2024 LEAP Challenge

LEAP

Project Host:

CEPE Di Tella

cepe
Evaluación de Políticas basadas en la Evidencia

Fellows:

Kaja Jasińska, Research Fellow

Rene Kizilcec, Research Fellow

Matt Lloyd-Rose, Team Lead, Social Entrepreneur Fellow

Swetha Prakash, Social Entrepreneur Fellow

TABLE OF CONTENTS

Executive Summary

Introduction

DECILE is a language evaluation tool, created by CEPE at the Di Tella university in Buenos Aires. It assesses the language skills of Spanish-speaking children with a hearing impairment aged 3 to 12 and is also suitable for neurodiverse populations.

By evaluating children's language attainment through simple tasks, it aims to provide educators with individual student metrics so they can make informed decisions about students' pedagogical paths. Still in the pilot phase, DECILE has been evaluated with over 200 children with and without hearing impairment in Buenos Aires.

Although English language assessment tools exist, there is currently no tool designed to test Spanish linguistic structures and dialects. DECILE aims to become the gold standard tool for assessing the language skills of Spanish-speaking children.

The DECILE team at CEPE Di Tella are engaging in the LEAP project to strengthen their solution and prepare to scale it.

Organisation's role & strength

Based at the School of Government at Universidad Torcuato Di Tella, the Center for Evaluation of Policies and Evidence-Based Policy-Making (CEPE Di Tella) specializes in conducting applied research to improve the quality of public policies through the generation and use of rigorous evidence.

CEPE Di Tella's Education and Behavioral Sciences Program focuses on the analysis of educational policies using social, pedagogical, economic, behavioral, and neuroscientific determinants of learning, with a particular emphasis on promoting equity and expanding opportunities for children from underserved populations.

Since 2017, CEPE Di Tella has collaborated with schools in Argentina to support the education of Deaf and Hard of Hearing (DHoH) children. Within this line of work, DECILE plays a central role as an innovative tool for assessing language development in Spanish-speaking children, addressing a critical gap in available assessment instruments for this population.

DECILE enables large-scale, evidence-based monitoring of linguistic abilities while collecting key demographic and environmental variables that influence language development. The tool is designed with scalability in mind, aiming to inform educational practices and policy at a national and regional level, while providing personalized recommendations for educators, therapists, and families.

The development of DECILE reflects CEPE Di Tella's interdisciplinary approach, combining expertise from education, behavioral sciences, linguistics, cognitive neuroscience, and public policy to advance both scientific understanding and practical interventions that foster inclusive, high-quality education.

Need summary

The DECILE team identified four priority needs for the LEAP project:

1. **Limited Validation & Benchmarks:** DECILE has been tested with a small user base, making its reliability and validity uncertain, especially in the absence of benchmarks for Spanish-speaking children.

2. **Scalability & Adoption Challenges:** While the team aspires to scale DECILE, there is no clear strategy for broader adoption, and its current design does not support independent use in classrooms.

3. **Low Practical Value for Teachers/ Students:** DECILE lacks a clear in-class usage protocol and does not yet provide meaningful insights or impact for teachers and students.

4. **Technical & Implementation Barriers:** The platform's current capabilities do not support large-scale, independent use by educators, limiting its effectiveness beyond a research setting.

Solution summary & next steps

To address these needs, the LEAP project has created four corresponding deliverables.

1. **A literature review of similar language assessments, and guide for psychometric analyses**, created through reviewing secondary research and mapping the landscape of Spanish language assessments.

2. **A Growth Toolkit**, containing a growth roadmap for the next five years, a landscape analysis summarising lessons from similar products, an overview of key growth levers, a go-to-market strategy and tools to support effective scaling.

3. **A proposal for a teacher-facing insights dashboard**, co-designed with teachers, developed following a review of best practices for dashboard development.

4. **Proposed UX flows for different stakeholders** to ensure that teachers, students, and other key stakeholders can use the app in a self-directed way

As a result of these LEAP project deliverables, DECILE will be able to:

1. **Expand Validation & Benchmarking of DECILE** with a larger, more diverse student sample and align it with best practices to establish reliability, validity, and effectiveness.

2. **Strategically prepare for growth** by refining the tool for wider implementation across Latin American schools, addressing diverse needs.

3. **Plan technological improvements** to support independent, large-scale use by teachers (incl. feedback dashboard, usage protocols).

These deliverables can also serve as a public good, with its tools and insights relevant both for others developing linguistic assessments or looking to scale a research-based ed-tech product.

Deliverable 1

Introduction

Deliverable 1 presents an overview of assessments and a strategy for establishing the reliability and validity of the DECILE assessment. This section provides a review of assessment goals and formats, including considerations for evaluating language abilities in deaf and hard-of-hearing (DHH) and bilingual children. Next, the section provides a comprehensive summary of reliability and validity of assessments, including a focus on specific analyses in service of establishing reliability over time and across test items (test-retest, internal consistency) and establishing criterion, construct, and content validity. The section includes a review of item response theory for optimizing test items in the DECILE assessment. Lastly, a roadmap for psychometric analyses of DECILE is provided, including steps for establishing proof of concept (reliable and valid assessment), assessment optimization, equitable language assessments, using DECILE for screening and progress monitoring in educational contexts, and establishing DECILE as a norm-referenced standardized assessment.

Purpose of Assessment

Assessment can have several purposes, including screening, diagnostics, progress monitoring, and outcome evaluation.

Screening

To determine a student's risk for language difficulty, and related learning challenges, and need for intervention. Children's language skills are foundational to learning (particularly learning to read in early primary school, which then permits children to read to learn in later primary school and beyond); it is critical to identify children who are struggling with language skills as soon as possible.

Screening measures are brief assessments of particular skills and typically focus on skills that are highly predictive of later outcomes. Screening can be administered several times per year (3-4) using standardized administration and scoring. The goal of screening is to quickly designate children into those who require intervention and those who do not. Data from screening can be used to make decisions about the need for interventions and the progress monitoring that follows.

Diagnostic Evaluation

To identify a student's language skills. A diagnostic evaluation would identify probable sources of academic challenges and identify if a student has a language disorder. In-depth,

time-intensive evaluations that are completely 1:1 by a trained specialist (psychologist, SLPs, educational specialist with advanced training). Typically administered once.

Progress Monitoring

To determine if a student's progress is adequate or if the student requires additional (or different) intervention.Progress monitoring is done to determine the student's response to an intervention as well as the rate of improvement. Progress monitoring assessments target skills that were previously identified (via screening or diagnostic evaluation) as in need of improvement. Progress monitoring should be frequent (weekly-monthly).

Progress monitoring assessments can be *interim assessments* (evaluate students' skills relative to a specific academic goal within a limited time frame) or *formative assessments* (ongoing evaluation to determine students' learning; used by teachers to provide feedback, modify instructions, or indicate areas needing further support). Relatedly, *summative assessments* are typically done at the end of the school year to evaluate students' performance relative to academic standards. Progress monitoring assessments often vary in reliability and validity.

Intervention Planning and Intervention Outcome Evaluation

To determine if an intervention is effective. For research—both basic and applied—to yield meaningful and applicable insights, the assessments used must be of high quality, aligned with the specific goals of the study, and appropriate for the population being studied. In research and evaluation contexts, a range of assessment tools is often used, including standardized tests, performance-based assessments, and tools specifically developed by researchers. For instance, assessments can be administered before and after an intervention (pre-post design) to children exposed to (treatment) or not exposed to (control) an intervention to evaluate whether an intervention approach is meeting its intended goals.

Types of Assessments

Standardized Reference Assessments

Standardized assessments are evaluation tools that are developed through research and designed to be both statistically reliable and valid. These tests require all individuals to respond to the same set of questions or tasks in a uniform manner and are scored consistently across all test takers. This standardized format allows for meaningful comparisons between individuals or groups. There are two main types of standardized assessments: norm-referenced, which compare a test taker's performance to that of a larger, representative group, and criterion-referenced, which measure an individual's performance against a fixed set of standards or learning objectives (American Speech-Language-Hearing Association, 2025).

Criterion-referenced Assessments

Criterion-referenced tests are standardized assessments that evaluate an individual's performance based on a specific set of predefined criteria or learning standards. Rather than comparing a test taker's results to those of others, these tests measure how well the individual has mastered particular skills or knowledge. The criteria are typically tied to developmental milestones or educational goals, outlining what a person is expected to know or be able to do at a certain stage of learning or grade level (American Speech-Language-Hearing Association, 2025).

Norm-referenced Assessments

Norm-referenced tests are standardized assessments created to compare and rank individuals based on how their performance relates to that of a larger group. These tests allow educators to evaluate a test taker's results in relation to a statistically selected sample—commonly made up of peers from the same age or grade level—who have previously taken the test. Scores from norm-referenced tests are typically expressed as percentile ranks, which indicate how a student performed compared to others in the norming group. E.g., a score in the 90th percentile means the student scored as well as or better than 90% of their peers in the comparison group (American Speech-Language-Hearing Association, 2025).

Dynamic Assessments

Dynamic assessment is an interactive approach to evaluating a child's learning potential by using a structured "test–teach–retest" model. This process begins with a baseline assessment (the initial "test"), followed by an instructional phase (the "teach") where specific skills or strategies are taught, and concludes with a second assessment (the "retest") to determine how much the child has learned and how effectively they responded to the instruction. Unlike traditional assessments that focus only on what a child currently knows, dynamic assessment provides valuable information about a child's ability to learn new concepts, their responsiveness to teaching, and the types of supports or strategies that are most effective for them. This model is particularly beneficial for DHH children, as it can help educators better understand how these students acquire new knowledge when provided with individualized support.

Dynamic assessments can be time-consuming, requiring careful planning, ongoing observation, and individualized instruction, which may be difficult to implement regularly in busy school settings. It also demands a high level of expertise from educators and specialists, who must be skilled in both the teaching and assessment components of the model.

Language Assessment in Deaf and Hard-of-Hearing (DHH) Populations

Testing Bimodal Bilinguals

Many DHH children use both spoken language and a signed language–bimodal bilingualism. Because these children are bilingual, considerations for language assessments of bilingual

children also apply for testing bimodal DHH children. It is important to evaluate the language skills of bimodal bilingual children in both of their languages to gain a complete and accurate understanding of their linguistic abilities. Assessing only one language can lead to underestimating their skills, potentially misidentifying typical bimodal bilingual development as a language delay or disorder. For example, bilingual children's total vocabulary (in both languages) is similar to or greater than the vocabulary of monolinguals (Bialystok et al., 2010). However, bilingual children usually know fewer words in one of their languages than monolingual speakers of those languages (Bialystok et al., 2010). Moreover, language skills across languages in a bimodal bilingual interact. Research on children who are simultaneously acquiring a signed (American Sign Language) and a spoken (English) language shows complex interaction between languages (e.g., Lillo-Martin et al., 2012 study of speech and sign Wh-question production in bimodal bilinguals).

Accommodations

Accommodations for DHH children in assessment settings can support equitable testing situations. Accommodations are assessment strategies designed to compensate for barriers inherent in testing, not to improve a child's performance beyond their actual abilities. The goal of high-quality accommodations is to provide access to tests and their content without changing the fundamental nature of the test. While various accommodations might be potentially useful for DHH children, they should not be applied universally. Instead, they should be considered on an individual basis to determine what a student would benefit from. Administration protocol may be necessary to ensure equity in the use of accommodations.

Accommodations include:

- Positioning: Preferential seating, separate locations for testing
- Use of assistive technology throughout the assessment
- Instructions: Repetition of directions, using a slower rate in delivering instructions
- Extended time

It is crucial to exercise caution in the selection and use of accommodations. If accommodations are chosen or implemented inappropriately, they can threaten the validity of the assessment by altering the skills or constructs being measured. The focus should always be on providing equitable access without compromising the integrity of the assessment.

Reliable and Valid Assessment

Several types of reliability and validity analyses are described below. Our recommendation is not that all of these analyses are completed or that they are completed in the precise manner in which they are described (other analyses may be appropriate beyond what is presented here). From our review of the assessment literature, few assessment tools will have undergone psychometric analyses that include all suggested approaches described below. Rather, the DECILE team should consider what is most appropriate to their goals and what is feasible at different stages of their growth roadmap.

Reliability

Reliability, the consistency of measurement, can be established across items (internal consistency), over time (test–retest), and across researchers (inter-rater).

Internal consistency

This method checks whether different items on the measure are measuring the same thing. If a test has several questions that assess the same skill, students' scores on those questions should be related. We can measure internal consistency in two main ways:

Split-Half Correlation: Here, we split the test in half (first half vs. second half) and see how students' scores on these two halves compare. A correlation above 0.80 is considered good.

Cronbach's Alpha: This is a more comprehensive way to assess internal consistency. It looks at all possible ways to split the test and calculates an average correlation. A Cronbach's alpha value above 0.80 indicates good internal consistency. A sample size of at least 30 can measure reliability using Cronbach's alpha.

Test-retest reliability

This method checks if students get similar scores when they take the same test more than once. To do this, a group of students takes the placement test twice, but the tests should be given close together in time (like within a week). This is important because if students learn more between tests, it might inflate their scores on the second test. We then look at how closely the scores from both tests match. A score correlation of 0.80 or higher indicates strong reliability, meaning the test consistently measures student abilities. For a correlation of $r=0.8$ ($R^2=.64$) with power $1-\beta=0.95$ and $\alpha=0.05$, a sample of 21 students is required.
Intra-class correlations coefficients (ICCs) can also be computed for determining test-retest reliability. ICC values less than 0.5 indicate poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability (Portney & Watkins, 2000).

Inter-rater reliability

This method checks whether results on the measure are reliable when administered by different examiners. Given that DECILE is a digital assessment, this type of reliability is not relevant for DECILE.

Validity

Validity, whether the assessment measures the target construct.

Criterion Validity

Criterion validity refers to whether scores on a new measure are correlated with an established standard of comparison. More specifically, criterion validity involves *concurrent*, *predictive*, and *retrospective* validity.

Concurrent Validity: whether scores on a new measure are correlated with another concurrent outcome collected at the same time.
Predictive Validity: whether scores on a new measure predict a future outcome.
Retrospective Validity: whether scores on a new measure are predicted by a previous outcome.

