with regard to their ability to see the student and decipher the student's response. No rubric was provided for this process. The criteria were simplified into one question "Is this video scorable?," to which the scorers either answered "yes, scorable" or "no, not scorable." Scorers provided qualitative feedback on reasons why videos were not scorable. Second, the scorers reviewed the response, compared the response to a provided answer guide, and answered the second question "Is the answer correct, incorrect, or not scorable?" Scorers used the scoring feedback form to provide qualitative feedback on the overall feasibility of scoring assessments and any challenges encountered.

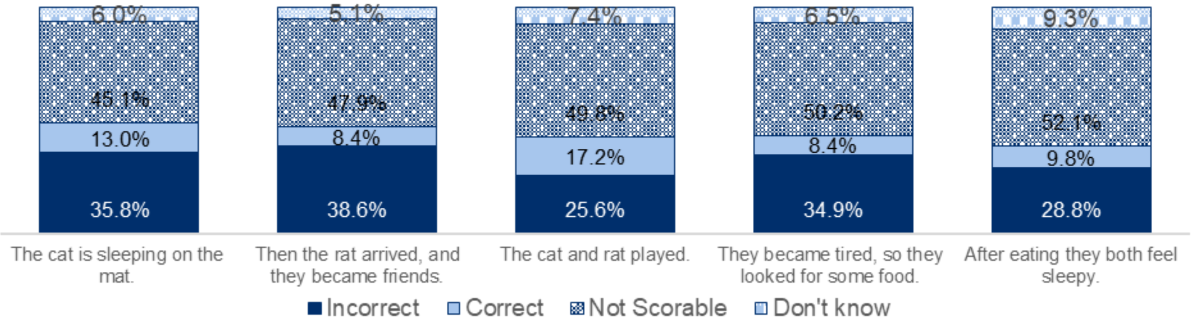
Qualitative feedback on scoring provided some detail as to why scorers had different perspectives on what made a scorable answer. Some scorers coded answers as not scorable if the student's hand went out of the camera view even if their answer was legible. Others were far more flexible on what they considered a complete submission. This further complicates the varying levels of success students had when filming their responses to expressive items. Ultimately, this underscores the need for comprehensive and specific scoring guidelines that remove any ambiguity on what should be considered a scorable answer from the perspective of the scorers. In **Figure 10**, the results for Expressive Vocabulary, which measures students' ability to produce the sign for common vocabulary words, show that on average 15.8 percent of students were scored as having signed the incorrect word and an average of 38.7 percent of all submitted video responses were marked as not scorable.

Figure 10. Expressive Vocabulary Item Scores



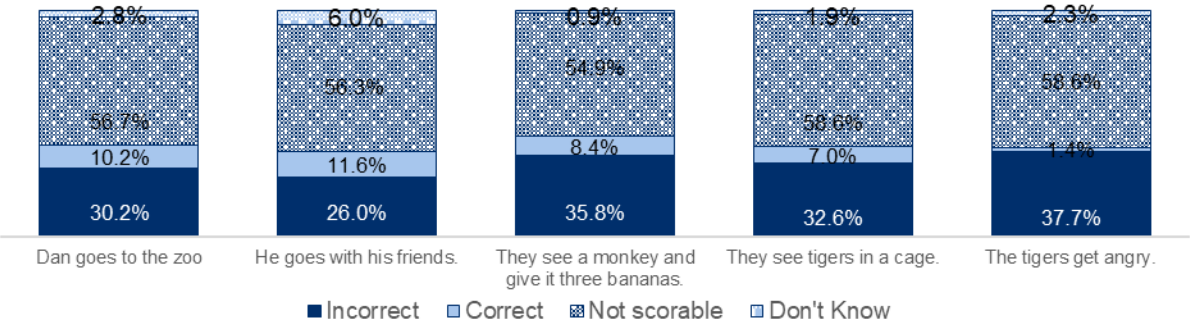
As the complexity of the levels of subtasks increased, so did the rates of answers marked as not scorable. The Sign Language Comprehension subtask measured students' ability to understand FSL grammar and comprehend sentences. Student responses were scored as not scorable an average of 49.1 percent across the five language comprehension questions (**Figure 11**).

Figure 11. Sign Language Comprehension Item Scores



Similarly, Sentence Reading Comprehension, which measures students' ability to read and comprehend connected questions, had the highest average of incorrect responses (32.5 percent) and not scorable responses (57.0 percent) shown in **Figure 11**.

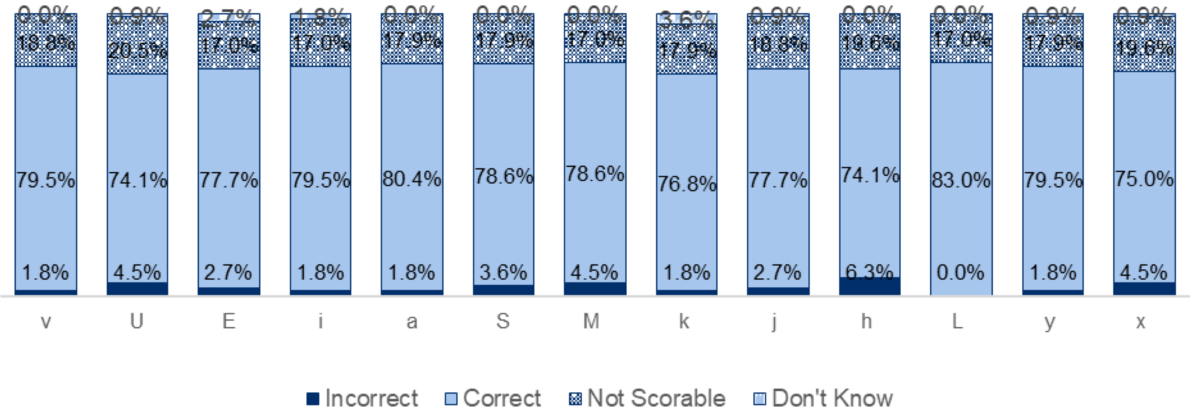
Figure 12. Sentence Reading Comprehension Item Scores



The receptive forms of the subtasks are a successful modality to overcome the high variance in IRR, especially for higher level subtasks. Students scored high on both expressive and receptive letter name subtasks with comparatively lower numbers of not scorable responses on the expressive subtasks. Students answered 12.2 out of 13 of the letter names correctly on the expressive form of the Letter Name Identification subtask. Students performed similarly on the receptive Letter Name Identification subtask, correctly answering 11.3 letters on average.

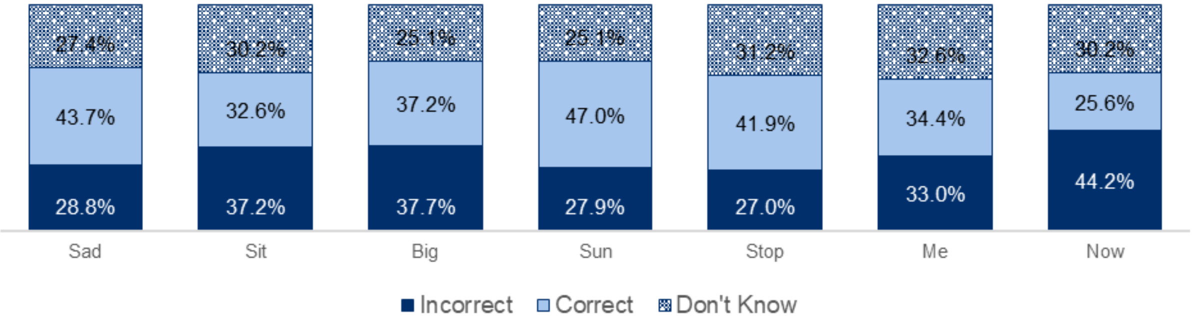
In comparison to the other expressive subtasks, the percentage of answers scored as not scorable for the expressive Letter Name Identification subtask was relatively low—on average only 18.0 percent of submitted answers were scored as not scorable (**Figure 13**).

Figure 13. Expressive Letter Name Identification Item Scores



In contrast, the Familiar Word Reading subtask saw notable differences between the scoring across the two forms (**Figure 14**). In Form 1, students were asked to record the correct sign for the given word. Whereas in Form 2, students were asked to select the correct word from a set of multiple-choice outcomes. The mean score (out of 7) is 2.9 for expressive answers and 3.4 for receptive answers. While we cannot make statistical comparisons between the two groups because students were not randomly assigned to either form but were programmatically sorted, we can highlight the higher average score among those responding to the receptive subtask.

Figure 14. Receptive Familiar Word Reading Item Scores



Findings from Duration Analyses

One of the main findings from the alpha test and a driving factor in testing receptive question modalities was that the learning assessment took too much time for learners to take. In order to make the asynchronous assessment tool accessible, the duration must be appropriate for the test population.

Performing a duration analysis also provides insight into the feasibility and limitation of asynchronous administration through the comparison of receptive and expressive question types. Theoretically, it would be expected that expressive questions would take longer, requiring students to record their answers in comparison to selecting them for receptive questions. However, the average duration difference between Form 1 (37.1 minutes), which contained only expressive question modalities, and Form 2 (36.1 minutes), which had two receptive question modalities, was small. Upon further inspection we can see that this difference was moderated by the students' FSL ability[12] (**Figure 15**). For students with low FSL ability, the assessment took on average 38 minutes. As student's FSL ability increases, the gap between the duration of Form 1 and Form 2 widens (**Figure 16**), suggesting that students with higher FSL ability were able to move more quickly through the receptive questions than the expressive questions.

The results from the duration analysis suggest that while receptive questions will mitigate scoring and IRR obstacles, they are unlikely to decrease the time of the assessment. Generally, the majority of students in early grades are unlikely to have high levels of FSL. Therefore, using receptive subtasks will likely take equivalent amounts of time as using expressive question types.

Figure 15. Length of Assessment by Form Type and Student FSL Ability

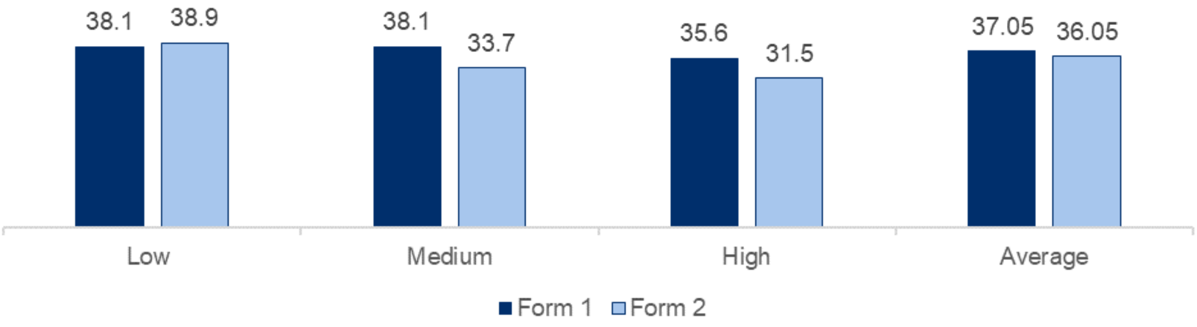
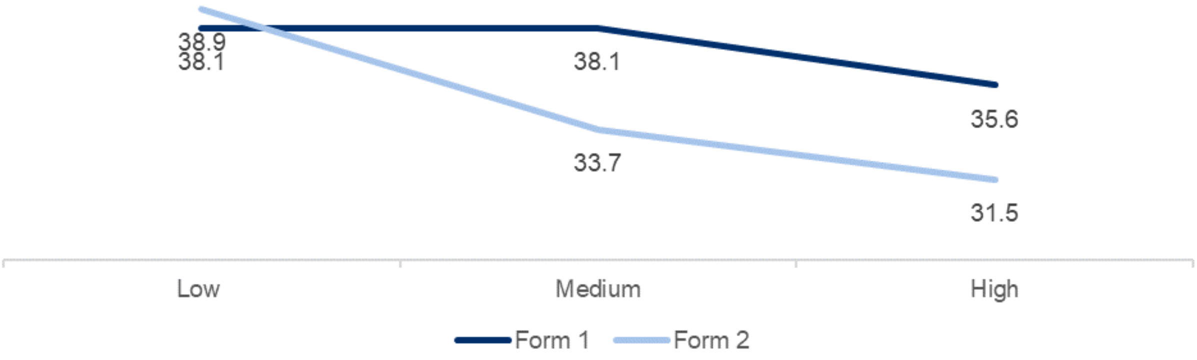


Figure 16. Length of Assessment by Form and Student FSL Ability



[12] Information on student FSL ability was provided by the student's teacher. As students are not typically assessed in FSL in the Philippines and due to the lack of standardized FSL assessment, this rating by the teacher is an approximation and may consider the student's ability to communicate, rather than represent a formal assessment.

Comparison of Expressive and Receptive Subtask Scores

A key facet of analyzing the use of receptive versus expressive testing modalities is to understand how students performed under each testing condition, and ultimately, whether they function comparably. During beta testing, two of the subtasks—Letter Name Identification and Familiar Word Reading—were given in either the receptive or expressive subtasks. Students performed well on the letter naming subtask with mean scores of 12.21 (out of 13) on the expressive version of the subtasks and 11.34 on the receptive subtasks, shown in**Table 7**. For the purposes of this report, more important than the students' high performance is that they performed equivalently on both modalities. Similarly on the more difficult subtask, familiar word naming, students who took both the expressive and receptive versions of the subtasks scored within similar ranges. Students' mean score on the expressive form was 2.87 (out of 7) and on the receptive form was 3.41.

Table 7. Subtasks Scores: Receptive and Expressive Comparisons

Subtask	Form	Mean
Letter Name Identification—Expressive	Form 1	12.2
Letter Name Identification—Receptive	Form 2	11.3
Familiar Word Reading—Expressive	Form 1	2.8
Familiar Word Reading—Receptive	Form 2	3.4

It is important to note why we did not perform any significance testing between the two scores in order to statically determine whether the scores were equivalent. Students were not randomly sorted into testing modalities due to pragmatic decisions. Therefore, the reasons scores were significantly different could be due to confounding variables—i.e., that the students who took Form 1 differed systematically from those who took Form 2 in a way that affected their scores across subtasks. **Table 8** reports the descriptive statistics for all the subtasks disaggregated by form. The mean percent scores suggest systematic differences between the abilities measured on the assessment between students in Form 1 and Form 2. Specifically, students taking Form 1 performed similarly but on average better than those who took Form 2. Notably, we can see that students on who took Form 2 performed lower on average in every subtask, except the receptive form of Familiar Word Reading, where students on average answered 48.8 percent of the questions correctly in comparison to 41.1 percent on Form 2.

Table 8.Summary Subtasks Percent Scores

Subtask	Form	Mean
Receptive Vocabulary	Form 1	72.6
Receptive Vocabulary	Form 2	68.5
Expressive Vocabulary	Form 1	62.9
Expressive Vocabulary	Form 2	51.7
Sign Language Comprehension	Form 1	19.2

Subtask	Form	Mean
Sign Language Comprehension	Form 2	18.2
Letter Name Identification—Expressive	Form 1	93.9
Letter Name Identification—Receptive	Form 2	87.2
Familiar Word Reading—Expressive	Form 1	41.1
Familiar Word Reading—Receptive	Form 2	48.8
Sentence Reading Comprehension	Form 1	19.2
Sentence Reading Comprehension	Form 2	5.0

A clear next step in understanding the uses and limitations of receptive subtasks is to use a quasi-experimental approach that can account for potential confounders. The treatment design would compare receptive and expressive question modalities. With randomization and a large enough sample size, the project could control for confounding variables that could also affect test size. In doing so, researchers can statistically analyze whether the receptive question modality affects test performance.

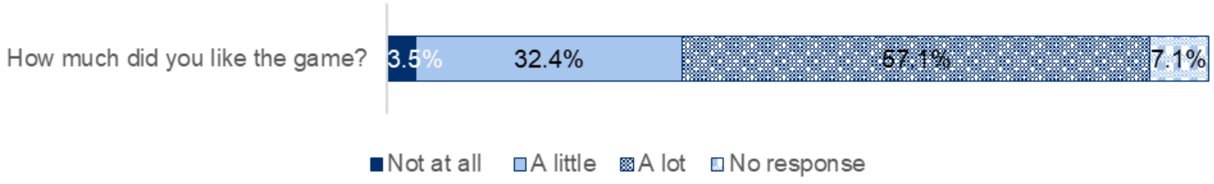
Findings From Student Feedback

Student responses to the feedback form indicate that the assessment was developed appropriately for the audience, with the majority of students understanding the FSL, enjoying the game, and asking for assistance when needed. Findings from the student feedback form suggest improvement could be made in shortening the length of the assessment which, would help with fatigue, we hope.

Of the 177 students who took the assessment either through Form 1 or Form 2, 172 students agreed to answer the follow-up questions. It is important to keep this in mind when interpreting the answers to the questions students were asked . It is very likely that the students who declined to answer the feedback questions are systematically different from those who opted to provide feedback that likely correlated with their experience with the assessment.

The overwhelming majority of students liked the game,[13] with over 85.5 percent of students selecting that they liked it either a lot or a little (**Figure 17**).

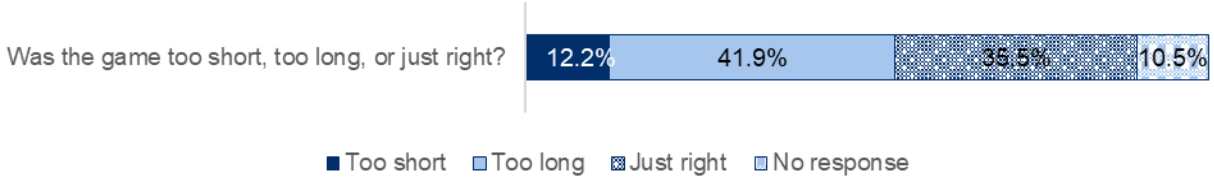
Figure 17. How Much Did You Like the Game?



[13] When speaking to the student, the assessment was referred to as a game, to mitigate any negative reaction or association a student may have to “assessment” or “application.”

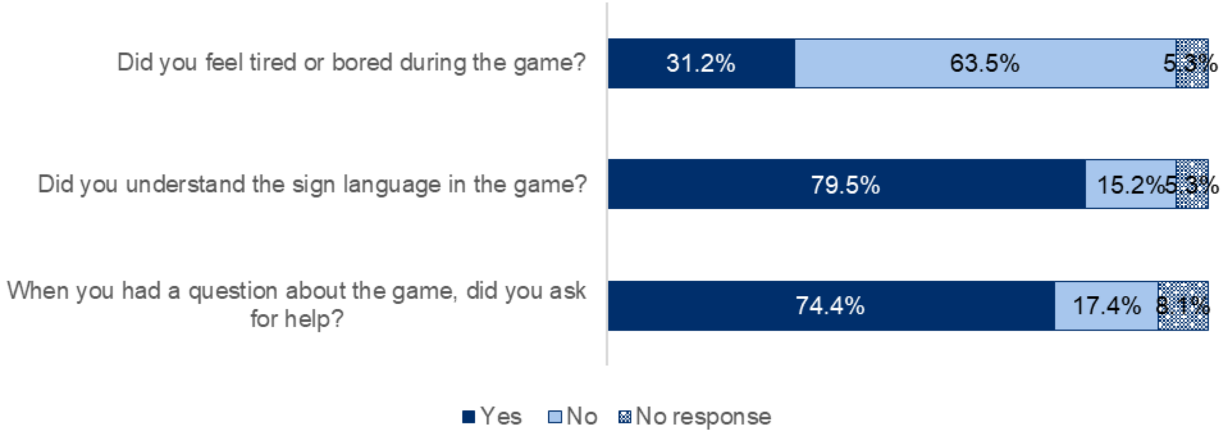
A little over half of the students said they did not feel tired or bored during the game, but this suggests room for improvement (**Figure 18**). This improvement could potentially be found in the length, as 41.9 percent of students felt the game was too long.

Figure 18. Was the Game Too Short, Too Long, or Just Right?



The majority of students understood the sign language in the game (79.5 percent) suggesting it was developed appropriately (**Figure 19**).

Figure 19. Student Feedback Survey Responses

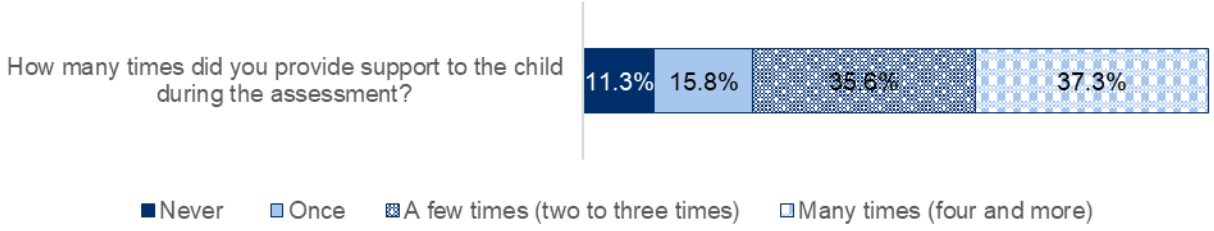


Findings From Proctor Feedback

Proctors played a pivotal role during the asynchronous assessment conducted in the beta testing. They were utilized often by students and mainly provided help in understanding the instructions and recording their responses.

The majority of proctors (72.9 percent) provided support two or more times (**Figure 20**).

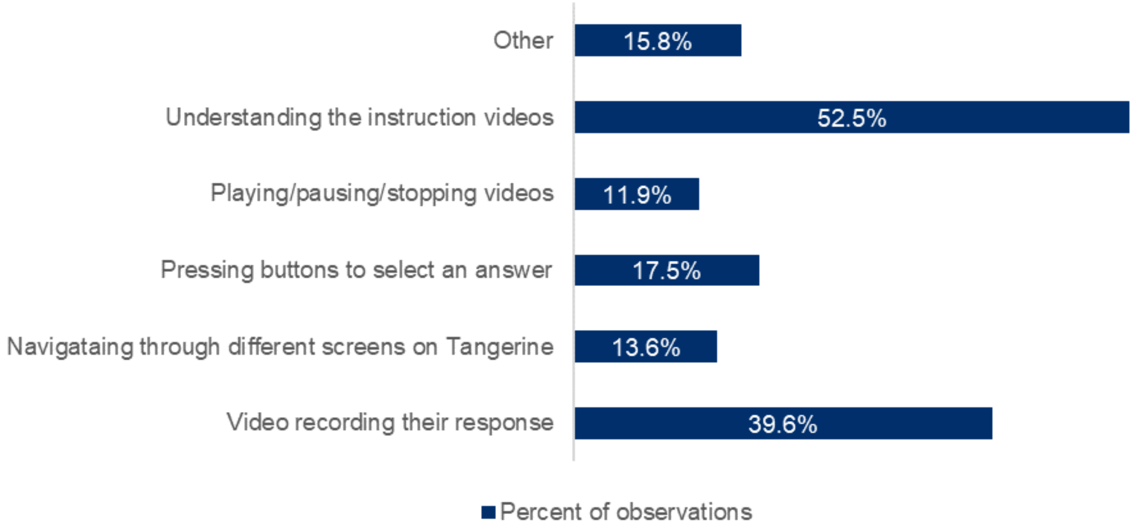
Figure 20. Frequency of Proctor Support



Despite the high rate of assistance, the types of assistance proctors indicated they provided were not very variable (**Figure 21**).

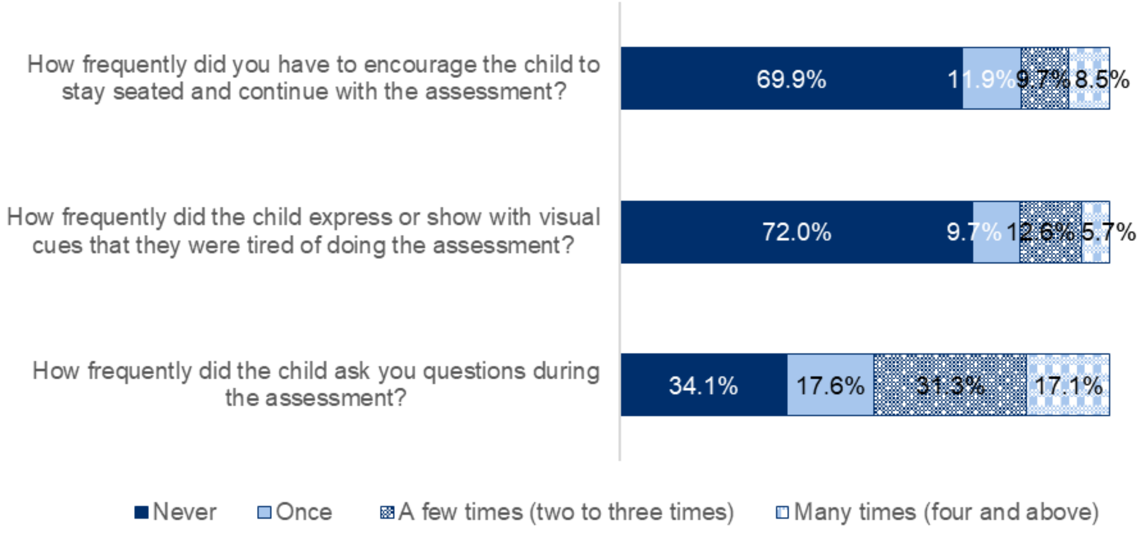
They most often provided assistance with understanding the instructions on the videos (52.54 percent) and with recording their responses (39.6 percent). Students were less likely to request help from the proctor to navigate through the different screens (13.6 percent), pressing buttons to select and answer (17.5 percent), and playing/pausing/stopping the videos (11.9 percent).

Figure 21. Frequency of Type of Proctor Support



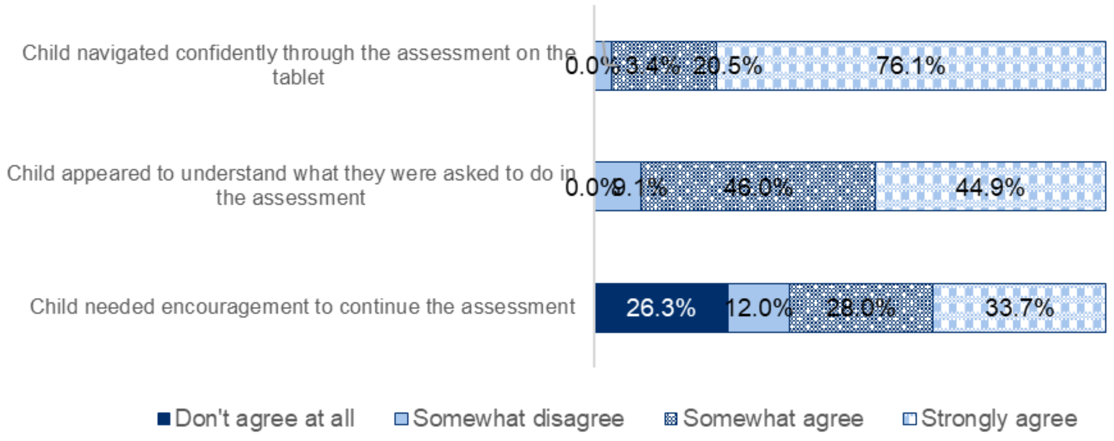
Proctors were less necessary when it came to test room management. The majority of the proctors (69.9 percent) did not need to encourage children to stay seated and continue with the assessment (**Figure 22**). Nor did they note that children were expressing that they were tired of doing the assessment (72 percent). Proctors' main function in the room remained as a facilitator for the assessment; 65.9 percent reported being asked a question during the assessment at least once and 48.3 percent two or more times.