To establish criterion validity, DECILE's sentence comprehension task should be significantly correlated with another measure of syntactic comprehension. For example, a current DECILE score should predict a future outcome related to syntactic comprehension, and prior standardized assessments of language abilities, specifically syntax, should predict a current DECILE score.

Construct Validity

Construct validity refers to whether scores on a new measure can measure the target construct or skill. For example, DECILE should measure children's syntactic skills, and not vocabulary. More specifically, construct validity involves convergent and discriminant validity.

Convergent validity: whether scores on a new measure show a strong statistical relation with scores on conceptually similar measures.
Discriminant validity: whether scores on a new measure are not correlated with an established measure of an unrelated construct.

Analyses of construct validity use correlations to determine both convergent and discriminant validity. To establish convergent validity, DECILE's sentence comprehension task should be significantly correlated with a related, established measure. For example, the Spanish language Clinical Evaluation of Language Fundamentals (CELF) sentence comprehension subtest is an appropriate comparison or the Test de Comprensión de Estructuras Gramaticales (CEG). For discriminant validity, DECILE's sentence comprehension task should be uncorrelated (or less correlated) with an unrelated measure. For example, children can be administered DECILE, CELF, and a non-verbal intelligence task (e.g., matrix reasoning). Strong positive correlations between DECLINE and CELF, but not between DECILE and the non-verbal intelligence task, would indicate appropriate construct validity. Suggested cutoffs for correlation coefficients for convergent and discriminant validity are $r > .07$ and $-0.2 > r < 0.2$, which correspond to > 50% and <5% of shared variance between the related and unrelated measures.

Content Validity

Content validity refers to whether a measure covers the construct of interest. This is established based on an analysis of the statistical relationships between the measure's content and the construct it is intended to measure. Exploratory factor analyses (EFA) can be used to examine how different test items relate to the constructs measured. EFA can identify items that do not

empirically belong to the intended construct and should be removed from the measure. Confirmatory factor analyses (CFA) can be used to determine whether items confirm the theorized model.

For example, the DECILE assessment focuses on morphosyntactic skills underlying sentence comprehension and question comprehension. Sentences in the sentence comprehension task vary in morphosyntactic constructs: Subject-Verb-Object (SVO), Object-Verb-Subject (OVS), Passive sentences, Sentences with subject relative clauses, and sentences with object relative clauses. Similarly, sentences in the question comprehension task vary in morphosyntactic constructs: simple active sentences with canonical SVO order, who subject questions, who object questions, which subject questions, and which object questions.

Examples of test items on DECILE's sentence and question comprehension tasks

Sentence Comprehension Task	**Question Comprehension Task**
5 syntactic structures: **SVO - Simple active sentences with canonical order** (subject-verb-object) *El mono abraza al koala.* *(The monkey hugs the koala.)* **OVS - Simple active sentences with non-canonical order** (object-verb-subject) *Al mono lo abraza el koala.* *(The koala hugs the monkey.)* **Pass - Passive sentences** *El mono es abrazado por el koala.* *(The monkey is hugged by the koala.)* **SR - Sentences with subject relative clause** *El mono que abraza al koala está contento.* *(The monkey that hugs the koala is happy).* **OR - Sentences with object relative clause** *El mono al que abraza el koala está contento.* *(The monkey that the koala hugs is happy.)*	**5 syntactic structures**: **SVO - Simple active sentences with canonical order** (subject-verb-object) *La madre seca a la alumna.* *(The mother dries the student.)* ***Who* subject questions** *¿Quién seca a la alumna?* *(Who dries the student?)* ***Who* object questions** *¿A quién seca la alumna?* *(Who does the student dry?)* ***Which* subject questions** *¿Qué madre seca a la alumna?* *(Which mother dries the student?)* ***Which* object questions** *¿A qué madre seca la alumna?* *(Which mother does the student dry?)*

In the sentence comprehension task, there are 12 test items per syntactic structure assessed (total of 60 test items). EFA can be used to examine how the 60 test items correspond to related constructs (i.e., the five syntactic structures assessed). Shared variance among items is assumed to represent the construct. In EFA, covariances between all 60 items are analyzed, and items sharing a substantial amount of variance are collapsed into factors. Determining the

underlying factors structure is, therefore data driven, rather than prespecified. Several models with different numbers of factors can be computed. The optimal model is the model with the best fit. In CFA, the covariances between items are prespecified to measure specific constructs. Here, specific items measuring SVO sentence comprehension (12 items) would be prespecified to measure the SVO construct. EFA and CFA models for DECILE's sentence comprehension task are conceptualized in Figure 1 below. If the CFA shows good model fit, it is an appropriate model. Standard model fit indices include CFI (comparative fit index), RMSEA (root-mean-square error of approximation), and others. CFI >.9 and RMSEA <.07 indicate good model fit.

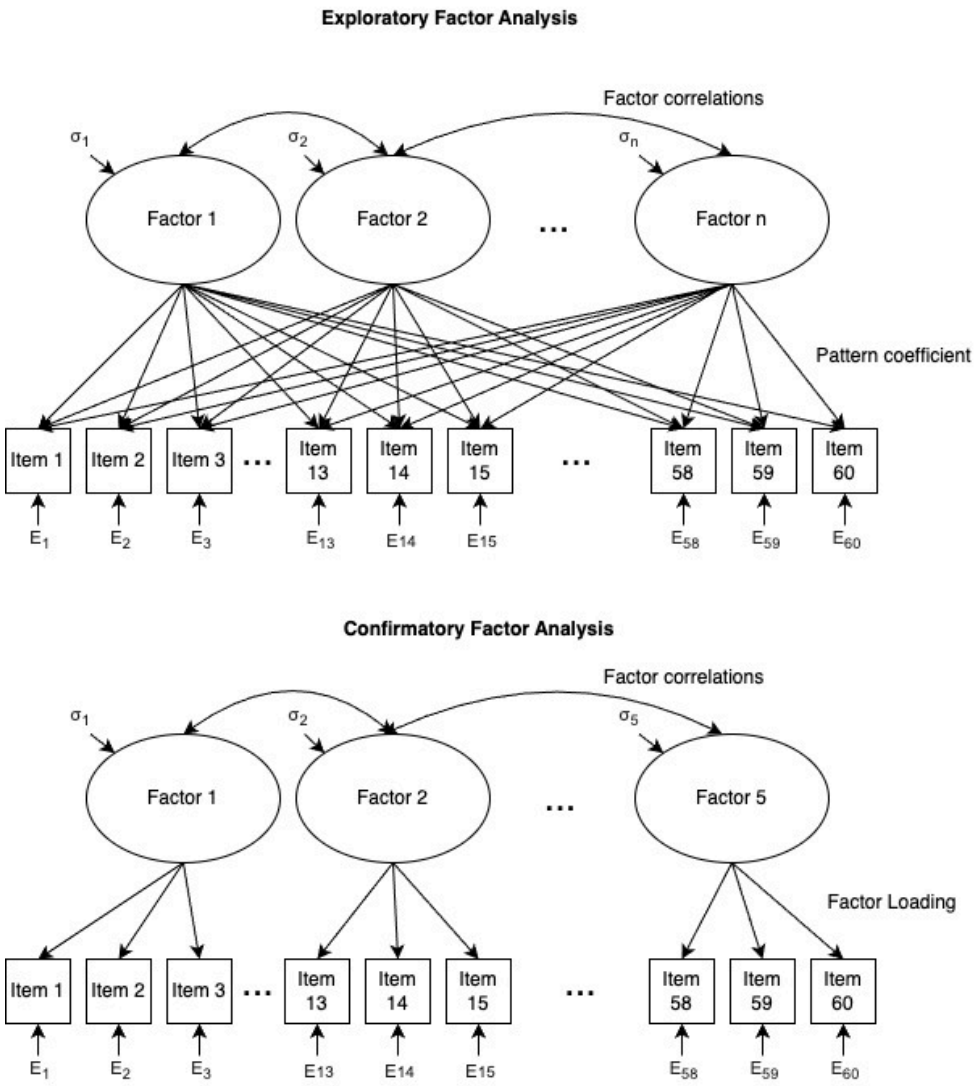


Figure 1: Exploratory and confirmatory factor analyses to establish content validity.

Note. DECILE sentence comprehension task includes 60 items (12 items measuring one of five syntactic structures). In EFA, covariances of all 60 items are analyzed (number of underlying factors is determined with analysis). In CFA, 12 specific items are prespecified to represent one of five constructs. This model is confirmed. σ represents the variance of each factor and E represents the unique error variance for individual items.

Factor analyses can establish whether the test items represent the target constructs that DECILE measures. Items that do not show significant association with the construct can be excluded from the measurement. EFA should be applied to an initial sample, and then CFA would be applied to a different sample. General guidelines suggest that a sample size of 100-200 participants is appropriate for factor analyses. However, the sample size depends on the number of factors, items per factor, size of factor loadings/pattern coefficients, and correlations between factors, among others.

Generally, high factor loadings for items are preferred. However, when several items have high standardized factor loadings (e.g., above .9), these items share a lot of variance, which suggests that the items may be too similar and not contribute unique information (see Clark and Watson, 1995). Low factor loadings mean that items share no or little variance with other items that theoretically should measure the same construct. There are no set rules for what factor loadings should be considered too low. A good guideline (Bandalos and Finney, 2010) is that items should explain at least 50% of the variance in a factor (R^2), which corresponds to standardized factors loadings of at least 0.7 (although lower values have also been used for content validity). If any item is removed, the EFA should be rerun to ensure that the original factor structure is still present (this can be done on the same data set).

Suggested: Several detailed guides and methodological papers exist for factor analyses in content validation. Knekta, Runyon, and Eddy (2019) provide a good primer.

Reliability and Validity of Digital Assessments

Computer-based assessments can present specific challenges for DHH children because they might rely on speech recognition or lack visual cues in the presentation of test items. Therefore, it is important to determine whether paper-based and digital versions of a task are similarly reliable and valid. Sorbes et al. (2023) provide an example of reliability and validity analyses conducted on paper-based and digital versions of a task. Children's performance across paper-based and digital versions of a task should be highly correlated. Reliability and validity metrics from each version of the task should be comparable.

Suggested Reading: Sobers, S. M., Whitehead, H. L., N'Goh, K. N. A., Ball, M. C., Tanoh, F., Akpé, H., & Jasińska, K. K. (2023). Is a Phone-Based Language and Literacy Assessment a Reliable and Valid Measure of Children's Reading Skills in Low-Resource Settings? *Reading Research Quarterly*, *58*(4), 733-754. https://doi.org/10.1002/rrq.511

Item Response Theory

Item Response Theory (IRT) assesses the validity of a measurement by describing the relationship between a latent trait (e.g., syntax abilities that are assessed with DECILE), the properties of test items in the assessment, and students' answers to individual test items. IRT

can be used to select the optimal subset of test items that measure language abilities reliably and efficiently (Wells, 2021).

IRT focuses on responses to individual test items and their characteristics. IRT describes how the respondent's ability and item characteristics contribute to the probability of a response (correct or incorrect). IRT models provide metrics of item difficulty and item discriminability that are useful for assessment development. IRT analyses provide item response curves and item information functions that are useful visualizations for determining whether any items should be excluded/revised.

A simple IRT model for binary responses (correct/incorrect) is the Rasch Model:

$$P(y_{ij} = 1) = \frac{exp(\theta_j - b_i)}{1 + exp(\theta_j - b_i)}$$

where $P(y_{ij} = 1)$ indicates the probability of a respondent (j) getting item i correct (score 1).
Two factors determine the probability of endorsing a correct answer:
Respondent j's ability (θj) and item i's difficulty (b_i)
exp represents the exponential function since a logistic regression form is adopted for probability

Item Discrimination (a_i) determines the rate at which the probability of endorsing a correct item changes given the respondent's ability levels. To design a precise measure, it is important to include items with high discrimination. Individual item discrimination generally range between 0-2 (although they theoretically range from -∞ to +∞). Items with poor discrimination should be revised. If an item has negative discrimination, this indicates that the probability of endorsing the correct answer decreases as the respondent's abilities increase (which should not happen).

Item Difficulty (b_i) determines the manner in which the item behaves along the ability scale. Items that are more difficult are less likely to be endorsed as correct by respondents with lower ability levels. Item difficulty is determined at the point of median probability (i.e., the ability at which 50% of respondents endorse the correct answer).

Item Response Curves

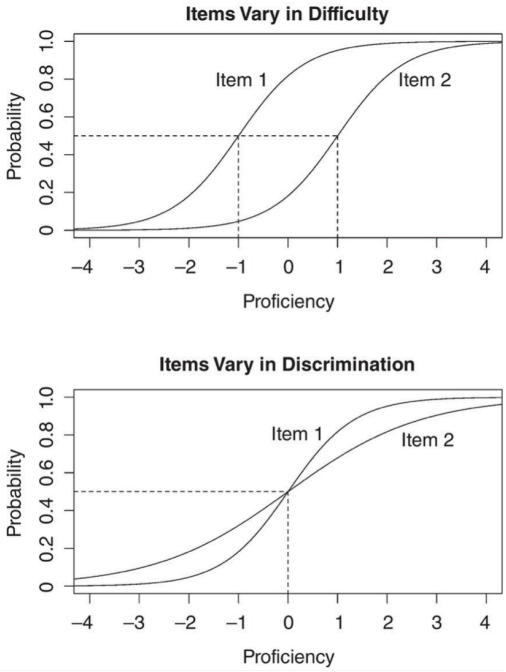


Figure 2: Example item response curves showing variation in item difficulty and discrimination (Wells, 2021).

In this example, there are 2 test items in the scale. The x-axis represents the respondents' ability (θ), and the y-axis represents the probability of endorsing a correct response. The point at which each test item crosses the horizontal line (e.g., $p(y_{pi}=0.5)$ corresponds to the item's difficulty (Item 2 is more difficult than item 1). Item discrimination is the slope of the line, with steeper slopes indicating higher discrimination (Item 1 has better discrimination than item 2).

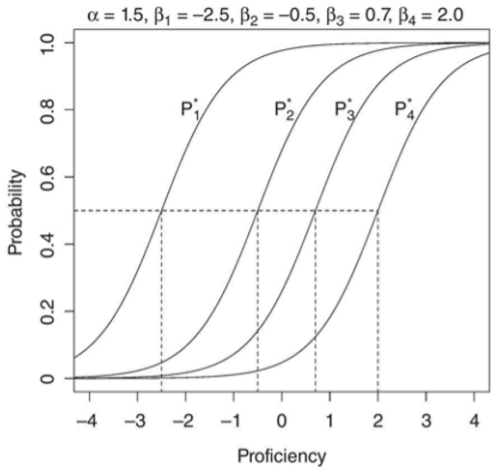


Figure 3: Example of four items with different difficulty (Wells, 2021).