Figure 22. Proctor Feedback Responses: Frequency of Assistance



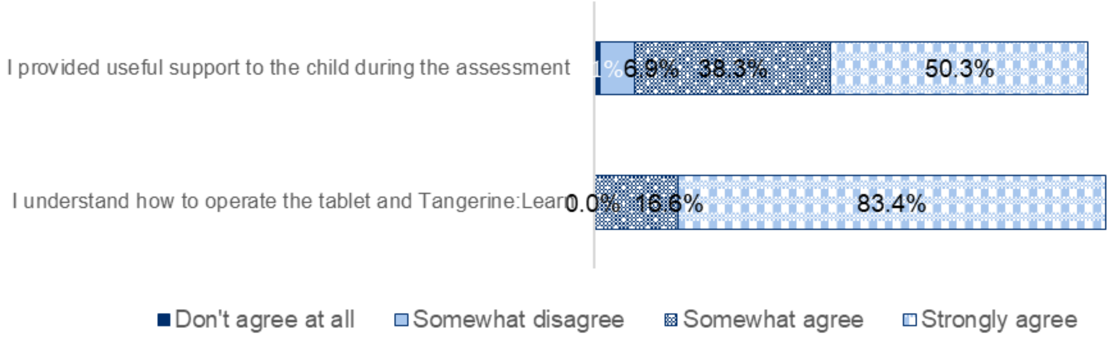
In addition to acting as facilitators throughout the assessment, proctors provided valuable insight on the experience of students during beta testing. Overall, proctors made positive observations on the students' experience during the learning assessment, which suggests the design of the asynchronous test is successful. The overwhelming majority of proctors agreed that learners navigated confidently through the assessment (**Figure 23**). In fact, only six proctors (3.4 percent) somewhat disagreed with that statement, and no proctors strongly disagreed. Proctors did note that more often than not students needed encouragement to continue the assessment (61.7 percent). It is likely that while the children were able to understand and navigate the assessment, they needed encouragement to finish possibly because of the length.

Figure 23. Proctors Assessment of Student Experience



Nearly 90.0 percent of proctors agreed that they provided useful support to the students during the assessment (**Figure 24**). Moreover, only six proctors said that the student did not need support, underscoring the importance of having a proctor for asynchronous assessments. The training provided during the beta test proved sufficient as all proctors agreed they understood how to operate the tablet and Tangerine:Learn.

Figure 24. Proctor Self Evaluation



Lastly, while proctors believed they were able to provide support during the assessment there is room for growth in their familiarity and experience with the assessment system, Tangerine:Learn. However, remote training appears successful and feasible. About 16.5 percent of proctors only somewhat agreed that they understood how to operate the tablet and Tangerine:Learn. In 83.0 percent of observer observations and 83.0 percent of proctor observations, surveys reported that proctors appeared to understand how to operate the tablet and application—a high proportion. Mechanisms, however, should ensure there are no knowledge gaps for proctors. Remote trainings should provide support to the minority of proctors who may need further confidence building with the technology.

Findings From Observer Feedback

Observers were utilized to report on the behavior of both learners and proctors during the assessment to answers the project's research questions and ultimately evaluate the scope and limitations of asynchronous assessment. It is important to note that observers did one form per assessment—i.e., per student—not per proctor.

Observers reported very positively on the behavior of proctors during the assessment, suggesting that the selection and training of proctors during beta testing was successful and can be used a guide for future assessments. Every observer noted that the proctors arranged the desk and chairs so that the tablets were at the learner's eye level, that they showed the learner the tablet and Tangerine:Learn application, and that the proctor indicated for the learner to press the first button to start the assessment (**Figures 25** and **26**). Additionally, they noted that the technology used during the assessment worked successfully in almost all cases. Only one observer reported that the proctor ended the assessment early because of tablet malfunction and only three noted that the proctor had to end the assessment early because the child continuously was unable to navigate the assessment or seemed too uncomfortable to continue.

Figure 25. Observation Data on Proctors

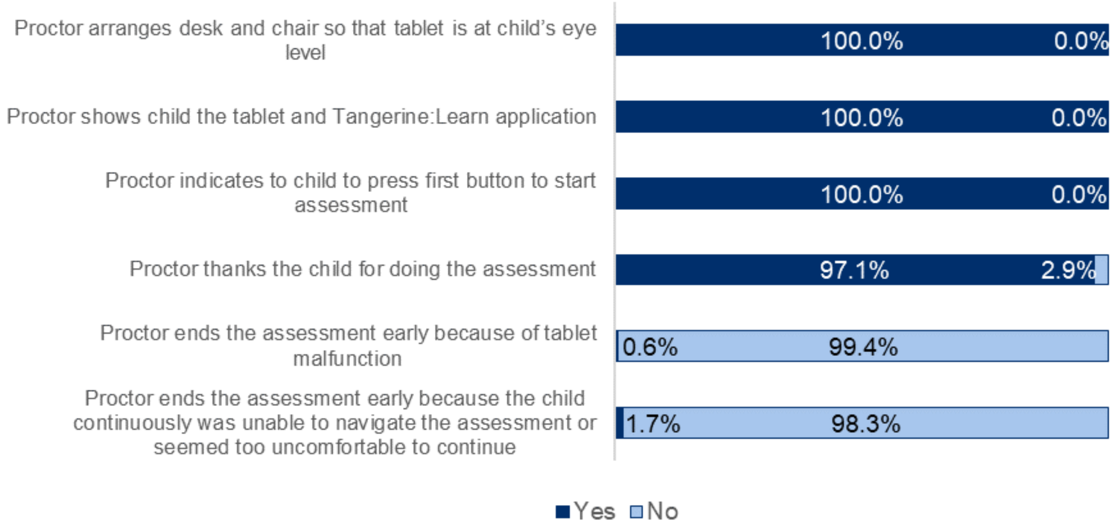
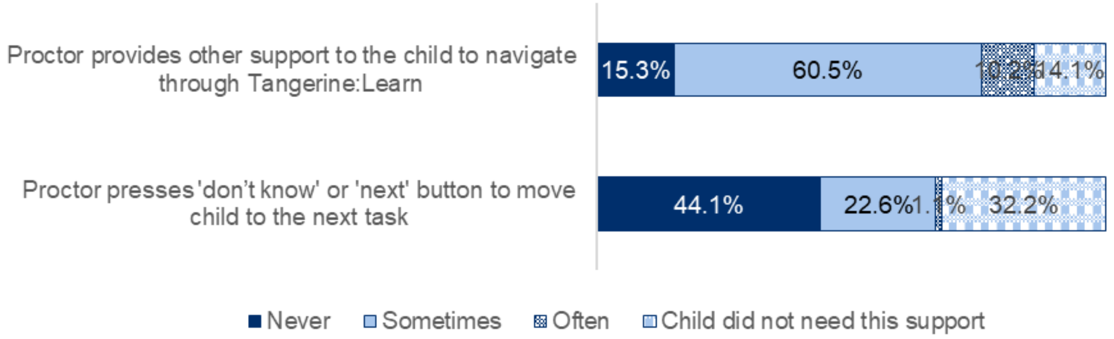


Figure 26. Additional Observation Data on Proctors



The data from observers mirror what was said by proctors: their primary role was not in test room management. Observers either noted that proctors never asked learners to stay seated and continue with the assessment (49.7 percent) or that they did not need this support (40.1 percent)—both indicating that overall learners did not need management during the assessment from proctors (**Figures 27** and **28**). While proctors did not need to exercise control over the learners during the assessment, observers highlighted that proctors facilitated the assessment by providing necessary encouragement to learners. Similar to levels reported by proctors, nearly half of observers agreed that children needed encouragement to continue the assessment.

Figure 27. Observer Data on Proctor Behavior Frequencies: Test Room Management

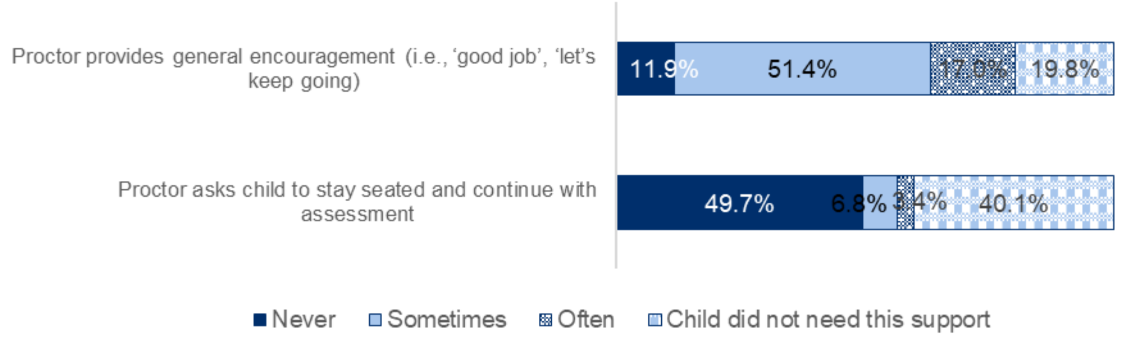
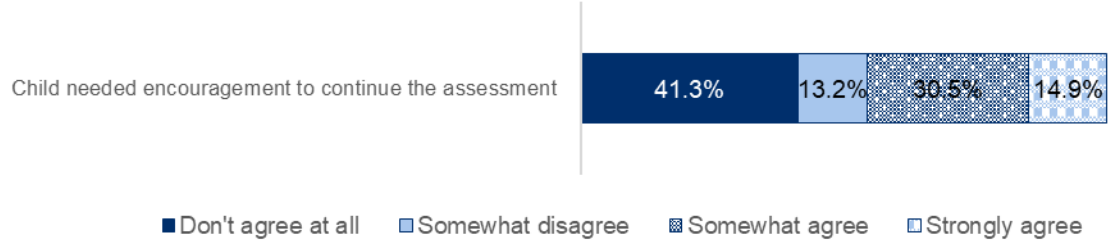
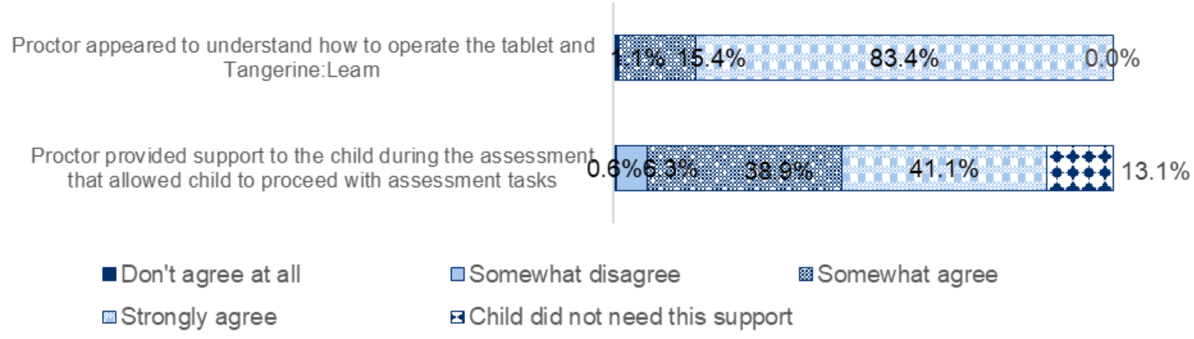


Figure 28. Additional Observer Data on Proctor Behavior Frequencies: Test Room Management



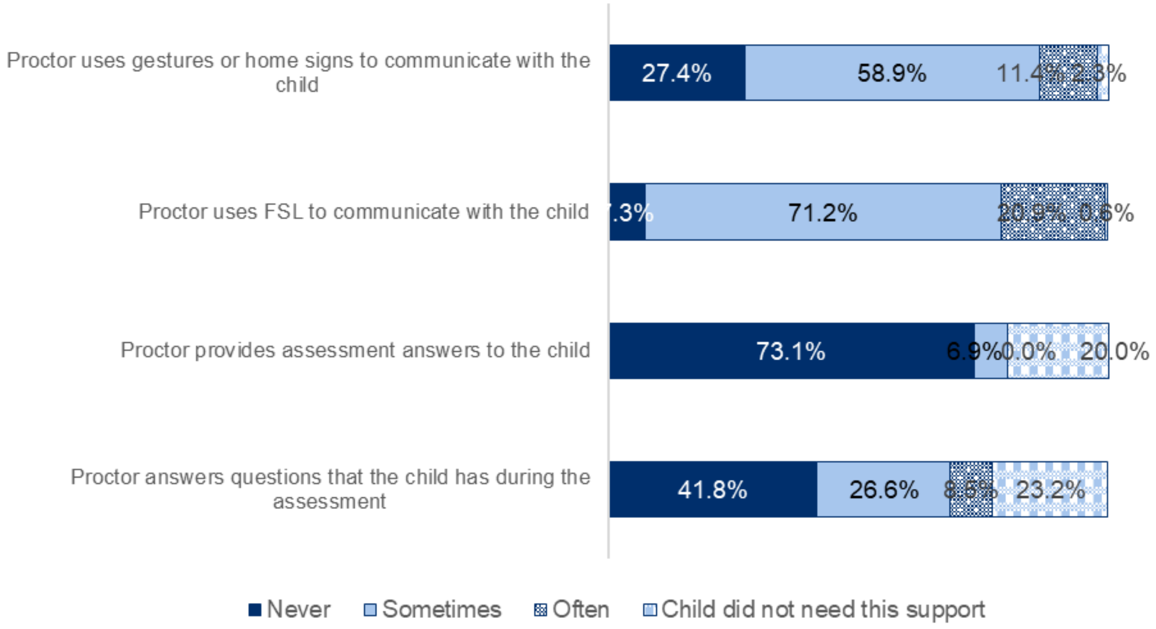
Observers also agreed with proctors that they were able to provide technical support to learners during the assessment. Ninety percent of observers agreed that proctors provided support during the assessment that allowed the child to proceed with the assessment. And nearly all observers agreed, except two, that proctors appeared to understand how to operate the tablet and Tangerine:Learn.