Item Information Functions (IFF)

IFFs indicate whether items are informative; items that have low discrimination are not very informative with respect to the respondent's abilities. Similarly, if all items have the same difficulty, then there is less information that they provide.

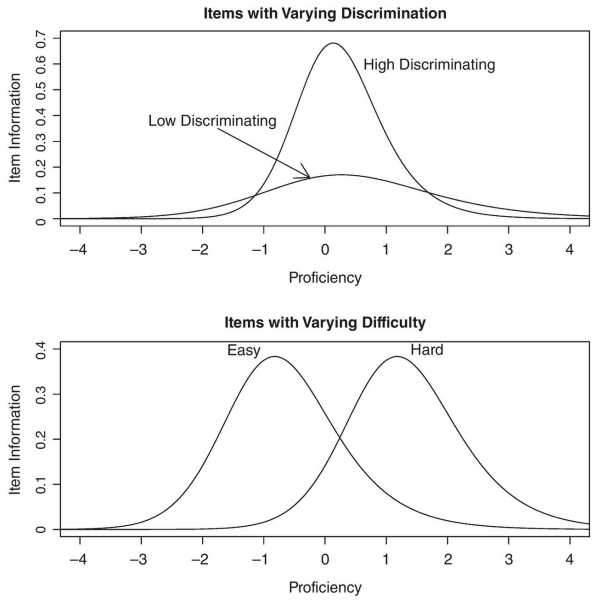


Figure 4: IFF showing differences in item informativeness (Wells, 2021).

Reliability and Validity Roadmap

The following sections outline a sequence of experiments to establish (1) the reliability and validity of DECILE (including across subgroups, and with potential accommodations for DHH), (2) optimize the DECILE assessment, (3) effective use of DECILE by an educator as a tool for screening and progress monitoring, and (4) establishing DECILE as a norm-referenced standardized assessment.

The results of the analyses suggested should be published in academic journals. Some examples include:
Barrington, E., Sarkisian, S. M., Feldman, H. M., & Yeatman, J. D. (2023). Rapid Online Assessment of Reading (ROAR): evaluation of an online tool for screening reading skills in a developmental-behavioral pediatrics clinic. Journal of Developmental & Behavioral Pediatrics, 44(9), e604-e610.
Yeatman, J. D., Tran, J. E., Burkhardt, A. K., Ma, W. A., Mitchell, J. L., Yablonski, M., ... & Richie-Halford, A. (2024, December). Development and validation of a rapid and precise online sentence reading efficiency assessment. In Frontiers in education (Vol. 9, p. 1494431). Frontiers Media SA.

Yeatman, J. D., Tang, K. A., Donnelly, P. M., Yablonski, M., Ramamurthy, M., Karipidis, I. I., ... & Domingue, B. W. (2021). Rapid online assessment of reading ability. Scientific reports, 11(1), 6396.
Gijbels, L., Burkhardt, A., Ma, W. A., & Yeatman, J. D. (2024). Rapid online assessment of reading and phonological awareness (ROAR-PA). Scientific Reports, 14(1), 10249.
Jasińska, K. K., Akpe, Y. H., Seri, B. A. D., Zinszer, B., Agui-Kouadio, R. Y., Mulford, K., Curran, E., Ball, M.-C., & Tanoh, F. (2022). Evaluating Bilingual Children's Native Language Abilities in Côte d'Ivoire: Introducing the Ivorian Children's Language Assessment Toolkit for Attié, Abidji, and Baoulé. Applied Linguistics, 43(6), 1116-1142. https://doi.org/10.1093/applin/amac025

Proof of concept

Examine the validity of DECILE against an existing assessment with excellent psychometric properties and widely-accepted validity. Students can complete the computerized DECILE assessment as well as undergo one-on-one person testing with an established assessment to examine the validity of DECILE digital assessment against the paper-based version of the task. Examine the reliability of DECILE, across time and across test items.

Established assessments for comparison:

1. Clinical Evaluation of Language Fundamentals - Fifth Edition - Spanish for Spain (CELF-5 Spain) This assessment has measures of sentence comprehension and norms were collected in Argentina.
2. Grammatical Structures Comprehension Test (CEG).
3. Woodcock-Muños Language Survey. Includes measures of language and reading.

Suggested Analyses:

1. Correlations to establish criterion and construct validity. Test whether DECILE scores are correlated with standardized assessment scores and whether scores predict outcomes.
2. Factors analyses to establish content validity.
3. Split half correlations and/or Cronbach's alpha to determine internal consistency.
4. Correlations to establish test-retest reliability.

Optimization

DECILE should aim to provide a valid, reliable, expedient, and automated assessment of language skills. The number of items in the assessment should be sufficient to provide an evaluation of children's language abilities, but not excessive as to make testing burdensome to teachers and students. Performance on DECILE assessments should not be affected by test effects such as the order in which subtests are administered.

Suggested Analyses:

1. IRT to optimize the test items for difficulty and discrimination and ensure items are informative with respect to children's abilities. The number of items can be reduced to include informative items at different levels of difficulty.

2. Order effect: Regression or ANOVA models can determine whether performance on DECILE assessments is affected by the order in which the assessments were done (e.g, BOSQUE versus SELVA task). Task order (1st or 2nd) can be specified as one of the independent variables. Non-significant results suggest task order is not related to performance.
3. After the test items are optimized, reliability and validity analyses should be done to ensure the assessment is reliable and valid with a reduced item set.

Equitable Assessment: Validity and Reliability within Subgroups

Once DECILE is determined to be reliable and valid with the respondent sample, it is important to ensure that the assessment is reliable and valid for all children. Reliability and validity should be compared across key subgroups of children. This can include:

- Different age groups
- DHH and typically hearing children
- Children with cochlear implants or those without
- Boys and girls
- Children from higher and lower SES backgrounds
- Children who completed the DECILE assessment with or without accommodations
- Bilingual, bimodal bilingual, monolingual children

DECILE for Screening and Progress Monitoring

For DECILE to be used by teachers, including dynamic assessments in the classroom, the DECILE team needs to consider how educators can use assessment data and develop appropriate technical manuals and test administration protocols.

Core reliability and validity analyses should be already completed to establish DECILE's reliability and validity before DECILE is used for screening and progress monitoring by teachers. This should minimally include validation of DECILE against an existing assessment with good psychometric properties.

How Educators Can Use DECILE Assessment Data

Student assessment data can be used to support instructional decision-making. Educators can make several types of instructional changes based on assessment data to improve student outcomes. These include:

1. Prioritizing instructional time
2. Additional individual instruction for students struggling with certain skills. E.g., adapt lessons or assignments in response to students' needs, modify student-grouping arrangements
3. Facilitate the identification of individual students' strengths and weaknesses, and identify interventions appropriate to specific students based on students' assessment scores
4. Evaluating the instructional effectiveness of classroom instruction
5. Refine instructional methods. E.g., alter classroom goals or objectives.
6. Use schoolwide data to inform curriculum adaptation

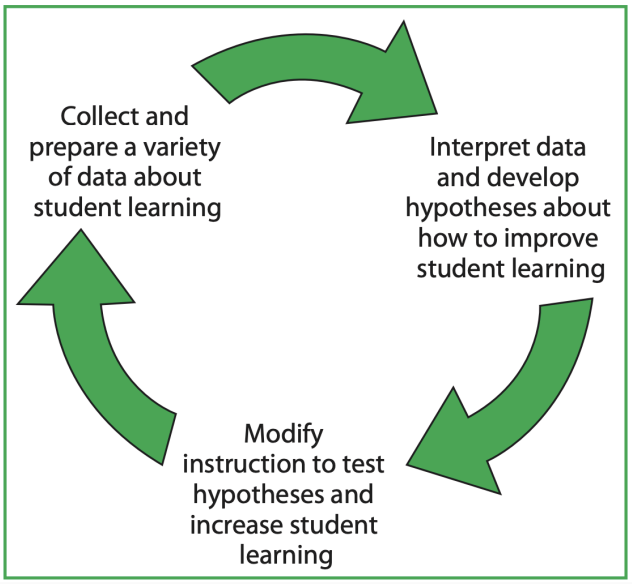


Figure 5: Teachers' process for using assessment data to inform instruction is visualized in the data use cycle (Institute of Education Sciences, 2009).

Teachers can use DECILE data about students' language skills in combination with other assessment data (interim, formative, and summative classroom assessments, diagnostic evaluations) to develop hypotheses about factors that contribute to students' performance and specific instructional actions the teachers can take to meet students' needs. Teachers can test these hypotheses by changing the instructional practices, followed by collecting performance data to determine the outcomes of instructional changes. It is recommended that teachers use several sources of data in this process to manage the limitations of each data type. For instance, diagnostic evaluation may be most comprehensive, but due to the time-intensive nature of these assessments, there is likely a longer lag between the assessment date and the student's current performance level. Interim, formative, and summative classroom assessments are more likely to not be reliable and valid. DECILE can provide a reliable and valid assessment that teachers can readily incorporate into their data sources.

Beyond using DECILE assessment data to respond to students' academic strengths and needs, schools can use DECILE as part of building longitudinal data systems. DECILE can be part of creating organizational and technological conditions that foster effective data use in education.

Suggested Further Reading:
IES Using Student Achievement Data to Support Instructional Decision Making.
Jang, E. E., Wagner, M., Hannah, L., & Kim, H. (2024). Diagnostic Assessment Feedback in the 21st-Century Technology-Rich Classroom. The Concise Companion to Language Assessment, 339.

Technical Manuals and Test Administration Protocols

A manual to accompany DECILE should broadly consist of two types of information: a user guide and a technical manual.

User Guide

A user guide must be developed to provide detailed instructions on how DECILE should be administered, how to navigate and use the dashboard, and how scores should be interpreted. The user guide should include recommendations for accommodation support. The user guide should also include test invalidity rules that describe the conditions under which the assessment is not scored. This typically concerns the number of missing items that prevent the assessment from being scored.

User guides can typically include the following components:

- Description of assessment and sub-tests
- General test administration (preparation for testing, administration, recommended accommodations)
- Scores and interpretation

Technical Manual

Technical manuals should also report the results of core reliability and validity analyses, and particularly highlighting any subgroups of children from whom DECILE was not reliable or valid.

Technical manuals can typically include the following components:

- Test design and development procedures
- Standardization and norm development procedures (if applicable–norming work described in the section below has been completed)
- Results of reliability analyses
- Results of validity analyses

Suggested examples:

BALANCE AI manuals and video resources for teachers.

Mather, N. & Wendling, B.J. (2014). Examiner's manual. *Woodcock-Johnson IV Tests of Cognitive Abilities*. Rolling Meadows, IL: Riverside Publishing.

McGrew, K. S., LaForte, E. M. & Schrank, F. A. (2014). Technical manual. W*oodcock-Johnson IV*. Rolling Meadows, IL: Riverside Publishing.

DECILE as a Norm-Referenced Standardized Assessment

For DECILE to become a norm-referenced standardized assessment of Spanish language skills, extensive psychometric work on a normative sample is necessary as well as the development of a comprehensive technical manual that reports the psychometric properties of the test.

Norm-referenced tests enable us to evaluate an individual's score by comparing it to a specific reference group. This reference group is established based on the sampling strategy for

obtaining a normative sample. Norms help determine an individual's position on the measured ability (e.g., language abilities) relative to others in the reference group (Miller, 2020).

Normative sample

Normative samples form the reference group to which individuals will be compared.

Normative samples should be representative and inclusive. Normative samples should be randomly sampled to safeguard against introducing systemic biases into a sample. It is important to try to avoid obtaining a sample by convenience, which will tend to systematically over- or under-represent certain characteristics (e.g., higher SES parents are more keen to have their child participate in a norming study and the references sample over represents traits that are linked with higher SES).

Normative samples are typically very large. For example, the CELF-5 (English language) included a normative sample of 830 children and adolescents. It is important to consider the age range for testing as an adequate sample in each age group would need to be included in normative work. Norms can be stratified by relevant characteristics such as age, gender, parent education level, or geographic region. Most standardized tests use a stratified sampling procedure where equal numbers of boys and girls in each age group are included (and other relevant characteristics). It can be helpful to track cells (e.g. age groups by gender) during sampling to avoid oversampling a particular cell. A suggestion is to use 50-75 participants per cell (based on normative studies in pediatric neuropsychology; Bridger & Holler, 2007). For example, the CELF-5 normative sample was divided into 15 age groups: 5:0–5:5, 5:6–5:11, 6:0–6:5, 6:6–6:11, 7:0–7:11, 8:0–8:11, 9:0–9:11, 10:0–10:11, 11:0–11:11, 12:0–12:11, 13:0–13:11, 14:0– 14:11, 15:0–15:11, 16:0–16:11, and 17:0–21:11. Ages 5–6 were broken down into 6-month intervals; ages 7–16 were broken down into 1-year intervals; and ages 17–21 were combined into one interval.

Norms and Scores

Individual respondents' scores are raw scores. This is the actual number of correct responses that a respondent provided on an assessment. Raw scores don't provide information that allows the teacher or test administrator to compare a child's performance to others. Raw scores need to be converted to a normed score. Scores can be computed based on large normative samples. This should be done for age groups and/or hearing age groups, if appropriate. For new respondents, a normed score can be calculated based on the test performance of their peer group.

Scores are typically divided into developmental or within group scores. Developmental scores include age and grade equivalent scores. Within group scores include percentile ranks and standard scores.

Age equivalent scores. The age equivalent of a students' raw scores is the chronological age of other students whose mean raw scores is the same. Child A scores 55, 55 is the average score of the 8 year olds in the norming sample. Child A's age equivalent score is 8.

Grade equivalent scores. The grade equivalent of a students' raw scores is the grade of the other students whose mean raw scores is the same. Child A scores 55, 55 is the average score of the 3rd grade students in the norming sample. Child A's grade equivalent score is 3rd grade.