Figure 29. Observer Data on Proctor Behavior Frequencies: Technological Assistance by Proctor



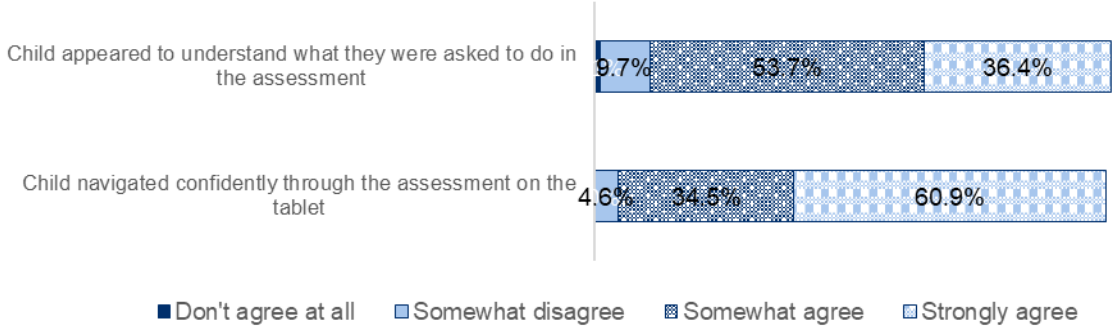
In addition to providing technological assistance, observers noted that proctors played a pivotal role in communicating with learners during the assessment. However, this experience was not universal and should be formalized in future assessments. Proctors were observed using both FSL (91.2 percent) and gestures or home signs to communicate with the learner (70.3). Observers also noted that while not common, some were observed providing learners with answers during the assessment (6.9 percent) (**Figure 30**). It is recommended that clearer communication guidelines be communicated with proctors that would provide a more uniform experiencer for learners, ensuring the validity of assessment results.

Figure 30. Observer Data on Proctor Behavior Frequencies: Communication with Learners



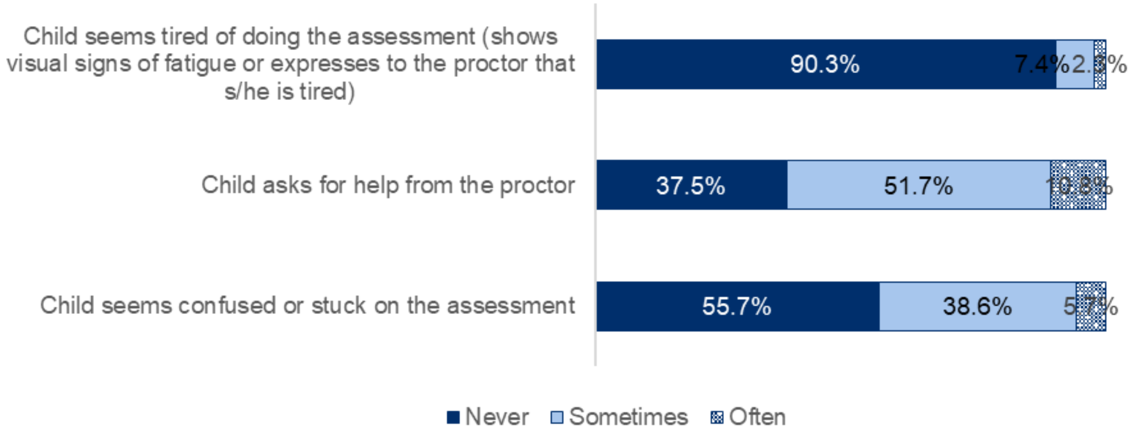
Observers also provided valuable insight on the experiences of learners during the assessment, ultimately suggesting learners potentially need more support from proctors, more exposure to information and communications technology (ICT) and tablets, and that language might still be a barrier for these learners. The majority of observers agreed that learners navigated confidently through the assessment (95.4 percent) and that they understood what they were asked to do during the assessment (90.1 percent) (**Figure 31**). While these percentages are high, they do not represent all learners, indicating that some learners did need more support during the assessment. We must also note that there is the possibility that learners did not ask proctors for assistance even when they needed it. Qualitative open-ended responses underscore that learners needed more exposure to ICT and the tablets. A common theme coming across open-ended responses highlights student confusion. But more than that, observers highlighted that as the assessment progressed the learner “gained confidence during the assessment.”[14]

Figure 31. Observer Agreement with Statements on Learners’ Experience



Furthering the point that not all learners were equally successful in navigating the assessment, 44.2 percent of observers indicated that a learner seemed confused or stuck. This noted confusion could be the result of understanding the assessment tool, but also could be the subtasks on the learning assessment questions themselves (**Figure 32**). Despite the confusion, at least one-third of observers stated that a child never asked for help from a proctor (37.5 percent). Further, more than 10.0 percent of observers witnessed learner behavior that suggested learners were tired. In order to fully understand the reason behind the high levels of noted confusion, future assessments should include a more rigorous investigation of learner experiences and whether difficulties came from the assessment tool or content.

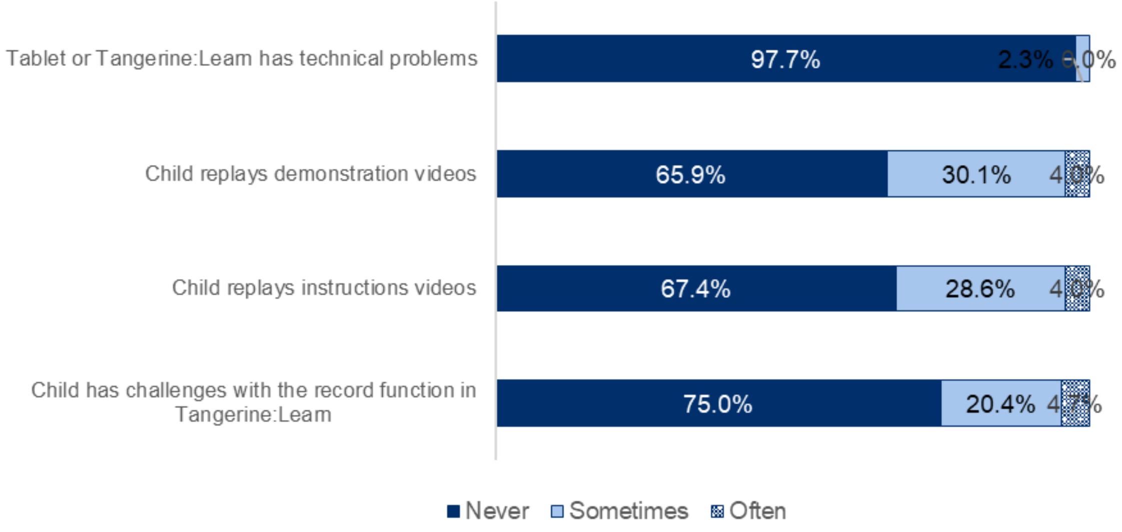
Figure 32. Observer Data on Proctor Behavior Frequencies: Learners



[14] Response from open-ended observer feedback form.

Students had notable difficulties with technological aspects of the assessment and often had to replay instructions or demonstrations. **Figure 33** shows 25.1 percent of observers saw students having sometimes or often having difficulties with the record function in Tangerine:Learn. Students also had to sometimes or often replay instruction videos (32.6 percent) and demonstration videos (34.1 percent). These difficulties might be based on the content delivered rather than with the tablet itself as only 4 observers noted technical problems with the tablet (**Figure 33**). Therefore, this is notable but not an outright impediment of asynchronous assessments, especially given that the learners in this sample had little to no access to tablets in school or out of school. Populations with greater exposure to tablets will likely not experience similar difficulties. In cases like this, greater exposure pre-assessment should be considered.

Figure 33. Observer Data on Proctor Behavior Frequencies: Learners' Experience with Assessment Medium



Other Notable Conclusions

While not specific to any particular measurement tool, it is important to note the overall positive reception participants had to the asynchronous assessment. Many students were eager and excited to participate, even when faced with technological learning curves. Proctors echoed their students' enthusiasm in their observations of their students' engagement with the application. One proctor reported that she saw her students attempting all questions because they enjoyed interacting with the application and watching the videos with FSL. In other observations, an FSL interpreter shared that he saw children's eyes "light up" when seeing the FSL signing in the videos.

Proctors themselves also shared enthusiasm about the assessment and the application's potential uses both as an assessment and as a teaching tool. One proctor reported that they were able to learn FSL by simply watching the FSL instruction videos.

Limitations

It is important to note some limitations to the beta test and its findings. These limitations should be kept in mind when considering the generalizability of the conclusions and recommendations as well as extensions of the design.

- Participants were not randomized into testing form groups, which limits the comparability of student performance between receptive and expressive subtask types.
- We were not able to score responses from Cavite Form 2 due to logistical difficulties surrounding scoring assignments and procedures.
- For the scope of the beta testing, heterogeneous effects driven by demographic characteristics of the participants were not investigated.

CONCLUSIONS AND RECOMMENDATIONS

The pre-test, alpha test, and beta test conducted by RTI, STS, and RBI tested asynchronous assessment for students who are deaf or hard of hearing. The results are highly informative in answering the research questions posed. This conclusion section will respond to them directly.

Research Question 1: Which subtasks from existing EGRAs for students who are deaf or hard of hearing (like the USAID Gabay EGRA) allow for asynchronous administration? Which, if any, subtasks that are not part of the existing EGRA could be considered?

All six subtasks tested adapted successfully to the asynchronous EGRA format. The subtasks were developed for use in the USAID Gabay EGRA for the USAID/Philippines Gabay (Guide): Strengthening Inclusive Education for Blind/Deaf Children project. These were: Receptive Vocabulary, Expressive Vocabulary, Sign Language Comprehension (Level 1), Letter Name Identification, Familiar Word Reading, and Sentence Reading Comprehension. Varying expressive and receptive formats of the subtasks over the course of the alpha and beta tests showed that subtask formats could be modified while maintaining high levels of learner interaction with the assessment.

Findings specifically from the beta test show that learners across both forms performed well with lower-level subtasks, and low scores across high-level subtasks reflect accurate learning levels. Subtasks where the average percent score was over 50 percent were Receptive Vocabulary, Expressive Vocabulary, and Letter Name. Students performed slightly poorer on Familiar Word subtasks with an average percent score between 41percent and 49 percent. These results suggest that these subtasks were appropriate for the learners who are deaf or hard of hearing through asynchronous administration. In comparison, students on average had a 20 percent score on language comprehension and sentence reading. These subtasks were more difficult, and scores reflect the reality of learning levels in this population.

There are two recommendations that come from findings related to the first research question.

Incorporating autostop protocols for receptive subtasks could mitigate IRR issues as well as assessment length. While expressive questionsrely on the use of a scorer post-assessment, receptive questions can be adjusted to have autostops built in as they do not require a scorer to review student responses. Additionally, autostops would reduce the number of items asked to the student, reducing assessment length. Skip logic, based on automated scoring of receptive subtasks, could also be incorporated to skip higher order expressive subtasks.

- Additional subtasks that were not tested in the project, including Fingerspelling Reproduction and Sign Language Comprehension (Level 2), could be included in future assessments. However, this would make the assessment longer.

Research Question 2: What type of asynchronous administration is operationally feasible, technically rigorous, and suited to the context of Deaf education in the Philippines?

Both the pre-test and the alpha test largely established the importance of proctors. Beta confirmed this, as proctors provided support in over 88 percent of observations. It is unlikely that this assessment could be self-administered without the support of a proctor on-site. Non-fluent in-person proctors were judged to be the best option for a scalable and feasible assessment model, given the limited availability and quantity of FSL-fluent proctors in this context. Further, non-fluent proctors proved to be just as effective as fluent proctors.

The online help desk proved less effective. Having FSL-fluent support present through an online help desk also requires high levels of resources, including on-site ICT support and stable Internet, both of which cannot be guaranteed in a school setting. Further, during the alpha test, there was a low level of student use of the help desk, and it proved disruptive to the student when the connection with the help desk was lost and had to be reset.

Unique findings from the beta test illuminate both the benefits and limitations to expressive and receptive question modalities. In terms of scoring, receptive questions lend themselves far better to consistency in scoring across students' assessments. The same standards by design—i.e., the automatic scoring inherent in this form—were applied to all answers. However, receptive questions can lead to students' lucky guessing, falsely inflating the scores. The validity of expressive assessments rests heavily on the scoring protocols developed and explained in scorer training.

Ultimately, the results across project phases suggest two recommendations:

- Proctors are necessary and assessment likely cannot be successfully self-administered without them. However, proctors can have low levels of FSL fluency as long as they receive sufficient training in how to proctor the EGRA, specifically experience with the application. The main functions they performed, like technological assistance, do not require high levels of fluency.
- If testing with expressive modalities, rigorous scoring protocols must be developed and trained across scorers.

Research Question 3: What are appropriate protocols for asynchronously administered subtasks? How do these diverge from protocols of the in-person administration of these subtasks? Protocols should consider preferred media platforms, suitable locations, length of testing, assess-ee identity, and data privacy, among other things.

The in-person synchronous EGRA is highly dependent on the "live" assessor's interaction and engagement with the learner. In contrast, the asynchronous administration with recorded instruction and demonstration videos is not. The latter encourages more autonomy of the learner. However, with an on-site proctor, the student still receives in-person support and guidance throughout the assessment, despite receiving instructions through the recorded videos. A notable distinction between in-person administration and asynchronous is the ability of the assessor to modify instructions or add examples—as allowed by EGRA guidelines—to assist the student in understanding the assessment. This may be difficult for an in-person proctor to do unless the proctor possesses FSL fluency to provide this support. However, as seen in the beta test, proctors with low levels of FSL were still able to provide sufficient support to students to complete the assessment.

As mentioned previously, proctors are necessary but do not necessarily need to be fluent. In fact, it is potentially more important that they be from the area or schools sampled. Having a proctor from the area or from the school could assist with knowledge of the regional FSL—even if the proctor has low levels of FSL, student familiarity and comfort with the proctor may reduce any anxiety during the assessment, and travel costs or any supplemental labor costs.

This project serves a definite proof of concept for the adaptation and utilization of the assessment on tablets—selected for this project due to their screen size and portability, which was a recommendation provided through consultations and landscape review. Further, Tangerine:Learn worked well for displaying, capturing, and storing videos in asynchronous scenarios. Regarding technology function, very few problems occurred. Videos loaded and played well, and the camera captured enough detail to be scored. Uploading videos to the server at the end of the day also worked well, given adequate bandwidth.

However, students' familiarity with tablets, or lack thereof in this case, did affect their level of comfort and confidence during the exam. During beta testing, about 25 percent of learners had challenges with the record function, which seems important to highlight, since this is critical for the assessment. Additionally, 44 percent of observations said that learner seemed confused or stuck sometimes or often. Despite the difficulties, students generally enjoyed the assessment. Therefore, tablets and Tangerine:Learn combined should not be a problem in the Philippines.

Common or simple classrooms should not pose any problem for both administrations, face to face or asynchronous. However, both need less visual clutter or people passing by, so as not to disrupt students' attention to the test. It is also important to ensure that the testing room does not have the sign language alphabet or any sign language of words or numbers posted on the wall, which may be common to classrooms for students who are deaf or hard of hearing. In the asynchronous administration, background of the child is very important because this may impact the scorer's ability to review and score the child's response for expressive subtasks.

Regarding the length of the assessment, the asynchronous EGRAs at beta test consumed an average of 30 minutes to 45 minutes. Analysis of receptive and expressive subtasks showed that receptive subtasks could potentially reduce assessment length in an asynchronous assessment. However, further research should be conducted to statistically analyze whether the receptive question modality affects test performance.

An important note about asynchronous evaluation and privacy, the assessment does require the recording of the child's face for expressive subtasks. For some videos, students were wearing face masks under the pandemic protocol. With face masks, the child's identity is protected with most of the student's face covered. However, Deaf mentors and FSL interpreters voiced that facial expressions were important to scoring, so for the comprehension questions, students removed their face coverings.

These videos were kept securely on the Tangerine:Learn server. However, scorers were granted access to this server when scoring. Other users with server credentials could potentially gain access to these videos. With regards to assess-ee identity and data privacy, server credentials should be shared only with individuals as necessary and download permissions should limited.

Protocols and best practices for data storage and management should be well defined and institutionalized. Further use of the asynchronous assessment should explore methods to improve how scorers access student response videos and document scores.

Ultimately, the results across project phases suggest three recommendations:

- Students should be introduced to the tablets and Tangerine:Learn before the assessment if the project is sampling from populations with lower levels of exposure to this type of technology. In order for this to be an appropriate type of assessment modality, USAID and DepEd should work to get learners more exposure to technology.
- Assessments should be held in classrooms that provide distraction-free environments with neutral backgrounds for video capturing.
- Scoring protocols should be examined further for areas of automation, specifically to address how scorers access data and provide scores for videos.

Research Question 4: Which factors are the most determinant drivers of the cost? Which factors impact the efficiency and effectiveness of asynchronous administration? Is the design scalable within the Philippines beyond the proof of concept?

The costs of asynchronous assessments are largely incurred on the creation of the assessment tool and in testing equipment. Specifically, the procurement of tablets, the video productions, the application's designs and features. However, these are often one-time costs. Video production, specifically, can require a large investment in video equipment, video editing software, and labor costs for highly skilled specialists. Equipment required for instruction and demonstration videos included backdrops—important to minimize distractions, particularly important for people who sign; camera; lighting equipment; and video editing software. Production of these videos also included many hours from FSL interpreters, FSL Deaf mentors, video editors, and video production specialists. While these are one-time costs, the investment in quality equipment and specialist can impact the quality of instructions and video prompts in the assessment.

In addition to these, proctors are a necessary but added cost. The proctors were recruited from the schools' teachers, who were on salary for their time. The project also provided training incentives to compensate the time spent outside of school learning the application and how to proctor the assessment.

While an Internet connection is not required during the assessment itself, it is necessary to upload student responses and to download any updates to the application. This can be difficult to do in the schools as some do not have stable access to the Internet or in the event of natural disaster or weather-related complications when the Internet is not available. Stable bandwidth is also required at the development stage of the application and assessment, as the instruction and demonstration videos need to be uploaded on the Tangerine:Learn server and programmed into the assessment.

Receptive subtasks have large cost-saving implications as they do not require the use of scorers and can be automatically scored.

Regarding scalability, assessor agreement and scoring are likely to be challenging using a majority of expressive subtasks. Additionally, the scoring of expressive subtasks would extend how much longer it would take to scale. The variability of the IRR results across locations during beta testing suggest that scalability—i.e., an increase in scorers—would require a rigorous scoring protocol that is easily taught and implemented across multiple populations.

The success of proctor training seen in beta, a marked improvement from alpha, suggests that training is effective for helping proctors successfully facilitate assessments. Improvements in training from the alpha to beta test meant that more proctors were able to ensure that students had ‘full and proper signing space’ while doing the assessment.

Splitting the forms into two versions—Form 1 and Form 2—demonstrated the role of receptive subtasks in an asynchronous assessment. Receptive subtasks could reduce time spent on scoring with fewer items to score manually and may require less external support for the learner. Continued exploration of a remote assessment with receptive subtasks could lead to a graduated assessment form, as the scoring could be automatized. Ultimately, further exploration would be needed into receptive/expressive as the assessment itself is validated.

Recommended Next Steps

While this version of Tangerine:Learn was developed as an assessment for primary-grade students who are deaf or hard of hearing, this application could have wider applications. It is possible that this application could function as a formative assessment, allowing teachers to see more about what areas of literacy challenge their students and the linkages between various literacy skills. Tangerine:Learn would also be appropriate as a summative assessment in some contexts, but would require keen attention to the items tested and their relation to what students are learning in the classroom.

In addition to its use in assessment, Tangerine:Learn provides opportunities for student practice in the classroom, informally with the application. Proctors noted during testing that the video playback was a useful tool because students could see themselves signing—allowing for instant feedback and self-correction.

With the proof of concept established, validation of the assessment should be considered, specifically examining expressive and receptive modalities. This validation should consider a quasi-experimental approach that can account for potential confounders between student populations. The treatment design would compare receptive and expressive question modalities. With randomization and a large enough sample size, the study could control for confounding variables that could also affect test size. In doing so, researchers can statistically analyze whether the receptive question modality affects test performance.

The assessment and application could also be improved based on the learnings from the beta test. These improvements could be as follows:

- Adapting the application, assessment, and modality for students with multiple disabilities—specifically students who are deaf/hard of hearing and are also blind or have low vision. Students who are deaf or hard of hearing and who have difficulty seeing participated in the beta test. During the assessment, these students had challenges navigating the assessment because of the size of the application components on the screen.
- Developing more rigorous scoring protocols and training; improving the scoring dashboard to include input fields to capture manual scoring for expressive subtasks.
- Including contextually appropriate art for images.
- Improving layout of application—eliminating scrolling screens, increasing the video sizes and other application components, eliminating overlapping buttons, increasing font size for included text.

In summary, the learnings from the pre-test, alpha test, and beta test of the asynchronous administrated EGRA can be summarized in the following points:

1. Non-FSL-fluent proctors are effective and scalable.
2. Stronger protocols for scoring expressive tasks are needed—both in definition of scorable responses and in the process of how scorers review responses.
3. Receptive tasks can reduce the scoring challenges.
4. The length of the assessment between receptive and expressive subtasks is on average equivalent, but as FSL level of the learner increases, the time of the assessment decreases.
5. Assessment delivery through tablets and Tangerine:Learn is user friendly and scalable, but students could use additional exposure to technology.

In the context of the Philippines, this project found an enthusiastic reception to this assessment modality. With enthusiasm from both students and teachers, an adaptation of this assessment into formative assessment or informal classroom or home practice could be an approach to increase technological exposure for future national testing and also provide FSL resources to a context where this support is much needed and desired.

ANNEXES

ANNEX A. LANDSCAPE REVIEW

Question 1: What assessments and assessment modalities are currently being used for learners who are deaf or hard of hearing, within and outside of the Philippines?

The availability of assessments for children who are deaf or hard of hearing is limited. As researchers noted in an October 2018 article in the *Journal of Deaf Studies and Deaf Education,* "[L]imited information exists on what signed language assessments are available, and if those available are quality assessments" (Henner, Novogrodsky, Reis, & Hoffmeister, 2018, p. 308). Reasons for the dearth of assessments in American Sign Language (ASL) include "challenges in creating tests that can adequately account for the linguistic features of ASL, the need for examiners to be highly trained and have strong language skills, and prohibitive costs associated with purchasing standardized tests and training examiners on those tests" (Pizzo & Chilvers, 2019, p. 233).

Still, researchers have created some assessment tools to test a variety of fundamental reading skills for students who are deaf or hard of hearing—and are currently developing new approaches. Because experts feel that "there is no one assessment that can provide a comprehensive portrait of a child's language and literacy abilities," sign language assessments may take many forms and structures (Pizzo & Chilvers, 2019, p. 225). These may be formal or informal and based on multiple different approaches.

Table A-1 summarizes many of the assessments currently being utilized in the United States to assess the literacy skills in ASL of students who are deaf or hard of hearing. Although surely incomplete, the table does capture all the assessments referenced in the reviewed sources cited at the end of this review.

When considering formal assessments, J. Henner et al. identify two options for producing formal assessment: (1) adapting an existing standardized test into signed language, or (2) creating a new test. Within these categories, assessments can be categorized as either "productive," wherein the test taker produces a language sample, or "receptive," wherein the test taker responds to a stimulus. Production assessments often take the form of a checklist. They utilize parents, teachers, or professionals familiar with the child to attest if they know a particular item or word. These tests can be subject to fluency limitations, inter-rater reliability issues, and inherent biases. Receptive assessments follow a format of exposing a test taker to a stimulus and then asking them to select the correct option from a multiple-choice test. J Henner et al. (2018) found that receptive tests are more likely to be normed using "classical test theory" but are not free from biases. Either approach can be applied to adapted or original-design assessments of signed language.

Pizzo and Chilvers (2019) describe informal assessments that can be used among students who are deaf or hard of hearing, including play-based assessments, performance-based assessments, and portfolio-based assessments. Pizzo and Chilvers find portfolio-based studies to be well suited for younger children. They lend themselves to iterations over time and can be monitored remotely thanks to the increased accessibility of videos. Unlike formal assessments, informal assessments have more flexibility in how they are applied and scored. This may be beneficial when working with populations of students who are deaf or hard of hearing. Informal assessments can be more easily adapted to smaller groups—even as small as a single classroom.

To evaluate the reading skills of students who are deaf or hard of hearing, researchers in the United States have used a wide variety of assessments adapted for the population and created specifically for them. For instance, in a study of 336 students who are deaf or hard of hearing in kindergarten, first, and second grade, researchers used a battery of different assessments to measure their progress in language, reading, and phonological awareness (Antia, et al., 2020). Antia et al. used a total of seven tests: four to measure students' language skills—vocabulary, receptive English syntax, expressive spoken English syntax, and receptive ASL syntax—two to measure students' phonological awareness—including spoken proficiency assessment and fingerspelling proficiency assessment—and one test for reading.

The majority of ASL assessments measure students' basic language skills, including phonology, vocabulary, morphology, and syntax, as noted in **Table A-1**. As Boston University researchers who recently developed a new ASL comprehension assessment noted, "Despite the importance of higher-order text comprehension skills, existing ASL assessments generally focus on basic proficiency in ASL vocabulary and grammar, and there is currently no means of evaluating the more advanced skills that are necessary for ASL text comprehension"[15] (Rosenburg, Lieberman, Caselli, & Hoffmeister, 2020, p. 2). Whether measuring basic or advanced language skills, most assessments may be self-administered on a web-based platform on a computer, with ASL instructions and items delivered via video.

In the wake of the coronavirus disease 2019 (COVID-19) pandemic, researchers working with students who are deaf or hard of hearing had to adjust their means of assessment to account for students' remote learning environments. In key informant interviews (KII), several researchers shared their experiences with assessing students during the pandemic. An associate professor in the department of curriculum and instruction at the University of Connecticut is currently studying how to measure performance of students in third to sixth grade who are deaf or hard of hearing. Researchers initially assessed students in-person, but they shifted the assessments online once the pandemic began, even though researchers recognized it would "jeopardize results." The switch to online testing required trial and error and resulted in varying levels of success. Researchers tried to simulate the in-person testing experience via Zoom for a common standardized test but then switched to asynchronous administration for a motivation survey, which they quickly discovered was difficult for students to complete without in-person or virtual help from someone fluent in ASL. Ultimately, researchers settled on a hybrid approach to assessment, with students controlling videos and taking assessments asynchronously. However, students could get in contact with an adult in real time to ensure understanding and get support. Adults were trained online, and all data collectors were native ASL users or proficient.

[15] ASL text comprehension is not the same as reading comprehension, with the authors defining an ASL text as "a composition expressed in ASL that is used to communicate information to others" (Rosenburg, Lieberman, Caselli, & Hoffmeister, 2020, p. 2).