Percentile Scores. A percentile rank shows where an individual child stands in comparison to the standardization sample. It reflects the percentage of that group who scored at or below a specific raw score. For instance, if a raw score of 55 corresponds to the 80th percentile, it means that 80% of the standardization group scored 55 or lower. In other words, the student performed as well as or better than 80% of the normative sample. The 50th percentile, or the median, represents the middle score in the distribution and indicates average performance. Percentile ranks do not increment proportionally with raw scores. A change from 50th to 60th percentile may correspond to an increase of 5 raw score points, but a change from 85th to 95th percentile may correspond to an increase of 10 raw score points because scores are normally distributed and therefore tend to cluster near the middle. Small raw score changes near the center of the distribution result in larger percentile changes, but larger raw score changes are required for equivalent percentile changes at the ends of the distribution.

Standard Scores. A standard score shows where an individual child's score is compared to the mean of the distribution of the standardization samples in standard deviation units. Z-score is the most common. Standard scores do increment proportionally with raw scores. The standard score has the same meaning throughout the range of scores.

Deliverable 2 - Growth Toolkit

Introduction

DECILE is a tech-enabled language assessment that has been developed with rigour and precision by a team of academics. While the team wished it to have the largest possible reach and impact, the development of DECILE was not driven by a commercial imperative, nor do the team have experience in growing and scaling EdTech products.

This second deliverable, therefore, is a **Growth Toolkit**. Its aim is to equip the team to scale DECILE, within Argentina initially, and then to other Spanish-speaking countries and to groups of children with needs other than hearing impairments.

This deliverable was initially conceived of as a Growth Roadmap. We later expanded our scope, recognising that a roadmap, though useful, is a static output, when what the team really needs is the dynamic capability to prepare for growth and respond to evolving circumstances. As such, we have instead created a Growth Toolkit. It includes five components:

1. An overview of **DECILE today**, including a SWOT analysis and newly drafted Theory of Change.
2. A summary of DECILE's key **Growth Levers**, including stakeholder mapping and the creation of user persona.
3. A **Growth Roadmap**, co-created with the team, which outlines a three-phase vision for success, as well as corresponding barriers and dependencies in each phase.
4. A **Scalability Checklist** to aid the team to assess readiness for scale, adapt to unforeseen events and reflect dynamically on scaling goals.
5. A **Landscape Analysis**, summarising ideas and inspiration from a range of comparator products.

Each of these outputs is introduced in more detail below.

DECILE Today

Interactions with the DECILE team provided great insights into what drives them to do this work. Every member of the team brings a different set of skills but when they come together, they are driven by a common purpose. Outlined below is their reason for existence (or the problem statement), Vision and Mission, reproduced verbatim as provided by the DECILE team.

DECILE's reason for existence

"Language is the bridge that connects minds to the vast world of ideas and emotions. It empowers children to understand others, express their thoughts, and unlock their full potential. But when linguistic challenges arise, a child's ability to connect, learn, and thrive can be compromised.

DECILE is dedicated to transforming how Spanish-speaking children's communication skills are assessed, providing an accessible, research-driven, and culturally responsive platform."

Vision for DECILE

"DECILE envisions a world where every child—regardless of linguistic background or ability—receives the support they need to develop strong communication skills. By leveraging cutting-edge psycholinguistic research and innovative digital tools, DECILE aims to ensure that no child is left behind due to language-related barriers."

DECILE's Mission

"To identify diverse language learning profiles early, reduce biases in assessment, and equip educators, health practitioners and caregivers with actionable insights. By bridging the gap between science and practice, DECILE strives to foster equitable learning opportunities, empowering all children to connect, express themselves, and reach their full potential."

DECILE's Theory of Change

A carefully formulated Theory of Change (ToC) can play a key role in helping the team stay true to their mission as they progress through different stages of the growth journey. A ToC defines the path and, as such, should undergo evolution at various stages of the organisation's journey. However, sometimes, the work of creating a ToC can become overwhelming and tedious, taking away valuable time and energy from the team. To simplify the process, we introduced the idea of a one line ToC (https://mightyally.org/blog/one-sentence-theory-of-change)

The template:

Because of _____ (WHY), we work in _____ (WHERE) to help _____ (WHO) by _____ (HOW), in order to achieve _____ (WHAT) by _____ (WHEN).

The team came up with many different versions of the ToC, each of them bringing in their own perspective. From this, it was evident that two ToCs were needed - one for the organisation and one for the product. The final one line ToCs are outputs of this exercise.

DECILE's one line Theory of Change (as an organisation)

Because of persistent inequities in language assessment and educational opportunities for DHoH children who experience linguistic, educational and social difficulties, we work in Spanish-speaking countries to help educators, health-practitioners, policymakers and families by providing accessible and evidence-based language assessment tools, in order to achieve more equitable language learning support and early intervention for children with diverse linguistic needs, reaching 65,000 children by 2030.

DECILE's one line Theory of Change (as a product)

Because of the disadvantages that hearing limitations impose on the educational development of Deaf and Hard of Hearing (DHoH) children, we work in schools and homes in Spanish-speaking countries to support educators, practitioners and parents by providing them a scientifically validated, AI-supported language assessment tool, in order to achieve more accurate linguistic evaluation and build personalized learning strategies for DHoH children.

Strengths, Weaknesses, Opportunities and Threats

In order to understand DECILE better, we carried out a SWOT analysis exercise with the team, where each team member identified, from their perspective, DECILE's strengths, weaknesses, opportunities for growth and threats. As it turned out, it offered an opportunity for the team to pause and reflect on their own beliefs while also getting a sense for the perspective of their colleagues.

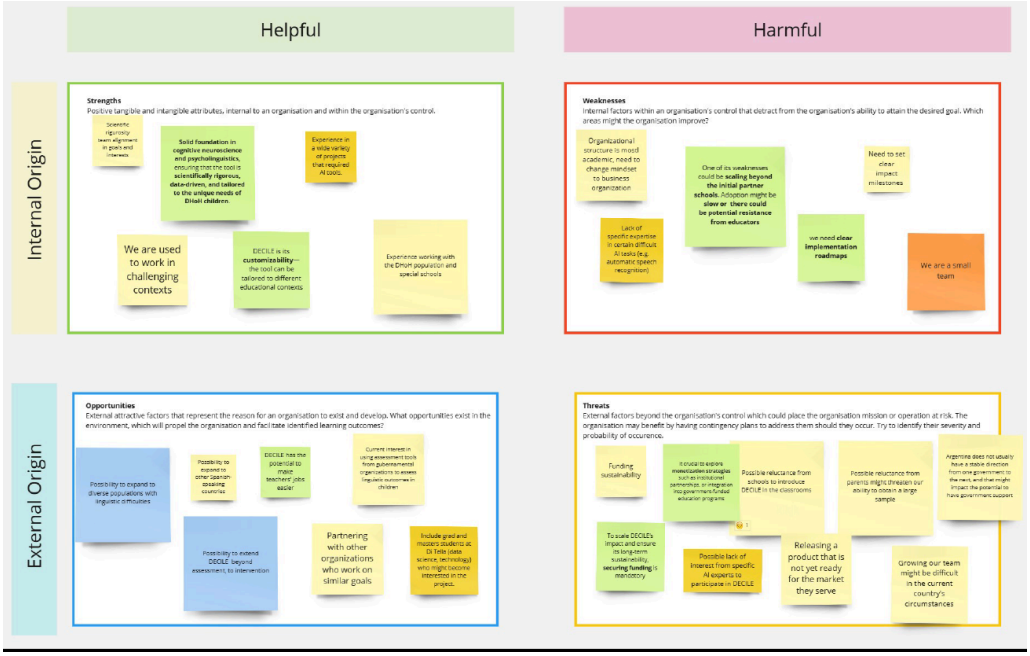


Figure 6: A glimpse of the SWOT exercise board with sticky notes from DECILE team members

The inputs from the team were consolidated and then split into two parts as with the ToC - one SWOT for the organisation and the other for DECILE as a product.

DECILE's Growth Levers

FOr DECILE to embark on a journey of growth, it is important for them to acknowledge the key levers that can help them reach their goals. For ease of reference, these levers have been collated under six broad categories

- Product-Market fit
- Customers
- Partnerships
- Team
- Business/Financial
- Marketing/Outreach

For each of these categories, we articulated what successfully drawing on these levers could look like for DECILE.

For example, successful partnerships with other institutions, government and service providers could mean:

- DECILE reaches 2x more students through partnership with institutions working with the target group
- The national government
    - recognizes DECILE as a credible tool for language assessment, particularly for DHoH children
    - allows DECILE to scale to 10x the number of students
- Collaborations with other service providers allows DECILE to
    - reach target populations outside of Buenos Aires
    - reach 3x more students by integrating into existing products being used by the target population.

Key Barriers to Growth

Along with identifying the levers for growth, it is important to recognise the barriers that could potentially derail DECILE from their growth journey. Based on our understanding of DECILE's ambitions, we identified six key barriers to growth.

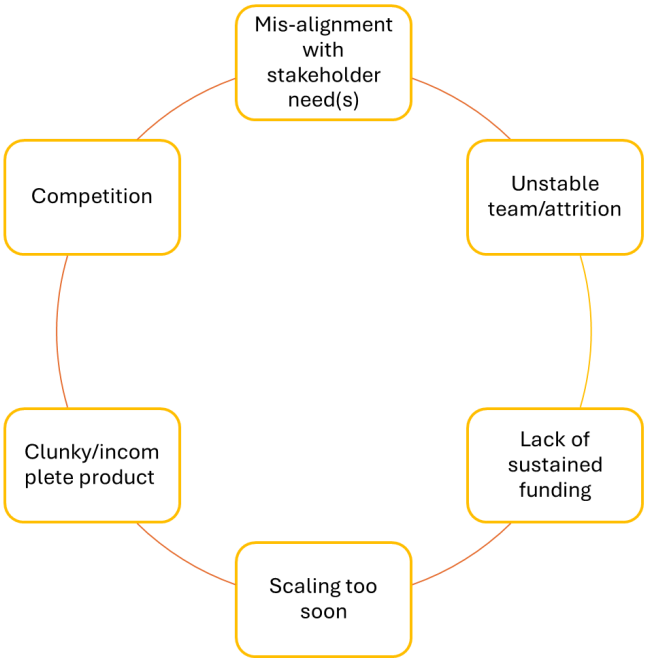


Figure 7: Barriers to DECILE's growth

For each of these, we then articulated why they are critical and what the team should watch out for, to ensure they stay on course. Figure 8 below showcases one example.

Scaling too soon

Getting the product in the hands of more children is exciting, but for sustainable growth, scale should come at the right time and for the right reasons.
Scaling too soon can lead to burnout, quality degradation and reputation damage.
Things to keep in mind:

- Determine whether you are ready to scale. Refer to the section on Scalability Checklist for an easy to use template.
- Scaling need not mean more reach. It can mean deeper engagement. Pick one.
- Scale comes at a cost - understand the choices and consequences before aspiring for scale.

Figure 8: Example of the information provided for one of the barriers to growth for DECILE

Stakeholder Mapping

As a first step in understanding DECILE and its users, we mapped out the key stakeholders for DECILE. As a team of researchers, understanding the students who would ultimately benefit from the program was essential. However, the product the team is building is meant to be used by educators in the classroom. From a product perspective, lack of clarity in identifying the primary user can lead to a cluttered product design that doesn't serve the intended purpose.

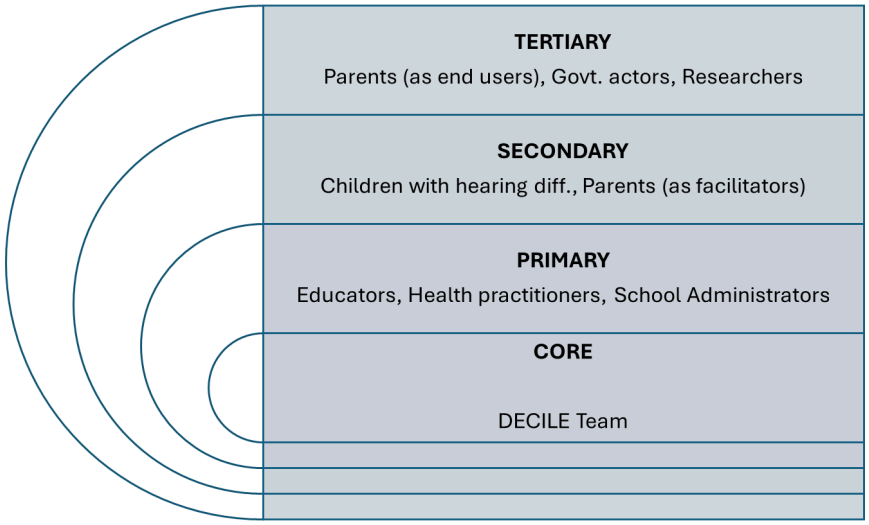


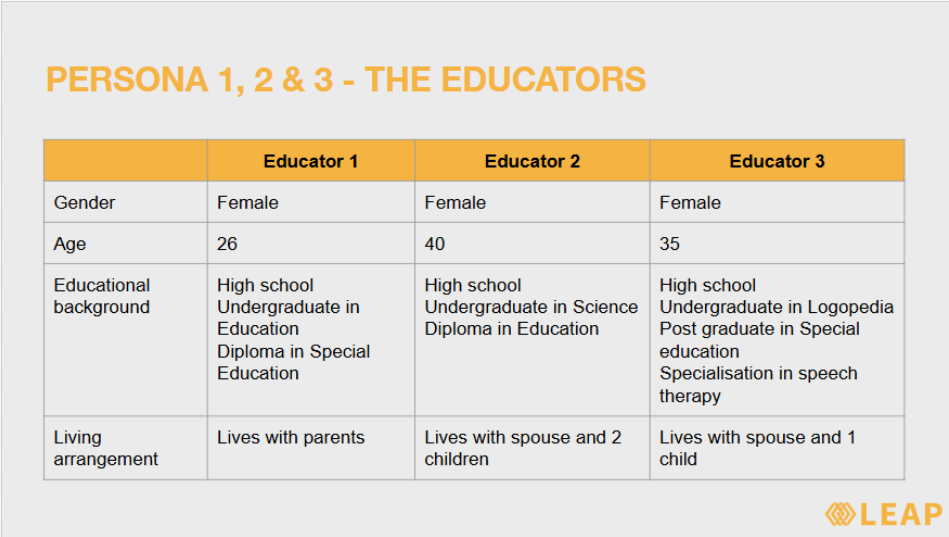
Figure 9: Stakeholder mapping for DECILE

User Personas

For DECILE, there are three key stakeholders -

1. The educator, the primary user for DECILE's product
2. The student, the secondary user for DECILE's product
3. The school official, the influencer and potential tertiary user of DECILE's product

For each of these stakeholders, we engaged with the DECILE team through a simple questionnaire, to understand and define the personas.



PERSONA 1, 2 & 3 - THE EDUCATORS

	Educator 1	Educator 2	Educator 3
Gender	Female	Female	Female
Age	26	40	35
Educational background	High school Undergraduate in Education Diploma in Special Education	High school Undergraduate in Science Diploma in Education	High school Undergraduate in Logopedia Post graduate in Special education Specialisation in speech therapy
Living arrangement	Lives with parents	Lives with spouse and 2 children	Lives with spouse and 1 child

LEAP

Figure 10: An example of part of the educator personas

Growth Roadmap

Although the team is ambitious, they had not crystallised their ambitions for DECILE or mapped out their desired growth trajectory. As such, we worked with the DECILE team to:

- Define their **overarching priorities** for growth - and the kind of organisation they hope to grow into.
- Articulate their **specific ambitions for growth** in the short/medium/long term & create a high level roadmap.
- Define **success indicators** for each growth stage.
- Highlight **dependencies and potential barriers** for each stage.

Overarching Priorities

To surface the team's overarching priorities for growth, we did a simple exercise that asked the team to rank their aims for DECILE as it grows. We provided the following list of options, based on what the team had already told us:

- Scaling across Spanish- speaking world
- Reaching the largest possible number of young people
- Becoming a profitable business
- Continued data generation for research
- Scaling to young people with needs other than hearing impairment
- Having maximum impact on young people's lives
- Equipping teachers to use the tool independently
- Scaling through government partnerships

We then asked each team member to rate, on a scale of 1-10, how much of a priority each of these eight potential aims is for DECILE as it grows (1- not important at all; 10 - of the utmost importance). They could only use each number once.

We then totaled up the team scores for each aim and ranked them. This enabled the team to see and discuss areas of alignment and divergence. This was the result of the exercise:

Ranking	Overarching aim	Score
1	Having maximum impact on young people's lives	49/50
2	Reaching the largest possible number of young people	36/50
3	Equipping teachers to use the tool independently	35/50
4	Continued data generation for research	34/50
5	Scaling to young people with needs other than hearing impairment	32/50
6	Scaling through government partnerships	28/50
7	Scaling across Spanish-speaking world	27/50
8	Becoming a profitable business	19/50

This is an awkward exercise, compelling people to make difficult choices. This is precisely why it is valuable: it reveals a hierarchy of priorities among goals that all feel important.

The results of the exercise showed that there is broad alignment and revealed that the team's priority is **maximum impact**: making DECILE available for as many children as possible, while generating ongoing research insights.

While the team is determined to become **financially sustainable**, commercial imperatives emerged as their lowest priority. They are aiming for a sustainable model that maximises impact and insight rather than profit, and which favours collaboration over competition.

Specific ambitions for growth

We asked each team member to map out their ambitions for growth across:

- The next 1-2 years

- The next 2-3 years
- The next 3-5 years

As well as including ambitions for each phase, we asked them to list success indicators, dependencies and potential barriers.

The team's roadmaps were similar and demonstrated strong existing alignment. We reviewed them in a Growth Workshop, working through divergences in order to create a single, synthesised growth roadmap, which expresses the team's ambitions for the next five years.

Here is the synthesised high level roadmap we generated:

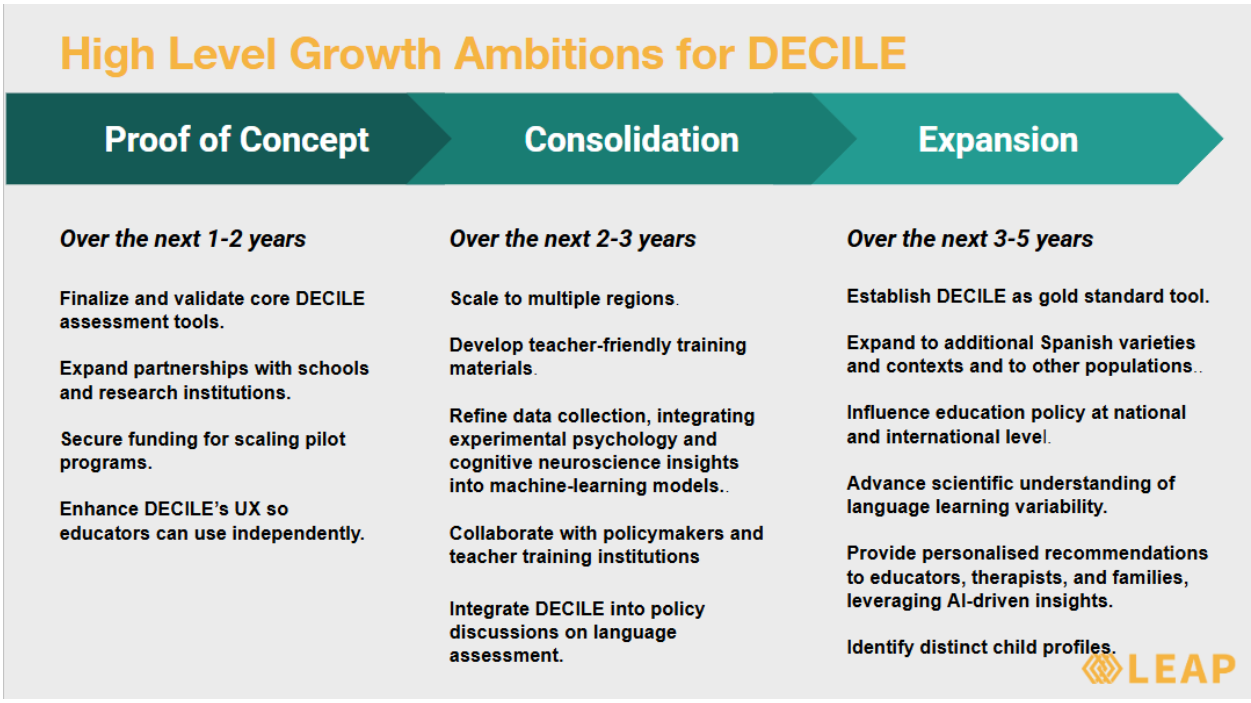


Figure 11: High level roadmap

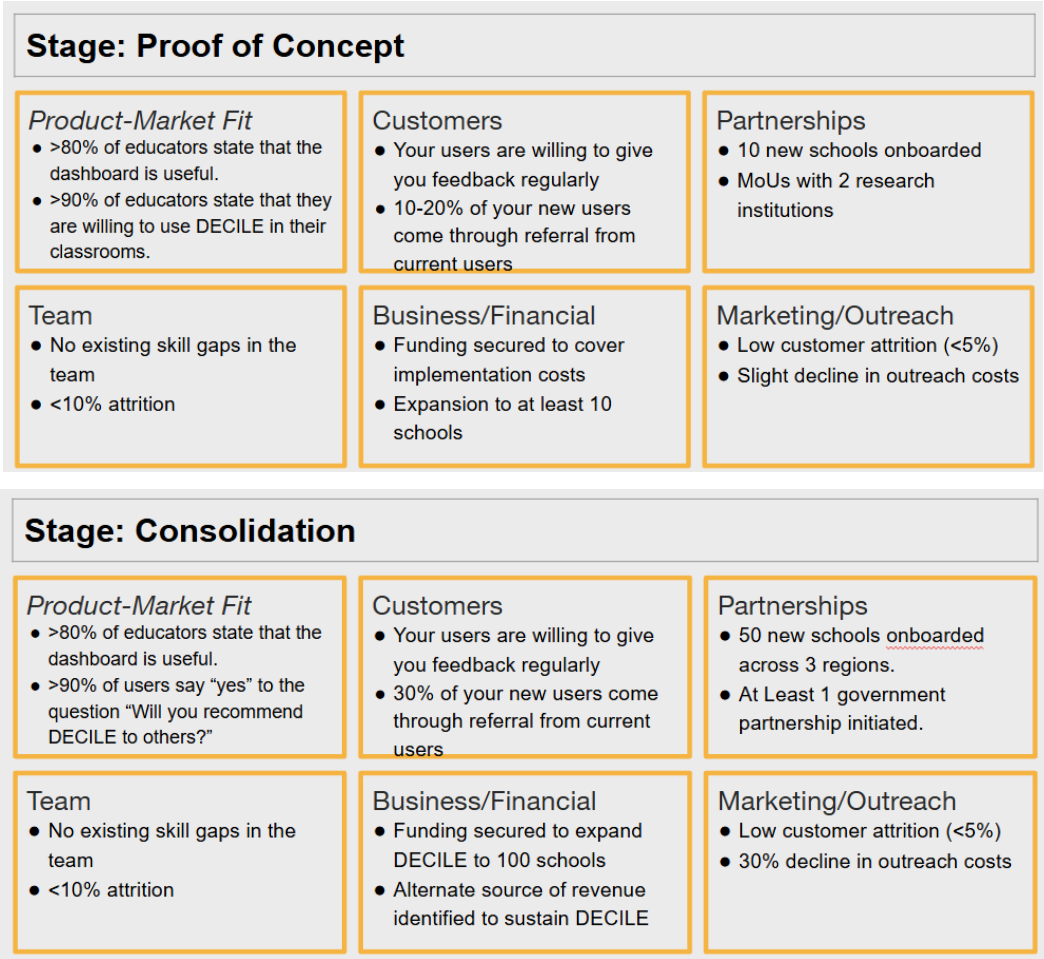
Although circumstances will likely evolve, this map can provide a north star for DECILE's planning and decision making.

Success Indicators

To support the team to reflect on progress at each phase, we have drafted some indicative success measures in relation to the following factors:

- Product-market fit
- Team
- Customers
- Business/financial
- Partnerships
- Marketing/Outreach

The DECILE team should review and adapt these in light of emerging circumstances, but these rough indicators give a sense of the basket of measures that could be assembled to evaluate progress.



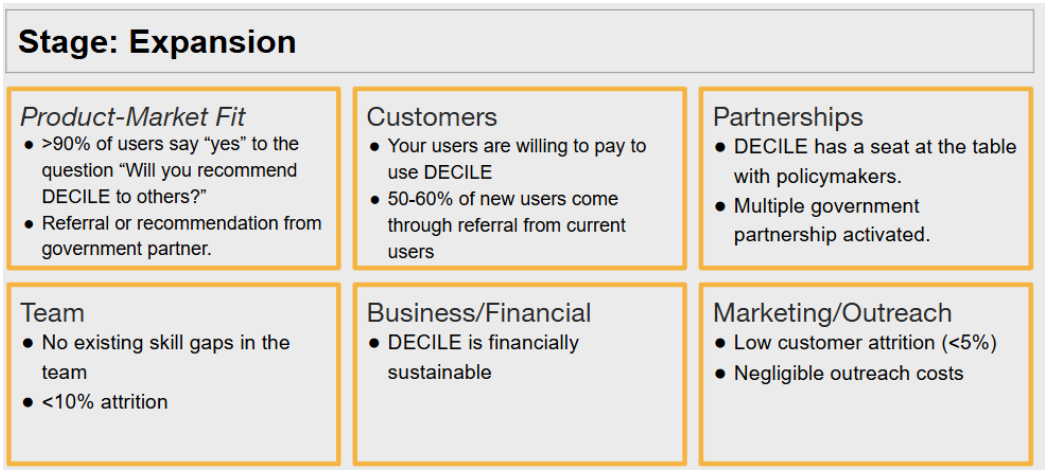


Figure 12: Proof of Concept, Consolidation, Expansion

Mapping dependencies and potential barriers

As part of our growth mapping exercises with the team, we asked them to identify barriers to growth and dependencies (what needs to be in place at each phase). Below is a synthesised summary of key factors they need to anticipate at each phase. These, too, are likely to evolve but this preliminary list will support the team as they plan for ambitious growth in the months and years to come.

WHAT MIGHT STAND IN YOUR WAY [BARRIERS / RISKS]	**Limited awareness** among stakeholders. **Data collection challenges** e.g. low sample size / low variability in participants. **Funding uncertainties**. **Resistance to change** in educators and health practitioners. **Low engagement** if teachers and therapists find the tool difficult to use. **Technical limitations in usability testing** that delay the validation process. **Challenge digitising** remaining assessments. **Mismatch between user priorities** and DECILE's growth plans.	**Resistance** to new assessment tools from schools or families. **Variability in educational policies** across regions. **Variability in infrastructure, connectivity, and resources,** preventing the use of a digital assessment. **Insufficient funding** for development and scaling. **Lack of governmental or institutional buy-in**, if educational authorities do not prioritize linguistic assessment for DHoH.	**Scaling & adaptation challenges** across diverse educational systems, diverse infrastructures and cultural contexts. **Maintaining scientific rigor** while expanding. **Long-term sustainability** beyond initial funding, to enable ongoing updates, training and wider adoption. **Ethical and data privacy concerns** when expanding the tool to a broader audience. **Inconsistent data quality** affecting accuracy and reliability of personalised machine learning recommendations.

WHAT NEEDS TO BE IN PLACE FOR THIS TO HAPPEN [CONDITIONS / DEPENDENCIES]	**Strong collaborations** with schools and researchers, generating rich feedback. **Engagement from teachers and health practitioners** who see the value in DECILE. **Ethical approval and logistical readiness** for data collection. **Effective outreach** and dissemination strategies. **Sufficient political stability** in the country. **A simple, accessible user experience** for practitioners and children. **Funding/grants for software optimization** including technical improvements, bug fixes, and user experience enhancements.	**Engaged educators and policymakers**. **Effective strategies to kindle and sustain school engagement**. **Robust data collection** by teachers/practitioners. **Adaptation** to new contexts. **Institutional partnerships for wider adoption**, including government agencies, NGOs, and educational institutions. **Pilot studies** demonstrating that DECILE improves assessment accuracy and informs better teaching strategies. **Continuous professional development** to help teachers and practitioners use DECILE effectively. **Sustained funding and grants** for scaling, additional research, and technology maintenance.	**Continued research and validation** (as knowledge on neuroscientific findings on language acquisition evolve). **Strong institutional and governmental partnerships**. **Sustainable business model or institutional support** for long-term impact. **Policy integration and national recognition** as a standard assessment tool for DHoH children. **Large-scale, robust data collection** for AI-driven insights. **Regional and linguistic adaptability**.