In 2020, a student outcome specialist at the California School for the Deaf conducted remote assessments in a variety of subjects of students in 2nd to 12th grade who lived near the Mexican border. Students used two devices simultaneously during the test—an iPad for taking the test itself, and a laptop with Zoom to speak with a teacher if support was needed. The specialist said such teacher support was needed because it was a “huge challenge” for students to access tests on the tablet. The testing environment also varied for students, with some students going to Starbucks or McDonald’s due to lack of Wi-Fi at home, and many students were disconnected from Zoom due to connectivity issues. The specialist said they “tested who we could and did the best we could,” but they were not necessarily considering the remote assessment data to be reliable based on all the challenges. Based on their experience, the specialist recommended one-on-one proctoring. They also suggested that students get exposure to the device to be used for the assessment ahead of time so that they know how to use them, especially so the devices are set up and operational once it is time to take the assessment.

As for students in the Philippines who are deaf or hard of hearing, Resources for the Blind, Inc. (RBI), School-to-School International (STS), and their partners developed an early grade reading and sign language assessment (EGRA) for the USAID Gabay (Guide): Strengthening Inclusive Education for Blind/Deaf Children project. In March 2020, 165 students in kindergarten to Grade 3 participated in a baseline EGRA that assessed students’ skills in Filipino Sign Language (FSL)—including receptive vocabulary, expressive vocabulary, and sign language comprehension—and English reading—including letter name identification, fingerspelling reproduction, familiar word reading, and sentence reading comprehension. STS, RBI, and partners opted for enumerators to sign test content to students live for the receptive vocabulary and language comprehension subtasks, rather than show videos of an enumerator signing the content, due to the fact that learners are “unfamiliar with testing environments” and “their nascent skills in FSL are better supported by live signing of subtasks, so they are better able to intuit context and expression” (School-to-School International, 2020, p. 6).

Table A-1. Summary of Assessments from Reviewed Sources

Name	Institution	Description	Intended population	Tech	Notes
ASSESSMENTS FOR LOWER-LEVEL SKILLS IN AMERICAN SIGN LANGUAGE					
ASL and Non-Linguistic Perspective Taking Comprehension Tests	n/a	For a single case study, David Quinto-Pozos and Lynn Hou developed a test to “assess perspective-taking skills with respect to the comprehension of classifiers within topographical space” (e.g., positional orientation) of two objects like a toy car and toy dog.	Children and adolescents (aged 7 to 20 years old)	Administered via computer	
ASL Communicative Development Inventory 2.0 (ASL-CDI 2.0)	Boston University and Wellesley College	Developed by professors at Boston University and Wellesley College, this vocabulary assessment is a recent update to the first version of the ASL Communicative Development Inventory developed about 20 years ago. It tests receptive and expressive vocabulary and includes a section about gestures and phrases.	Children 5 and younger	Online in beta form; it can be administered by someone without formal training in sign language	
ASL Online Vocabulary Exam (ASL-OVE)	Language Acquisition and Assessment Laboratory (LAA) at the Rochester Institute of Technology (RIT)	LAA director Dr. Peter C. Hauser has developed and tested this ASL proficiency test in the past year. According to the RIT website, researchers are currently writing a peer review manuscript of their work with the ASL-OVE.	n/a	n/a	This research lab is part of RIT's National Technical Institute for the Deaf Research (NDIT) Center on Culture and Language (CCL).
ASL Phonological Awareness Test (ASL-PAT)	University of Alberta	Developed by the University of Alberta's Dr. Lynn McQuarrie, this 49-item online test aims to assess ASL phonological awareness in children aged 4 to 7.	Children aged 4 to 7	n/a	

Name	Institution	Description	Intended population	Tech	Notes
ASL Proficiency Assessment (ASL-PA)	University of South Florida, University of Illinois, and University of Arizona	Researchers collaborated to design an ASL proficiency test for children aged 6 to 12. An assessor rates a child's ASL proficiency as either Level 1, 2, or 3 after watching the child's 30-minute ASL sample recorded on video.	Children aged 6 to 12	Video	
ASL Receptive Skills Test	Northern Signs Research	Available through Canada-based Northern Signs Research, the test "measures children's understanding of ASL grammar, including number/distribution, negation, non/verb distinction, spatial verbs (location and movement), size/shape specifiers, handling classifiers, role shift and conditionals." The assessment was adapted from a receptive skills test for British Sign Language.	Initial piloting was conducted with children from the ages of 3 to 14	Online	
ASL Vocabulary Test (ASL-VT)	University of Roehampton (United Kingdom) and City University London	Three researchers adapted a British Sign Language test into ASL and piloted it with 20 native ASL speakers.	n/a	Web-based	
Fingerspelling and Number Comprehension Test (FaNCT)	LAA at RIT	LAA director Dr. Peter C. Hauser has developed and tested this ASL proficiency test in the past year. According to the RIT website, researchers are currently writing a peer review manuscript of their work with FaNCT.	n/a	n/a	This research lab is part of RIT's NDIT CCL.

Name	Institution	Description	Intended population	Tech	Notes
Visual Communication and Sign Language (VCSL) Checklist	Gallaudet University	This standardized checklist assesses young children's ASL development from birth to age 5. Its purpose is to document "the developmental milestones of children from birth to age 5 who are visual learners and are acquiring sign language regardless of level of hearing. It is presented in a user-friendly format that is accessible to parents and teachers, as well as specialists and experts." It is an "observational tool used to document language in natural environments."	Birth to age 5	Checklist may be completed with paper and pencil	
ASSESSMENTS FOR HIGER-LEVEL SKILLS IN ASL					
ASL Assessment Instrument (ASLAI)	ASL Ed Center in Framingham, Massachusetts	Administered exclusively by the ASL Ed Center in Massachusetts, this computer-based test assesses students in 10 areas of ASL vocabulary and grammar. As noted in his academic profile, Boston University professor emeritus Dr. Robert Hoffmeister initially developed this instrument.	PreK to Grade 12; students aged 4 to 21; adults have also taken the assessment for research	Computer-based; students view ASL instructions on their own and answer multiple-choice questions	It is not clear how ASLAI transitioned from Hoffmeister's research to being administered by the ASL Ed Center. Information is very limited about ASLAI on the ASL Ed Center website, with only one paragraph briefly providing an overview.

Name	Institution	Description	Intended population	Tech	Notes
ASL Comprehension Test (ASL-CT)	RIT	RIT's Dr. Peter Hauser led the development of an online ASL comprehension test made up of 30 multiple-choice items that can be administered without highly trained interviewers and raters.	Pilot conducted with college-aged students, so it is not clear without further research if the test is appropriate for children or adolescents	Web-based	
American Sign Language Text Comprehension Task (ASL-CMP)	Boston University	A team of researchers recently developed a new ASL reading comprehension test by adapting three texts from two reading assessments. Children answered three literal and two inferential multiple-choice questions about each text.	Children aged 8 to 18	Self-administered on a computer	

Question 2: What technologies are used in the Philippines by people who are deaf?

Focus group participants and key stakeholders repeatedly reported that people who are deaf or hard of hearing utilize the same technologies as those in hearing communities, apart from those who rely on audiological technologies. This response is promising because it allows for a relatively wide selection of devices, applications, and software. However, it also means that people who are deaf or hard of hearing likely face the same issues of poor connectivity and limited capabilities on their devices.

Devices: Overall, only one in four households has a communal cell phone in the Philippines, according to a national information and communications technology (ICT) survey conducted by the government in 2019. In addition, while four of five respondents reported using a cell phone in the previous three months, only three in 10 said they had used a computer (Department of Information and Communication Technology, 2022). While the perceptions of the availability of cell phones, tablets, and computers for students who are deaf varied widely among respondents in KIIs and focus group discussions (FGDs), the responses seemed to mirror the national ICT survey findings that cell phone use and ownership were more prevalent than computers or tablets. Some reported that most learners had their own cell phones. Others stated that their students did not have access to a cell phone or, if they did, they must share it with their families. Only one teacher reported that students had tablets or laptops, while others in the group identified organizations that provided tablets to learners in the past. Several respondents shared that teachers generally had access to laptops. The Department of Education in Manila provided laptops to some teachers in response to COVID-19 restrictions and virtual learning. According to the Gabay Assessment of Distance Learning Delivery Modalities (DLDM) Report (2021), the absence of laptops, computers, and mobile phones was cited as the primary technology-related issue.

Software and Applications: Zoom (with and without annotations), PowerPoint and Slido, PDFs, Facebook, and Google classroom were all mentioned as examples of technologies teachers have used to engage students who are deaf or hard of hearing in remote learning over the past year. Most respondents shared that teachers used these technologies to enable modular lessons for asynchronous learning—although results varied. Respondents reported that students were generally comfortable using Google and other social media platforms; however, not all students could access the lessons regardless of how they were presented due to the lack of Internet connectivity and devices. Teachers of students in urban areas and middle grades reported higher confidence levels in their students' fluency and capabilities using software and appliances.

Apart from technologies used for virtual lessons, respondents also named several social media sites and apps popular with the deaf community. Among these were three video messaging services: Glide, MarcoPolo, and Line. One respondent also spoke favorably of RIT's WorldAroundYou learning platform in the Philippines.

Internet Signal: Unstable Internet access was reported by almost all respondents as a considerable challenge facing students who are deaf or hard of hearing, in terms of remote learning and assessments. Scholarship in the area has also identified a “lack of Internet signal” as one of the most significant technological challenges facing students who are deaf or hard of hearing (Gabay, Resources for the Blind Inc., 2021). According to KIIs and FGDs, the Internet is the strongest and most reliable in metro Manila. Rural and mountainous areas face a lack of signal. Respondents felt that most students' households in these areas did not have Internet access and instead relied on television and radio. Rural regions along the coast face additional issues as many households lost electricity in recent typhoons.

Assistive Technologies: Apart from general technologies, students who are deaf or hard of hearing also use and need a variety of assistive technologies to thrive in the classroom. Unfortunately, current local government services for children with disabilities are "sparse, isolated and disjointed," according to a 2018 policy brief put forward by the United Nations Children’s Fund (Taparan, 2018, p. 1). The Gabay Assessment of DLDM Report (2021) also advises that learners should be provided with more appropriate gadgets—and a greater number of them—in order to participate in lessons. In particular, electronic sign language dictionaries, hearing aids, and tablets were recommended for students, while Wi-Fi and printers were identified as needs for teachers

REVIEWED SOURCES

Agbon, A. D., & Mina, C. D. (2017). *School Participation of Children with Disability: The Case of San Remigio and Mandaue City, Cebu, Philippines.* Philippine Institute for Development Studies.

Angrist, N., Bergman, P., Evans, D. K., Hares, S., Jukes, M. C., & Letsomo, T. (2020). *Principles for Phone-Based Assessments of Learning.* Center for Global Development.

Antia, S. D., Lederberg, A. R., Easterbrooks, S., Schick, B., Branum-Martin, L., Connor, C. M., & Webb, M.-Y. (2020). Language and Reading Progress of Young Deaf and Hard-of-Hearing Children. *The Journal of Deaf Studies and Deaf Education*, 334–350.

Broekhof, E., Bos, M. G., Camodeca, M., & Rieffe, C. (2018). Longitudinal Associations Between Bullying and Emotions in Deaf and Hard of Hearing Adolescents. *Journal of Deaf Studies and Deaf Education*, 17-27.

Bustos, M. T. (2008). Exploring Emergent Literacy Behavior of Filipino Deaf Children. *The Asia-Pacific Education Researcher* , 149-164.

Corina, D. P., Hafer, S., & Welch, K. (2014). Phonological Awareness for American Sign Language. *Journal of Deaf Studies and Deaf Education*, 530-545.

Data Reportal. (2021, February 10). *Digital 2021 Philippines (January 2021)*. Retrieved January 2022, from https://www.slideshare.net/DataReportal/digital-2021-philippines-january-2021-v01

Day, L. A., Adams Costa, E. B., & Raiford, S. E. (2015). *Testing Children Who are Deaf or Hard of Hearing.* NCS Pearson.

Department for Information and Communication Technology. (2022). 2019 Survey on Information and Communication Technology. Republic of the Philippines. Retrieved from https://psa.gov.ph/content/2019-survey-information-and-communication-technology-sict-non-core-ict-industries

Education for All 2015 National Review Report: Philippines. (2015) World Education Forum/UNESCO.

Estella, P., & Loffelholz, M. (2020). *Media Landscapes: Philippines*. Retrieved January 2022, from https://medialandscapes.org/country/philippines

Gabay, Resources for the Blind Inc. (2021). *Assessment of Distance Learning Delivery Modalities (DLDM) Report.* USAID & RBI.

Hauser, P. C., Paludneviciene, R., Riddle, W., Kurz, K. B., Emmorey, K., & Contreras, J. (2016). American Sign Language Comprehension Test: A Tool for Sign Language Researchers. *Journal of Deaf Studies and Deaf Education*, 64-69.

Henner, J., Novogrodsky, R., Reis, J., & Hoffmeister, R. (2018). Recent Issues in the Use of Signed Language Assessments for Diagnosis of Language Disorders in Signing Deaf and Hard of Hearing Children. *Journal of Deaf Studies and Deaf Education*, 307-316.

Holmer, E. (2016). *Signs for Developing Reading: Sign Language and Reading Development in Deaf and Hard-of-Hearing Children.* Linköping: Linköpings University, Department of Behavioural Sciences and Learning.

Kemp, S. (2022, January). *Digital 2022: Global Overview Report*. Retrieved January 2022, from https://datareportal.com/reports/digital-2022-global-overview-report

Luft, P. (2018). Reading Comprehension and Phonics Research: Review of Correlational Analyses with Deaf and Hard-of-Hearing Students. *Journal of Deaf Studies and Deaf Education*, 148-163.

Muega, M. A. (2016). Inclusive Education in the Philippines: Through the Eyes of Teachers, Administrators, and Parents of Children with Special Needs. *Social Science Diliman*, 5-28.