Scalability Checklist

An ambitious roadmap can bring clarity, energy and alignment. It focuses the team around shared goals and provides a shared reference for team discussions and decision making. But it also needs to be flexible enough to respond to unpredictable events.

As well as supporting the DECILE team to outline tangible ambitions for growth, we wanted to equip them with tools to think about scaling DECILE on an ongoing basis - and change course as required.

One of the most valuable tools for thinking rigorously and creatively about how to scale education products is the Brookings Education Scalability Checklist: a tool, in English and Spanish, that can be employed annually, or at key decision points, to systematically consider readiness for scale and issues to address.

The checklist and user guide can be found here in English:

The Education Scalability Checklist [EN]
The User Guide to the Education Scalability Checklist [EN]

And here in Spanish:

La 'Lista de verificación de la escalabilidad para la educación' [ES]
La guía del uso de la 'Lista de verificación de la escalabilidad para la educación' [ES]

Brookings describe the tool like this:

> "The newly launched Education Scalability Checklist (ESC) is meant to help determine the ease of scaling a particular education initiative—not as a one-off evaluative exercise but as a dynamic diagnostic tool to help identify areas where an initiative might require further consideration and adaptations in order to increase the likelihood of scaling."

The dynamism, and ongoing value, of this tool is where its value lies for the DECILE team. The tool can be used at key review and decision points to provide insight on seven areas:

A. Convincing scaling strategy
B. Credibility of initiative
C. Strength of support for initiative
D. Relative advantage
E. Ease of transfer and adoption
F. Fit with education system
G. Sustainability of funding

Completing the Scalability Checklist is a collective exercise for the team. Each person answers 32 questions, encompassing these themes. For example, 'Is there a sustainable source of funding?' Team members select A, B or C to show which best describes the current state and provide a short written rationale for each answer.

Item	A	<<<<< Scaling up is easier	B	Scaling up is harder >>>>>	C
30.	--	Budget implications are clear, predictable and justifiable for those who are expected to bear the costs	--	Budget implications are unclear and / or difficult to predict and justify for those who are expected to bear the costs	X
31.	--	Taking the initiative to scale requires small amounts of funding that can easily be mobilized internally or externally	X	Taking the initiative to scale requires large commitment of funds that are difficult to mobilize either internally or externally	--
32.	--	Cost of implementation at scale can be integrated in government budget and / or in budget of key implementing organizations	--	Cost of implementation at scale cannot be integrated in government budget and / or in budget of key implementing organizations	X

Figure 13: Scaling Checklist example.

These individual answers can then be combined into an overall group score and a list of insights, which become a prompt for discussion, highlighting areas of alignment and divergence, as well as areas that may have been overlooked within the team's scaling strategy.

The instructions for the checklist are in the user manual. However it is worth noting that:

- Answering the questions thoroughly (including a short written rationale for each point) can take 1-2 hrs per participant. Likewise, synthesising the team's comments and scores can take several hours.
- The checklist works on the assumption that the aim for a product is to be adopted by government institutions. The wording of these questions (Section E) can be tweaked to more accurately describe your aims.
- It's important to watch out for confirmation bias and to look for evidence to support different claims.

Completing the checklist takes 1-2 days. It is not a massive investment of time, given the value it can add to team planning, but time does need to be earmarked to do it properly. Using the checklist:

- Ensures systematic reflection on key considerations when scaling, surfacing any potential blind spots or risks that may have been underestimated.
- Gathers and synthesises the team's views, providing rich information for discussion and decision making.
- Generates a clear, high quality action plan with a relatively low time investment.

The checklist exercise can be repeated annually or at key decision points, with the results of each round showing change in scaling readiness over time.

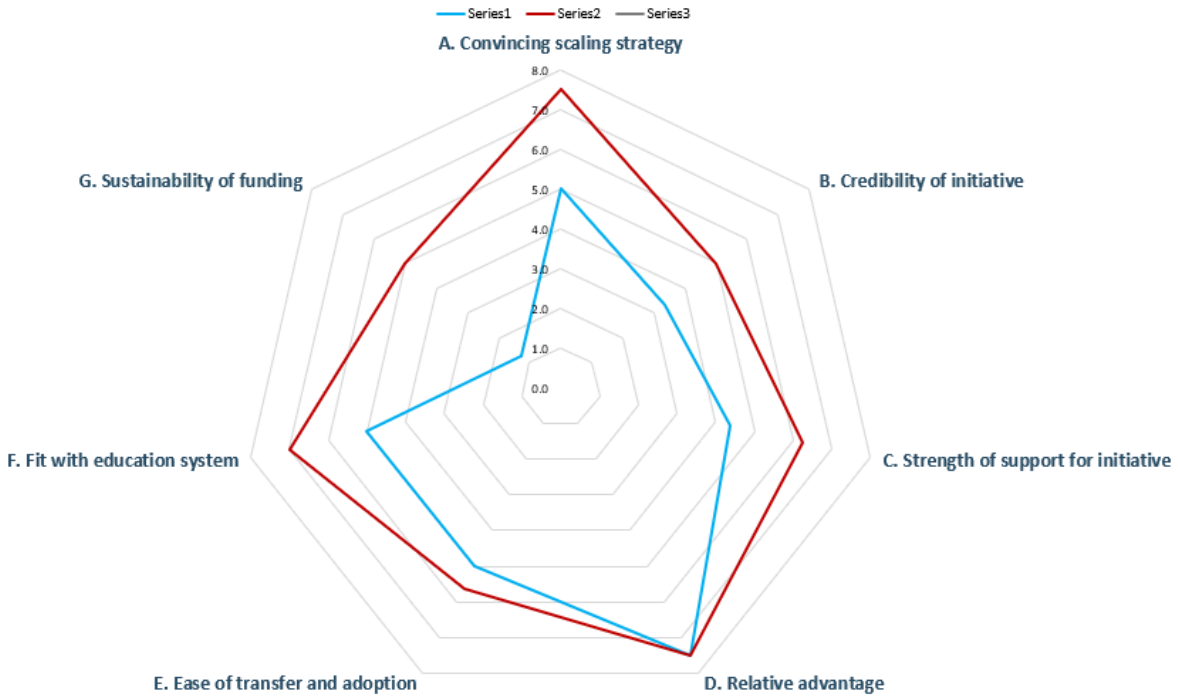


Figure 14: Scaling Checklist results example.

Landscape Analysis & Market Research

The DECILE product does not exist in a vacuum. There is a growing number of digital hearing or language assessment products on the market, from highly commercial products to low- or no-cost products with a strong research focus.

To underpin our work with DECILE on growth, we conducted a piece of Landscape Analysis and Market Research, to look at the design, functionality and business models of eight comparator products in both the Spanish- and English-speaking worlds.

This resource is designed to serve as a stimulus for reflection. Our rapid investigation yielded some insights, ideas and inspiration which could have value for DECILE as the team shape the path forward. The DECILE team could take this work further by investigating featured products in more depth and potentially contacting the most interesting organisations to learn more about their models and their journey to scale.

The full Landscape Analysis can be found in Annex 2 of this document. Here we cover some key highlights.

Headline insights from the Landscape Analysis

Having reviewed eight comparator products in depth, and looked briefly at many others, these were our headline reflections:

- We found multiple hearing assessment apps but **we didn't find a direct English-language equivalent for DECILE**.
- **There seem to be far more assessments focused on *how much children can hear*** (and using word understanding as a proxy measure) rather than on *which aspects of language* children have difficulty hearing.
- The assessments we found generally **require 1:1 adult supervision**.
- **We didn't encounter speech recognition or the use of AI** in any of the comparator products we found.
- **We didn't spot any apps that adjust accents** according to location.
- **The functionality, tech features and business models of some comparator products could have relevance** for DECILE and might provide valuable ideas or inspiration.

Specific ideas from the Landscape Analysis

Here is a brief overview of some of the specific ideas from other products, which DECILE could look at and replicate if appropriate:

Specific product ideas	Business model ideas
Adapt the assessment to the age and ability of the child, to ensure it is appropriate for all children. Draw on insights from mobile game technology to **add further gamification** into assessments. Trial and **adapt for children with Down's Syndrome**, who are particularly susceptible to glue ear. **Auto-assess the volume levels** in a room and recommend a move to somewhere quieter if it's too noisy. **Design the app for use with bone conduction headphones**, which have been shown to help children with glue ear. **Offer ideas for intervention**, linked to	

insights dashboard - blending assessment and pedagogical aims. **Use songs and audiobooks** for assessment purposes.	

Visual inspirations

Within the Landscape Analysis we also pulled out some visuals that illustrate interesting features of comparator products. Here is a small sample:

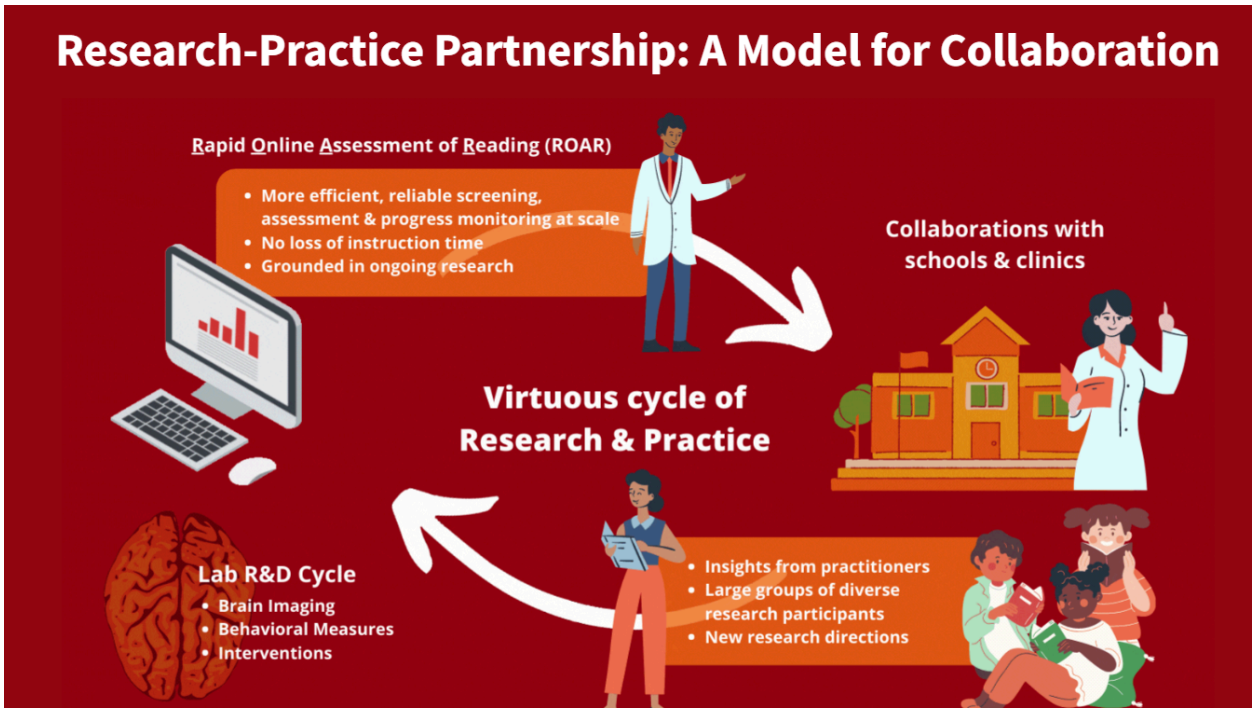


Figure 15. Roar's Iterative Research-Practice Model.

ROAR (from Stanford University) is a digital reading assessment with an 'open science' ethos. This graphic shows the iterative learning cycle they establish with schools and clinics who use their product.

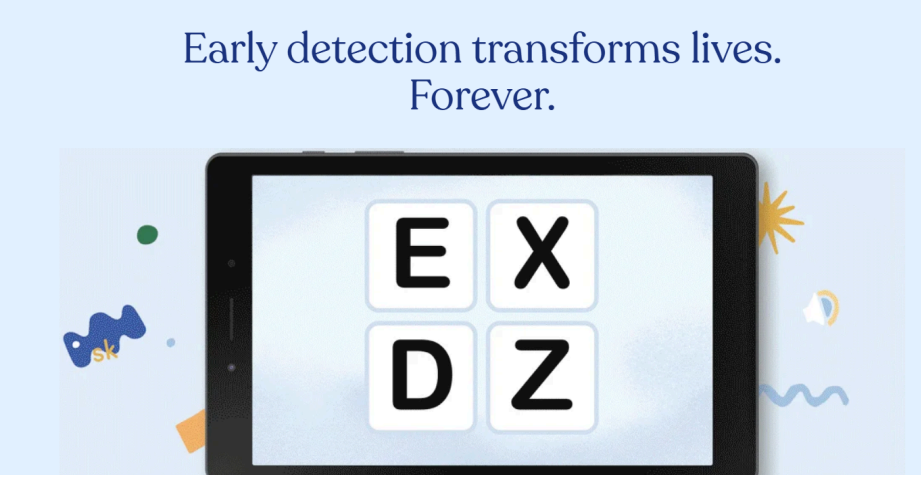


Figure 16: Early Bird's Messaging

Early Bird, an assessment tool that identifies reading challenges like dyslexia, uses simple, emotive messaging to communicate the value of their product.

Figure 17: Hear Glue Ear's gamification and use of songs and stories

Hear Glue Ear, a free app for families to help children with glue ear, showcase the child-friendly nature of their product, highlighting their use of songs, games & audiobooks.

Figure 18: Early Bird's User Dashboard

Early Bird's class-level dashboard has a simple, appealing, colour-coded interface for teachers using their product.

Deliverable 3 - Insights Dashboard

Goals of Insights Dashboards

We co-designed insights dashboards for the DECILE application to facilitate the effective and independent use of the application by teachers and school officials. The dashboards provide clear, actionable information about student language skills assessment results, helping educators make informed decisions about instructional strategies and interventions. Adding these dashboards is an important step towards scaling the application as they effectively close the feedback loop, enabling schools and teachers to utilize DECILE independently, without direct involvement from the DECILE research team. This empowers educators and school officials to proactively plan interventions, support student development, and take immediate action based on assessment data.

The dashboards were developed collaboratively with inputs from the DECILE team and teachers, ensuring alignment with best practices and real classroom needs. They represent an initial design phase and are subject to further iterations based on user feedback and real-world implementation. The current designs serve as practical guides for DECILE engineers to implement front-end user experiences and structure backend data systems.

Student List View Dashboard

The List View Dashboard offers a comprehensive overview for school officials and teachers, displaying essential performance metrics of students organized by groups or classes. Key features include performance averages (scores averaged across all assessments) as well as recent performance (date and results of the most recent assessments), with key assessment metrics (scores, percentage of skipped trials, and the number of reminders required which indicates the level of assistance provided during assessments. The list can be filtered by group or grade, and sorted by any column, enabling easy identification of students requiring immediate attention. This dashboard ensures teachers and school officials can quickly identify patterns, track progress over time, and prioritize interventions effectively.

The following tables show how the list view dashboard would be presented. Here we split it into two tables for ease of presentation but in the application it would be one larger horizontal table that contains both the overall average and the most recent information.

Show All	OVERALL AVERAGE	
Filter List by Group	Task 1 Sentence comprehension	Task 2 Question comprehension

Name	Group ⌄	Score	Reminders	Skipped	Score	Reminders	Skipped
Xxx Yyyy	Grade 5	20/30	3	3	5/20	0	3
Xxx Yyyy	Grade 4	12/30	0	0	3/20	1	5
Xxx Yyyy	Grade 3	21/30	1	2	15/20	0	2

Show All		MOST RECENT							
Filter List by Group		Task 1 Sentence comprehension				Task 2 Question comprehension			
Name	Group ⌄	Date	Score	Reminders	Skipped	Date	Score	Reminders	Skipped
Xxx Yyyy	Grade 5	5/20/2025	20/30	3	3	5/20/2025	5/20	0	3
Xxx Yyyy	Grade 4	5/10/2025	12/30	0	0	5/10/2025	3/20	1	5
Xxx Yyyy	Grade 3	3/25/2025	21/30	1	2	3/25/2025	15/20	0	2

Student In-Depth Dashboard

The in-depth student dashboard provides information on individual student profiles, including deeper insights into each student's performance and the ability to leave notes. This dashboard combines quantitative data with qualitative insights, empowering teachers and school officials to deliver personalized, data-driven educational support.

Based on initial rounds of feedback, we propose that the student in-depth dashboard include the following information:

- **Profile Information:** Name, age, grade, and details such as diagnosis, cochlear implant usage, training history with recorded voices, assessment frequency, and next reminder dates (a reminder email will be sent to perform an assessment).
- **Detailed Performance Visualizations (overall and sentence-level):** Graphical representations of performance across each assessment type, highlighting strengths and specific areas requiring support. Figure 19 presents an example of how detailed feedback can be provided using a bar plot.
    - Visualization with one bar for each type of sentence (SVO, OVS, Pas, RS and RO for the BOSQUE task and SVO, QuienSuj, QuienObj, QueSuj and QueObj for the SELVA task)
    - Visualization with number of reminders, number of skipped trials
- **Comparative Metrics:** Contextual benchmarks showing the student's performance relative to typically developing peers and other deaf children within similar hearing age groups and educational settings.

- Overall performance of the child in comparison to a group of typically developing children of the same hearing age for that task
- Overall performance of the child in comparison with other deaf children within the hearing age group (same grade/school)

- **Teacher and School Official Annotations:** Space for educators to leave interpretations, recommendations, interventions, and context-specific notes to enrich understanding and support collaborative decision-making.

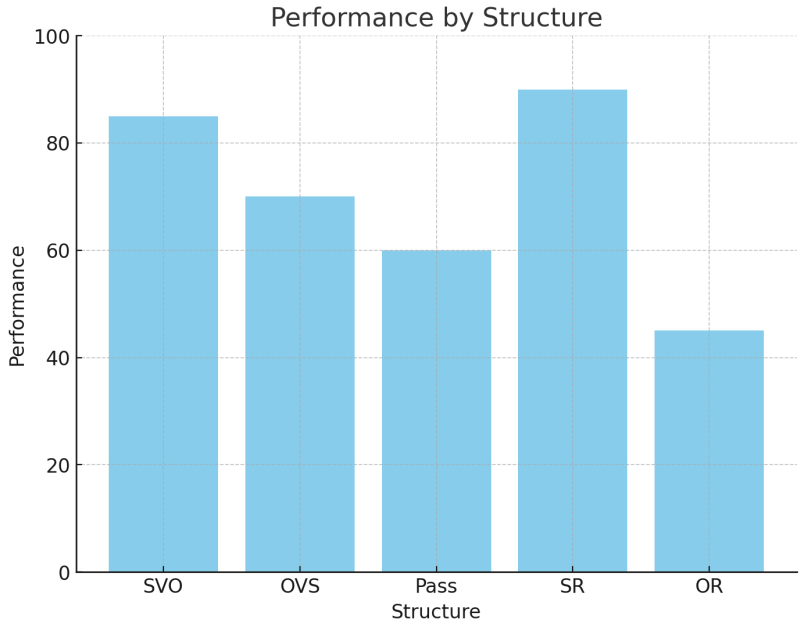


Figure 19: Example of Bar Plot for the Individual Student View Dashboard Showing the Performance of a Student on Each Type of Sentence.

Deliverable 4 - User Experience Flow

Creating a User Experience (UX) for Scale

The UX flows we developed for the DECILE application aim to create a scalable and intuitive user experience that supports independent and efficient use by different stakeholders: school officials, teachers, and students. These flows illustrate clearly defined navigation paths within the application, ensuring that users can effectively perform tasks without needing continuous guidance from the DECILE research team. These UX flows represent preliminary designs, developed with input from teachers and DECILE team members, and are subject to ongoing iteration based on user feedback and implementation testing.

To achieve larger scale implementation, the DECILE application will need to allow schools to use it independently and without direct oversight from the DECILE research team. The current version of the application (as of writing this report) supports conducting the actual assessment tasks, but it does not have the necessary user interfaces for setting up students, setting up teachers, viewing the results of the assessment, and analytics across students and over time. The dashboards in Deliverable 3 address the need for viewing assessment results and analytics. We worked together with the DECILE team to develop an overarching user structure and flows for each user category to enable the most important use cases for scale.

Overarching User Structure

Figure 20 illustrates an example of the user structure for DECILE's implementation, highlighting roles and interactions between school officials, teachers, and students. This exemplary structure clarifies user roles, responsibilities, and interdependencies to streamline system implementation. School officials are expected to set up teacher and student profiles. Students can be assigned to teachers to streamline assessment procedures and analytics. School officials and teachers have access to assessment data and analytics. Student profiles are set up so that assessments can be performed without the need to enter user profile information for each student, streamlining the assessment process that enables in-class assessments with a larger group of students.

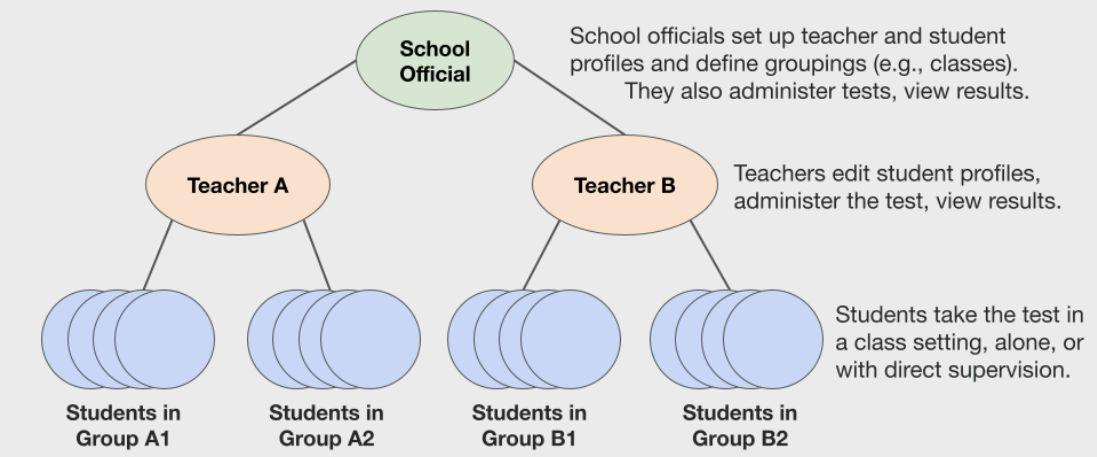


Figure 20: Example of Simple User Model Structure.

To allow for multiple school officials within a school and more flexible assignment of students into groups, the user structure we propose a concrete data model in Figure 21 that has schools as the top-level entity (organizations table) with multiple members of the organization, including school officials and teachers (users table). Individual students (students table) are then grouped into one or more groups (these could be classes but also groups defined by special needs, including a group of just one student; groups table), and any number of teachers and/or school officials would be assigned to a group of students (relations table). This type of mapping would provide the highest level of flexibility, and it can be implemented efficiently with any type of SQL database (e.g., Postgres). A single relations table in the data model can hold all types of connections between organizations, users, students, and groups.

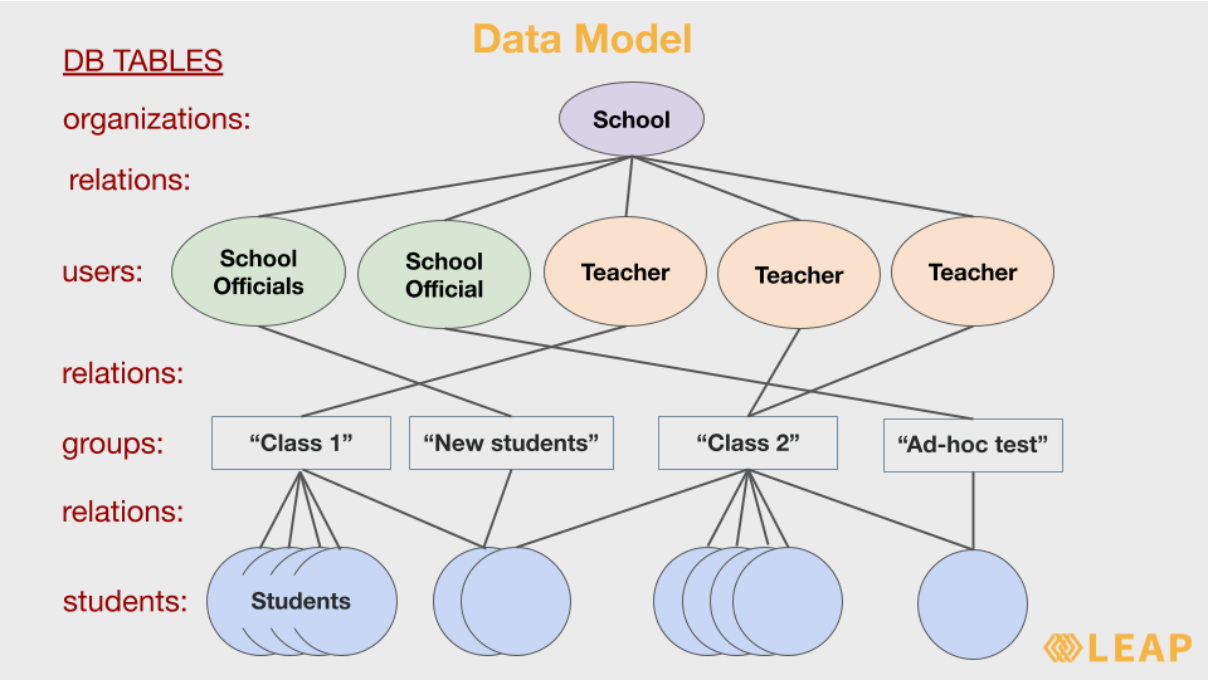


Figure 21: Data Model with Suggested Database Table Names.

UX Flow for School Officials

Figure 22 illustrates the UX flow of the DECILE application for school officials. School officials use the DECILE application primarily for administrative oversight and management of user profiles. Their accounts are created by the DECILE team, and they have the ability to create and manage teacher and student profiles within their organization, assign students to groups or classes, and designate specific assessment activities. School officials also have access to dashboards that aggregate assessment results across classes or school-wide, allowing them to monitor overall student progress and performance.

By having a clear view of both individual and collective student data, school officials can identify broader trends, strategically allocate resources, and support teachers in implementing targeted educational interventions. This flow ensures that school officials maintain high-level oversight and can effectively coordinate the use of DECILE across the institution.

Figure 23 shows the wireframe for the login page, landing screen with the main navigation buttons, and the screen for managing student profiles. These will be the same or very similar for teachers and school officials (depending on user feedback, teacher access to creating new teacher or student profiles, and new groups, could be restricted; teacher access could also be restricted to only see information for students assigned to them via groups).

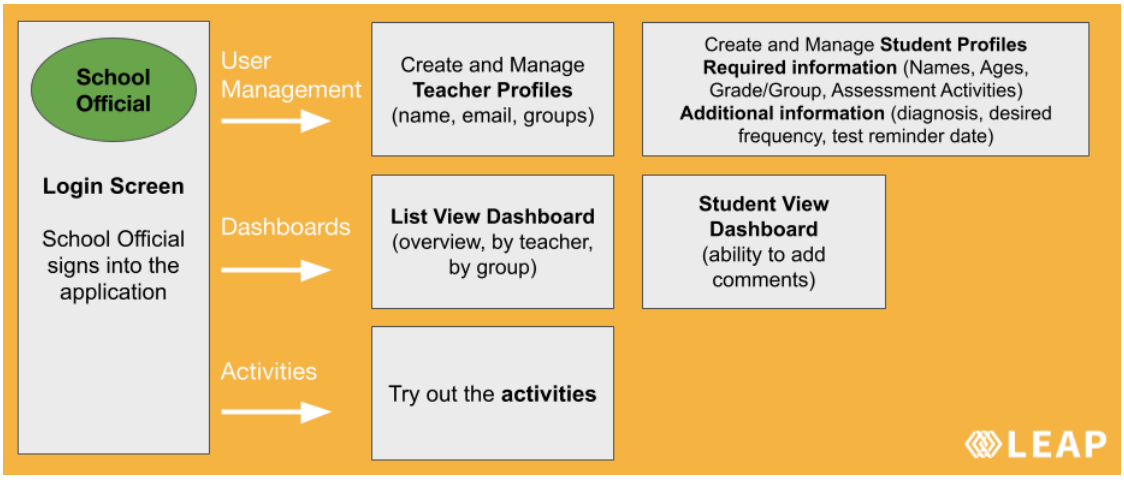


Figure 22: How School Officials will use the DECILE application.