Ocampo, D. S. (2017). *Mother Tongue Based-Multilingual Education in the Philippines.* Presentation.

Pizzo, L., & Chilvers, A. (2019). Assessment of Language and Literacy in Children Who Are d/Deaf and Hard of Hearing. *Education Sciences*, 223.

RIDBC Thomas Pattison School. (2018). *The development of an Online Assessment Tool for Auslan.* Royal Institute for Deaf and Blind Children.

Rosenburg, P., Lieberman, A. M., Caselli, N., & Hoffmeister, R. (2020, May 12). The Development and Evaluation of a New ASL Text Comprehension Task. *Frontiers in Communication, 5*, 1-12.

RTI International. (2018). *All Children Reading-Asia Desk Review Activity 2.7.* USAID.

School-to-School International. (2020). *USAID Gabay: Strengthening Inclusive Education fro Blind/Deaf Children: Pilot and Baseline Training Report.* Washington, DC: USAID.

Social Impact, Inc. (2017). *Basa Pilipinas Cost Analysis.* USAID/Philippines.

Taparan, S. (2018). *Children with Disabilities: Finding the Way to an Inclusive Service Framework.* UNICEF.

ANNEX B. ALPHA TEST ROLES AND RESPONSIBILITIES

Role	Responsibility: Before (each) Assessment	Responsibility: During Assessment	Responsibility: After Assessment
Proctor (teacher)	• Find an appropriate space for the assessment (limited distractions, good natural light)[16] • Arrange desk and chair for child so that tablet can be at child eye level[17] • Arrange the desk and chair so that that video on the tablet can be captured clearly (any light should be facing child, not behind or on the side of the child)[18] • Ask child for permission to video record the assessment • For scenario 2: ensure that the online help desk tablet is functioning and ready • Note start time of assessment on the registration form	• Introduction script: **"Hello, my name is [NAME]. You're here today to help us test out a new game for children who are deaf. This is not a test – it's just practice. You will use this tablet to play the game. Do you have questions? Let's get started!"** • When child sits down, show them the tablet. Make sure the child is comfortable and the tablet is at eye level to the child. If they are ready, press the assessment button to start. • Note the child's unique Tangerine ID on the registration form • **For scenario 2**: introduce the online help desk person to the child by pointing to the help desk tablet • If the child is stuck on a page and isn't sure how to move on, press "don't know" button and "next" button to move the child to the next task • If child cannot operate the record function, show the child how to	• Tell the child thank you and good job and direct the child to the observer for the feedback survey • Note end time of assessment on the registration form • Respond to observer feedback survey

[16] This should happen once at the start of the day, but the proctors should check and adjust the assessment conditions if needed between assessments

[17] This should happen once at the start of the day, but the proctors should check and adjust the assessment conditions if needed between assessments; for example, in between each assessment, the proctor should adjust the height of the tablet based on the height of the child.

[18] This should happen once at the start of the day, but the proctors should check and adjust the assessment conditions if needed between assessments; for example, the proctor should rearrange the learner setup if the light has changed based on time of day

Role	Responsibility: Before (each) Assessment	Responsibility: During Assessment	Responsibility: After Assessment
		press the "record" and "stop" buttons. If the child does not know the answer, press "don't know" button and "next" button to move the child to the next task. • Encourage the child to stay seated and continue with the assessment • If child is unable to interact with the assessment or is persistent in wanting to end the assessment, terminate the assessment • Answer any questions the child has during the assessment; do not provide any assessment answers	
Online help desk support (scenario 2 only)		• When the proctor introduces the child, Sign the introduction script to the child: **"Hello my name is [NAME]. What is your name?...Like your teacher told you, you will be playing this game on your tablet. I will be here to answer any questions you have about the game or anything else. If you have a question, please ask me at any time during the game."** • Answer any questions the child has during the assessment; do not provide any assessment answers	• Provide qualitative feedback on the child's level of understanding of sign language
Observers + sign language interpreter[19]	• At beginning of day: coordinate with head teacher and classroom	• Fill out observation checklist	• Administer student feedback survey

[19] The observers and sign language interpreters will not directly interact with the student, proctor, or online help desk support during the administration of the assessment. Their roles will be to provide insight in the student, proctor, and online help desk support's involvement in the assessment.

Role	Responsibility: Before (each) Assessment	Responsibility: During Assessment	Responsibility: After Assessment
	teacher; get a list of the children who will take part in the assessment and their basic demographic information (grade, age) and enter on the registration form • Ensure that tablet has sufficient memory to capture videos[20] • Sit to the side of the child and proctor with sufficient space to not interrupt; but should be able to see both the child's interaction with the tablet and see the proctor		• Accompany child back to their classroom and bring the next child for the assessment • Administer proctor survey • Collect feedback from online help desk support • At end of day: sync Tangerine:Learn data, and keep track of amount of time it takes to synch

[20] Should be done prior to each assessment

ANNEX C. ALPHA TEST OBSERVER CHECKLIST, LEARNER FEEDBACK, AND PROCTOR FEEDBACK FORM

REMOTE EGRA FOR LEARNERS WHO ARE DEAF OR HARD-OF-HEARING
ALPHA TEST
OBSERVER CHECKLIST

Section I.

Please fill in the following details for this observation.

a. Date (dd/mm/yyyy)	__ __/ __ __/ __ __ __ __		
b. Observer name:	____________________	c. Interpreter name:	____________________
d. Proctor's name:	____________________	e. Region:	____________________
f. School name:	____________________	g. Child's sex:	F / M
h. Child's grade:	K / G1 / G2 / G3 / G4 / G5 / G6 / Non graded	i. Child's age:	__ __

Section II.

In the following section, please mark "Yes" with an "x" if the action took place at any point during the assessment. Mark "No" if the action did not take place at any point during the assessment. Mark "N/A" if the action is not applicable. Please add comments to explain your answer.

ACTION	EVIDENCE (mark with x)			COMMENTS
	Yes	No	N/A	
a. Proctor arranges desk and chair so that tablet is at child's eye level				
b. Proctor confirms that online help desk tablet is functioning and ready				
c. Proctor shows child the tablet and Tangerine:Learn application				
d. Proctor indicates to child to press first button to start assessment				
e. Proctor introduces online help desk person to child by pointing to the tablet				
f. Online help desk person introduces themselves to the child				
g. Proctor thanks the child for doing the assessment				

ACTION	EVIDENCE (mark with x)			COMMENTS
	Yes	No	N/A	
h. Child or proctor ended the assessment early				

Section III.

In the following section, please mark the frequency of an action during the assessment. Mark "Often" if the action occurred regularly during the assessment. Mark "Sometimes" if the action occurred a few times during the assessment. Mark "Never" if the action did not occur during the assessment. Mark "N/A" if the action is not applicable. Please add comments to explain your answer.

ACTION	RESPONSE (mark with X)				COMMENTS
	Often	Sometimes	Never	N/A	
a. Proctor presses "don't know" or "next" button to move child to the next task					
b. Proctor encourages child to stay seated and continue with assessment					
c. Proctor answers questions that the child has during the assessment					
d. Proctor provides assessment answers to the child					
e. Proctor uses FSL to communicate with the child					
f. Child seems confused or stuck on the assessment					
g. Child asks for help from the proctor					
h. Child asks for help from the online help desk					
i. Child needs proctor's support to navigate through Tangerine:Learn					
j. Child has challenges with the record function in Tangerine:Learn					

ACTION	RESPONSE (mark with X)				COMMENTS
	Often	Sometimes	Never	N/A	
k. Child replays instructions videos					
l. Child replays demonstration videos					
m. Child needs encouragement to continue the assessment					
n. Online help desk person answers questions that the child has during the assessment					
o. Proctor encourages the child during the assessment					
p. Tablet or Tangerine:Learn has technical problems					

Section IV.

In the following section, please rate your general observations from the assessment.

ACTION	RESPONSE (mark with X)					COMMENTS
	Strongly agree	Somewhat agree	Somewhat disagree	Don't agree at all	N/A	
a. Child navigated confidently through the assessment on the tablet						
b. Child appeared to understand what they were asked to do in the assessment						
c. Proctor provided useful support to the child during the assessment						
d. Proctor appeared to understand how to operate the tablet and Tangerine:Learn						
e. Online help desk person provided useful support to the						

ACTION	RESPONSE (mark with X)					COMMENTS
	Strongly agree	Somewhat agree	Somewhat disagree	Don’t agree at all	N/A	
child during the assessment						
f. Child demonstrated their FSL and English reading skills during the assessment						
g. Child needed encouragement to continue the assessment						

Section V.

Please provide any general comments or feedback about this observation that you would like to share.

Notes

Please use this space to make notes during the observation. These notes should help you fill out the checklist after the observation.

REMOTE EGRA FOR LEARNERS WHO ARE DEAF OR HARD-OF-HEARING
ALPHA TEST
STUDENT FEEDBACK SURVEY

Section I.

Please fill in the following details for this observation.

a. Date (dd/mm/yyyy)	__ __ / __ __ / __ __ __ __		
b. Observer name:	______________________	c. Interpreter name:	______________________
d. Proctor's name:	______________________	e. Region:	______________________
f. School name:	______________________	g. Child's sex:	F / M
h. Child's grade:	K / G1 / G2 / G3 / G4 / G5 / G6 / Non graded	i. Child's age:	__ __

Section II.

In the following section, please place an "x" in the appropriate response category per the learner's feedback. Please add additional comments if applicable. This section should take approximately 10 minutes to administer.

CONSENT		
Thank you for playing the game! I want to talk with you for a few minutes about the game. You do not have to answer these questions and can go back to class if you would like. Would you like to talk with me about the game?	Yes	____
	No	____

QUESTION	RESPONSE		COMMENTS
Q1. How much did you like the game?	A lot	_____	
	A little	_____	
	Not at all	_____	
	No response	_____	
Q2. Was the game fun?	Yes	_____	
	No	_____	
	No response	_____	
Q3. Did you understand the sign language in the game?	Yes	_____	
	No	_____	
	No response	_____	
Q4. When you had a question about the game, did you ask for help?	Yes	_____	
	No	_____	
	No response	_____	
Q5. How would you replay the video if you wanted to see it again?	Learner replays successfully	_____	
	Learner does not demonstrate replay	_____	
Q6. If you didn't want to answer a question, can you show me how you would skip the question?	Learner skips successfully	_____	

QUESTION	RESPONSE		COMMENTS
	Learner does not demonstrate skip	_____	
Q7. Would you like to play this game at school?	Yes	_____	
	No	_____	
	No response	_____	

Section I.

Please fill in the following details for this observation.

a. Date (dd/mm/yyyy)	__ __ / __ __ / __ __ __ __		
b. Observer name:	______________________________	c. Interpreter name:	______________________________
d. Proctor's name:	______________________________	e. Region:	______________________________
f. School name:	______________________________	g. Child's sex:	F / M
h. Child's grade:	K / G1 / G2 / G3 / G4 / G5 / G6 / Non graded	i. Child's age:	__ __

CONSENT

Thank you for proctoring the assessment. I'd like to ask you a few questions about your experience proctoring for the last child, so we can better understand their experience and your experience. This information will be confidential and anonymous; we will not use your name when sharing out your feedback. You can skip any question you'd like. This should take about 5-10 minutes. Do you consent to this survey?

Enumerator note: If the proctor consents, move to Section II. If they do not consent, end survey.

Section II.

Enumerator note: Please ask the proctor the following questions. Add any comments that they provide in addition to their response. Mark "N/A" if the question is not applicable.

QUESTION	RESPONSE		COMMENTS
How much support did you provide to the child during the assessment?	A lot	_____	
	A little	_____	
	None	_____	
What parts of Tangerine:Learn did the child need the most support on? *(mark all that apply)*	Recording their response	_____	
	Navigating through different screens on Tangerine:Learn	_____	
	Pressing buttons to select an answer	_____	
	Understanding the instructions videos	_____	
	Other:	_____	
Please describe what type of support they needed and what support you provided:			
How frequently did you have to encourage the child to stay seated and continue with the assessment?	Many times	_____	
	A few times	_____	
	Once	_____	

QUESTION	RESPONSE		COMMENTS
	Never	_____	
How frequently did the child ask you questions during the assessment?	Many times	_____	
	A few times	_____	
	Once	_____	
	Never	_____	
Please describe what types of questions the child asked you:			
How frequently did the child interact with the online help desk?	Many times	_____	
	A few times	_____	
	Once	_____	
	Never	_____	
	Not applicable	_____	

Section III.

Now I will read you a few statements, and I want you to tell me whether you agree, somewhat agree, somewhat disagree, or don’t agree at all with them.

ACTION	RESPONSE (mark with X)					COMMENTS
	Strongly agree	Somewhat agree	Somewhat disagree	Don’t agree at all	N/A	
a. Child navigated confidently through the assessment on the tablet						
b. Child appeared to understand what they were asked to do in the assessment						
c. I provided useful support to the child during the assessment						
d. I understand how to operate the tablet and Tangerine:Learn						
e. Online help desk person provided useful support to the child during the assessment						
f. Child demonstrated their FSL and English reading skills during the assessment						

ANNEX D. ALPHA TEST SCORING FEEDBACK FORM

Remote EGRA for Students Who Are Deaf or Hard of Hearing

Alpha test Scoring Feedback Form

Questions:

1. How long does it take you to review each assessment?

2. What issues are present in scoring the videos? Are there any issues with the quality of the videos or the size of the videos?

3. Did you have difficulty accessing the videos?

4. Did you have difficulty navigating the Excel file?

5. Did you encounter any other issues in this exercise?

6. For this alpha test, only 10 students were assessed per school. For beta test, we will be assessing over 100 students in total. Do you have any suggestions for improvements in this scoring process?

ANNEX E. ALPHA TEST DEBRIEF

ALPHA TEST DEBRIEF

REMOTE EGRA FOR STUDENTS WHO ARE DEAF OR HARD OF HEARING

June 3, 2022

Goals

- Debrief the experiences of observers, proctors, and the helpdesk support person across all three scenarios of the alpha test
- Gather feedback on user experience, learner engagement, and assessment modalities to guide design of beta version of Tangerine:Learn and beta testing

Discussion Questions

1. From your observations, what were difficulties that learners encountered with the assessment?

2. Proctors, how well were you able to support the child to do the assessment? What were the factors that made it easier or harder for you to support the learner? *(E.g., FSL ability, training, familiarity with Tangerine or the tablet)*

3. Scenario 2: Describe your experience with the help desk
   a. When did the learners engage with the help desk and when with the proctor?
   b. Help desk support, how well were you able to offer quality support over zoom? Why or why not?

4. Given your experiences during the alpha test, which scenario would you recommend we use in the next round of testing?

5. What changes could be made to Tangerine that would make the assessment more accessible for learners? *(e.g., navigation, instruction videos, length, etc.)* What changes could be made to the presentation of specific subtasks?

6. What other recommendations do you have on the assessment conditions or application that we should consider?

7. Do you think there are any limitations with this type of administration?

8. What were the qualities of learners who were able to understand what they needed to do on the assessment and navigate Tangerine? What were the qualities of learners who needed a lot of support to understand what they needed to do and navigate Tangerine? *(E.g., FSL ability, previous exposure to technology)*

ANNEX F. BETA TEST ROLES AND RESPONSIBILITIES

REMOTE EGRA FOR STUDENTS WHO ARE DEAF OR HARD OF HEARING
BETA TEST ROLES AND RESPONSIBILITIES

PROCTOR:

Responsibility: Before (each) Assessment

- Set up assessment using checklist
- Prior to beginning assessments for the day, review roles and responsibilities, proctor feedback form, and any questions on tablet navigation with observer.

Responsibility: During Assessment

- Introduction script:
  "Hello, my name is [NAME]. You're here today to help us test out a new game for children who are deaf. This is not a test – it's just practice. You will use this tablet to play the game. Do you have questions? Let's get started!"
- When child sits down, show them the tablet. Make sure the child is comfortable and the tablet is at eye level to the child. If they are ready, press the assessment button to start.
- Note the child's unique Student ID on the teacher survey form and share with observer
- If the child is stuck on a page and isn't sure how to move on, press "don't know" button and "next" button to move the child to the next task
- If child cannot operate the record function, show the child how to press the "record" and "stop" buttons. If the child does not know the answer, press "don't know" button and "next" button to move the child to the next task.
- Encourage the child to stay seated and continue with the assessment. If child is unable to interact with the assessment or is persistent in wanting to end the assessment, terminate the assessment
- Answer any questions the child has during the assessment; do not provide any assessment answers

Responsibility: After Assessment

- Tell the child thank you and good job and direct the child to the observer for the feedback survey
- Complete Proctor Survey

OBSERVER + SIGN LANGUAGE INTERPRETER

During the administration of the assessment, the observers and sign language interpreters will not directly interact with the student or proctor. Their roles will be to provide insight in the student and proctor's involvement in the assessment and to administer the student feedback survey.

Responsibility: Before (each) Assessment
• Prior to beginning assessments at each school, review proctor roles and responsibilities, proctor feedback form, and any questions on tablet navigation with proctor. • Prior to each assessment, ensure that tablet has sufficient memory to capture videos
Responsibility: During Assessment
• Sit to the side of the child and proctor with sufficient space to not interrupt; but should be able to see both the child's interaction with the tablet and see the proctor • Fill out Observer Checklist
Responsibility: After Assessment
• Administer student feedback survey • Accompany child back to their classroom and bring the next child for the assessment • At end of day: sync Tangerine:Learn data and fill out the Tablet Memory and Uploads Tracker

ANNEX G. BETA TEST OBSERVER CHECKLIST, LEARNER FEEDBACK, AND PROCTOR FEEDBACK FORM

Section I.

Please fill in the following details for this observation.

a. Date (dd/mm/yyyy)	__ __ / __ __ / __ __ __ __		
b. Observer name:	____________________	c. Interpreter name:	____________________
d. Proctor's name:	____________________	e. Region:	____________________
f. School name:	____________________	g. Child's sex:	F / M
h. Child's grade:	K / G1 / G2 / G3 / G4 / G5 / G6 / Non graded	i. Child's age:	__ __
j. Student ID:	____________________	k. Form:	Form 1 / Form 2

Section II.

In the following section, please circle "Yes" if the action took place at any point during the assessment. Circle "No" if the action did not take place at any point during the assessment. Please add comments to explain your answer.

ACTION	RESPONSE (circle response)		COMMENTS
a. Proctor arranges desk and chair so that tablet is at child's eye level	Yes	No	
b. Proctor shows child the tablet and Tangerine:Learn application	Yes	No	
c. Proctor indicates to child to press first button to start assessment	Yes	No	
d. Proctor thanks the child for doing the assessment	Yes	No	
e. Proctor ends the assessment early because of tablet malfunction	Yes	No	
f. Proctor ends the assessment early because the child continuously was unable to navigate the assessment or seemed too uncomfortable to continue	Yes	No	
g. Child refused to participate	Yes	No	

Section III.

In the following section, please mark the frequency of an action during the assessment. Circle "Often" if the action occurred regularly (four or more times) during the assessment. Circle "Sometimes" if the action occurred a few times (one to three times) during the assessment. Circle "Never" if the action did not occur during the assessment. Circle "Child did not need this support" if it seems like the child didn't need support from the proctor. Please add comments to explain your answer.

ACTION	RESPONSE (circle response)				COMMENTS
a. Proctor presses "don't know" or "next" button to move child to the next task	Often	Sometimes	Never	Child did not need this support	
b. Proctor provides other support to the child to navigate through Tangerine:Learn	Often	Sometimes	Never	Child did not need other support	Please describe support provided.
c. Proctor asks child to stay seated and continue with assessment	Often	Sometimes	Never	Child did not need this support	
d. Proctor provides general encouragement (i.e. 'good job', 'let's keep going', etc.) to the child	Often	Sometimes	Never	Child did not need this support	
e. Proctor answers questions that the child has during the assessment	Often	Sometimes	Never	Child did not ask questions	
f. Proctor provides assessment answers to the child	Often	Sometimes	Never	Child did not ask for answers	
g. Proctor uses FSL to communicate with the child	Often	Sometimes	Never	Proctor did not communicate with child	
h. Proctor uses gestures or home signs to communicate with the child	Often	Sometimes	Never	Proctor did not communicate with child	
i. Child seems confused or stuck on the assessment	Often	Sometimes	Never		

ACTION	RESPONSE (circle response)				COMMENTS
j. Child asks for help from the proctor	Often	Sometimes	Never		
k. Child has challenges with the record function in Tangerine:Learn	Often	Sometimes	Never		
l. Child replays instructions videos	Often	Sometimes	Never		
m. Child replays demonstration videos	Often	Sometimes	Never		
n. Child seems tired of doing the assessment (shows visual signs of fatigue or expresses to the proctor that s/he is tired)	Often	Sometimes	Never		
o. Tablet or Tangerine:Learn has technical problems	Often	Sometimes	Never		

Section IV.

In the following section, please rate your overall observations from the assessment.

ACTION	RESPONSE (circle response)					COMMENTS
a. Child navigated confidently through the assessment on the tablet	Strongly agree	Somewhat agree	Somewhat disagree	Don't agree at all		
b. Child appeared to understand what they were asked to do in the assessment	Strongly agree	Somewhat agree	Somewhat disagree	Don't agree at all		
c. Proctor provided support to the child during the assessment that allowed child to proceed with assessment tasks	Strongly agree	Somewhat agree	Somewhat disagree	Don't agree at all	Child did not need this support	
d. Proctor appeared to understand how to operate the tablet and Tangerine:Learn	Strongly agree	Somewhat agree	Somewhat disagree	Don't agree at all		
e. Child needed encouragement to continue the	Strongly	Somewhat	Somewhat	Don't		

ACTION	RESPONSE (circle response)					COMMENTS
assessment	agree	agree	disagree	agree at all		

Section V.

Please provide any general comments or feedback about this observation that you would like to share.

Notes

Please use this space to make notes during the observation. These notes should help you fill out the checklist after the observation.

ASSENT		
Thank you for playing the game! I want to talk with you for a few minutes about the game. You do not have to answer these questions and can go back to class if you would like. Would you like to talk with me about the game? *Enumerator note: If the student consents, move to Section II. If they do not consent, end survey.*	Yes	____
	No	____

Section II.

In the following section, please place an “x” in the appropriate response category per the learner’s feedback. Please add additional comments if applicable.

QUESTION	RESPONSE (mark with X)		COMMENTS
a. How much did you like the game?	A lot	_____	
	A little	_____	
	Not at all	_____	
	No response	_____	
b. What did you like about the game?			
c. What did you not like about the game?			
d. Did you feel tired or bored during the game?	Yes	_____	
	No	_____	
	No response	_____	
e. Was the game too short, too long, or just right?	Too short	_____	
	Too long	_____	
	Just right	_____	
	No response	_____	
f. Did you understand the sign language in the game?	Yes	_____	
	No	_____	
	No response	_____	

QUESTION	RESPONSE (mark with X)		COMMENTS
g. When you had a question about the game, did you ask for help?	Yes	_____	
	No	_____	
	No response	_____	
Thank you for answering my questions. Great job!			

Section I.

Please respond to the following questions in your capacity as proctor. Add any additional comments as necessary. Mark “N/A” if the question is not applicable.

QUESTION	RESPONSE (mark with X)		COMMENTS
a. How many times did you provide support to the child during the assessment?	Many times (four and above)	_____	
	A few times (two to three times)	_____	
	Once		
	Never	_____	
b. What parts of Tangerine:Learn did the child need the most support on? *(mark all that apply)*	Video recording their response	_____	
	Navigating through different screens on Tangerine:Learn	_____	
	Pressing buttons to select an answer	_____	
	Playing/pausing/ stopping the videos		
	Understanding the instructions videos	_____	
	Other:	_____	

QUESTION	RESPONSE (mark with X)		COMMENTS
c. Please describe what type of support they needed and what support you provided:			
d. How frequently did you have to encourage the child to stay seated and continue with the assessment?	Many times (four times and above)	_____	
	A few times (two to three times)	_____	
	Once	_____	
	Never	_____	
e. How frequently did the child express or show with visual cues that they were tired of doing the assessment?	Many times (four times and above)	_____	
	A few times (two to three times)	_____	
	Once	_____	
	Never	_____	
f. How frequently did the child ask you questions during the assessment?	Many times (four times and above)	_____	
	A few times (two to three times)	_____	
	Once	_____	
	Never	_____	
g. Please describe what types of questions the child asked you:			

Section II.

Please indicate whether you agree, somewhat agree, somewhat disagree, or don’t agree at all with the following statements.

ACTION	RESPONSE (circle response)					COMMENTS
a. Child navigated confidently through the assessment on the tablet	Strongly agree	Somewhat agree	Somewhat disagree	Don’t agree at all		
b. Child appeared to understand what they were asked to do in the assessment	Strongly agree	Somewhat agree	Somewhat disagree	Don’t agree at all		
c. I provided useful support to the child during the assessment	Strongly agree	Somewhat agree	Somewhat disagree	Don’t agree at all	Child did not ask or did not need support	
d. I understand how to operate the tablet and Tangerine:Learn	Strongly agree	Somewhat agree	Somewhat disagree	Don’t agree at all		
e. Child needed encouragement to continue the assessment	Strongly agree	Somewhat agree	Somewhat disagree	Don’t agree at all		
Thank you very much for your feedback.						

ANNEX H. BETA TEST SCORING FEEDBACK FORM

Remote EGRA for Students who are Deaf or Hard of Hearing

Beta test Scoring Feedback Form

Questions:

1. How long does it take you to review each assessment?

2. What issues are present in scoring the videos? Are there any issues with the quality of the videos or the size of the videos?

3. Did you have difficulty accessing the videos?

4. Did you have difficulty navigating the Tangerine web browser?

5. Did you encounter any other issues in this exercise?

6. Do you have any suggestions for improvements in this scoring process?

ANNEX I. BETA TEST DEBRIEF

BETA TEST DEBRIEF

REMOTE EGRA FOR STUDENTS WHO ARE DEAF OR HARD OF HEARING

October 1, 2022

Goals

- Debrief the experiences of observers, proctors, and interpreters
- Gather feedback on user experience, learner engagement, and assessment modalities for future use of remote assessments or tablet-based learning

Discussion Questions

General observations:

9. From your observations, what worked well during the assessments? Is there any positive feedback on the assessments that you would like to share?

10. Proctors – how did you support your students during the assessment? Did you feel like you were able to support your students well?

11. Different levels of FSL of proctors? How impacted assessment?

12. Proctors – what were factors that made it easier or harder for you to support your students (e.g., FSL ability, training, familiarity with tablets or Tangerine)?

13. Proctors – would additional training help you in providing support to the student? And if so, what would you like training on?

14. From your observations, what were difficulties that learners encountered with the assessment?

15. What are factors that may have impacted or hindered the student's ability to understand the assessment and navigate Tangerine? For example, with the pandemic, many students were not able to attend school and therefore, may have not had many interactions in FSL while out of school.

16. What are any limitations with this type of administration?

17. How feasible would it be to implement this type of assessment throughout the Philippines? (e.g., internet connection, tablets)

18. What are other ways that this type of this type of tablet-based learning could be utilized/useful for children who are Deaf or hard of hearing and their SPED/HI teachers in the Philippines?

19. What changes could be made to Tangerine that would make the assessment more accessible for learners? (e.g., navigation, instruction videos, length) What changes could be made to the presentation of specific subtasks?

20. What other recommendations do you have on the assessment conditions or application that we should consider?