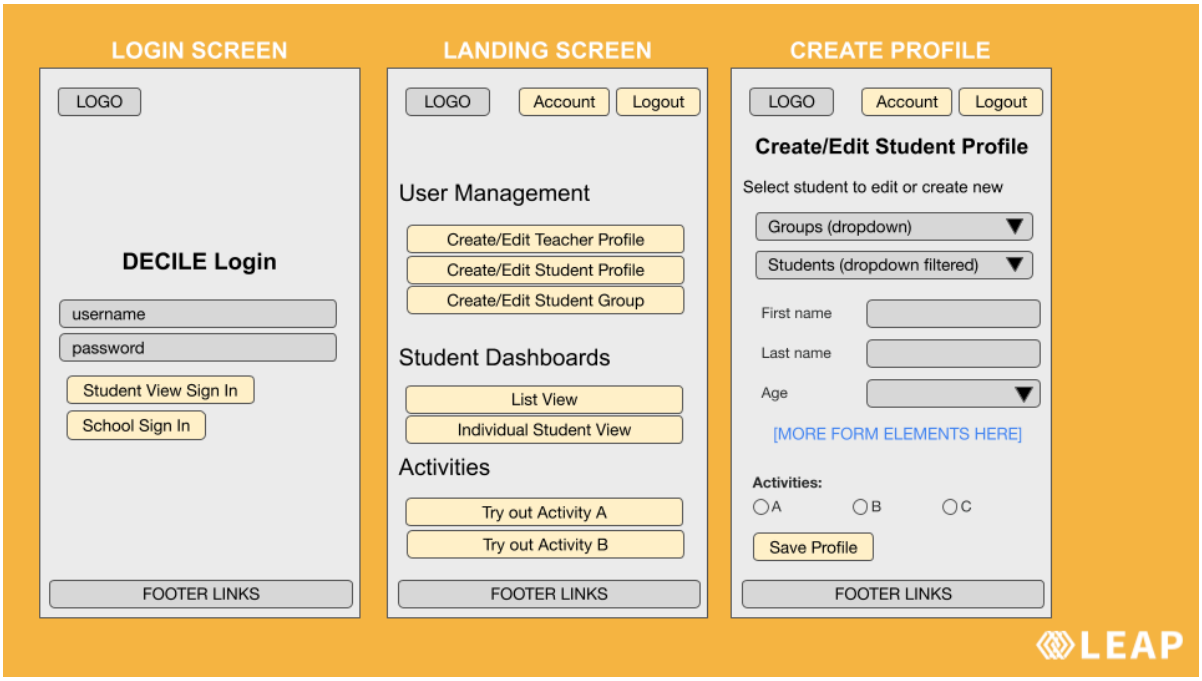


Figure 23: Wireframes for Key Pages of the School Official and Teacher Flow.

UX Flow for Teachers

Figure 24 illustrates the UX flow of the DECILE application for teachers. Teachers interact with the DECILE application to support their instructional responsibilities. Upon logging into the teacher view of the application (which is almost identical to the school official view), they can manage student profiles, including editing necessary and supplementary information such as diagnostic details, frequency of assessments, and additional contextual data. Teachers have the flexibility to select individual students or groups for assessment and configure assessment settings according to the educational context (group, 1:1, or individual sessions).

To help teachers efficiently run an in-class assessment, they can set up all student details in advance so that they can quickly set up individual devices (laptops or tablets) for each student by simply selecting each student profile from a list (see UX Flow for Students).

Teachers can use the list view and student view dashboards to track student assessment performance, assess progress in language skills, and plan immediate or future interventions based on detailed analytics. By enabling teachers to manage these tasks independently, DECILE significantly enhances their ability to tailor instruction and respond swiftly to student needs.

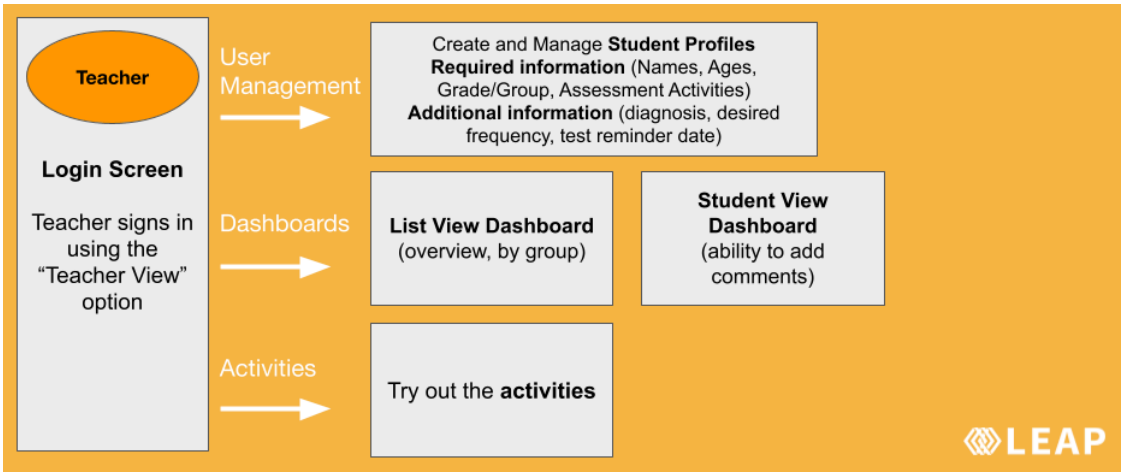


Figure 24: How Teachers will use the DECILE application.

UX Flow for Students

Figure 25 illustrates the UX flow of the DECILE application for students with set-up assistance by teachers (and also school officials for 1-1 testing outside the classroom). The student flow starts with the teacher setting up the device used by an individual student for the assessment. This is a simplified, user-friendly flow to reduce the time it takes to set up a device for the assessment, considering that a teacher will need to do this for each student in the class that takes the assessment. The UX flow begins with logging into a student-specific view using the teacher credentials. They then see a list of groups/classes available to them. Upon selecting a group, they see a list of all students associated with that group and select one. Once a student is selected, the interface shows profile information for that student, the pre-selected set of assessment activities for that student, and lets the teacher confirm the assessment context (in-group, one-on-one, or self-directed assessment). The teacher can make changes if necessary but the goal is for this to be a quick confirmation step because the details would have been populated by the teacher or school official in advance. Once confirmed, the application shows a ready screen. Students will be given the device and told to press the start button, which will begin the selected assessment. Figure 26 shows the wireframe design for these pages to guide the implementation.

After completing assessments, all students receive immediate positive feedback, reinforcing their engagement and confidence. This streamlined experience encourages active participation, ensures clarity of purpose, and helps students feel comfortable during assessments, ultimately contributing to more accurate and representative performance data.

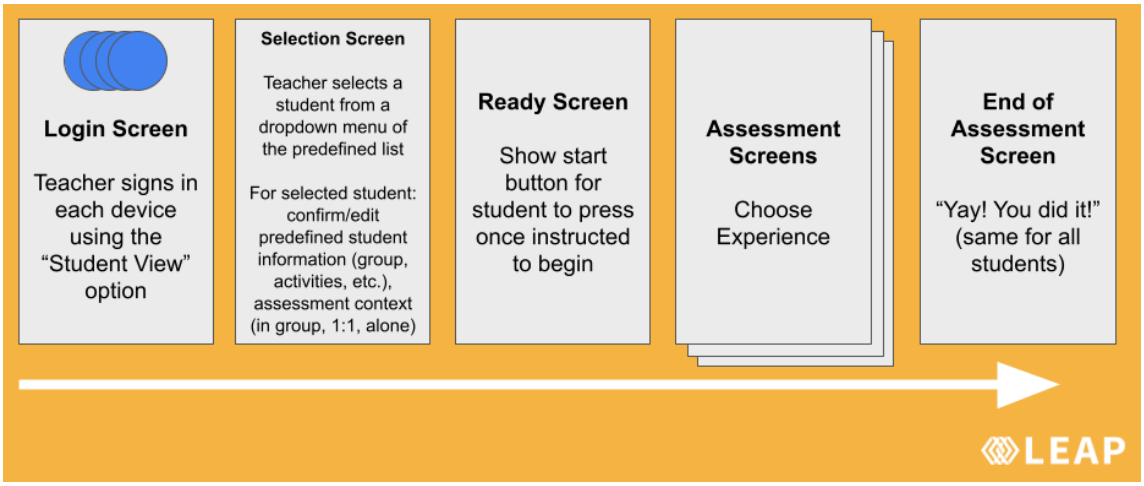


Figure 25: How Students will use the DECILE application with Teacher Setup.

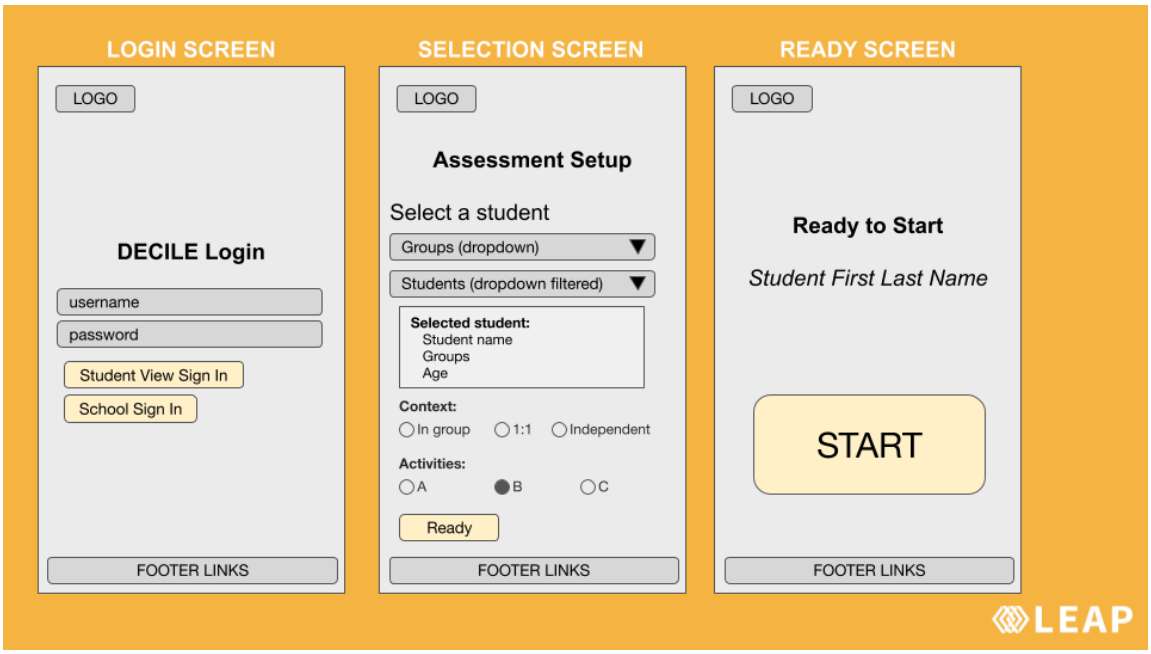


Figure 26: Wireframe for Student Flow Screens.

References

American Speech-Language-Hearing Association (2025). Assessment Tools, Techniques, and Data Sources. https://www.asha.org/practice-portal/resources/assessment-tools-techniques-and-data-sources/

Barrington, E., Sarkisian, S. M., Feldman, H. M., & Yeatman, J. D. (2023). Rapid Online Assessment of Reading (ROAR): evaluation of an online tool for screening reading skills in a developmental-behavioral pediatrics clinic. Journal of Developmental & Behavioral Pediatrics, 44(9), e604-e610.

Bialystok, E., Luk, G., Peets, K. F., & Yang, S. (2010). Receptive vocabulary differences in monolingual and bilingual children. Bilingualism: Language and Cognition, 13(4), 525-531.

Clark, L. A., & Watson, D. (2016). Constructing validity: Basic issues in objective scale development.

Gijbels, L., Burkhardt, A., Ma, W. A., & Yeatman, J. D. (2024). Rapid online assessment of reading and phonological awareness (ROAR-PA). Scientific Reports, 14(1), 10249.

Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). Using student achievement data to support instructional decision making (NCEE 2009-4067). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practiceguides/.

Jang, E. E., Wagner, M., Hannah, L., & Kim, H. (2024). Diagnostic Assessment Feedback in the 21st-Century Technology-Rich Classroom. The Concise Companion to Language Assessment, 339.

Jasińska, K. K., Akpe, Y. H., Seri, B. A. D., Zinszer, B., Agui-Kouadio, R. Y., Mulford, K., Curran, E., Ball, M.-C., & Tanoh, F. (2022). Evaluating Bilingual Children's Native Language Abilities in Côte d'Ivoire: Introducing the Ivorian Children's Language Assessment Toolkit for Attié, Abidji, and Baoulé. Applied Linguistics, 43(6), 1116-1142. https://doi.org/10.1093/applin/amac025

Knekta, E., Runyon, C., & Eddy, S. (2019). One Size Doesn't Fit All: Using Factor Analysis to Gather Validity Evidence When Using Surveys in Your Research. CBE life sciences education, 18(1), rm1. https://doi.org/10.1187/cbe.18-04-0064

Lillo-Martin, D., Koulidobrova, H., de Quadros, R. M., & Pichler, D. C. (2012). Bilingual language synthesis: Evidence from WH-questions in bimodal bilinguals. In Proceedings of the 36th Annual Boston University conference on language development (Vol. 302, pp. 302-314). Cascadilla.

Mather, N. & Wendling, B.J. (2014). Examiner's manual. *Woodcock-Johnson IV Tests of Cognitive Abilities*. Rolling Meadows, IL: Riverside Publishing.

McGrew, K. S., LaForte, E. M. & Schrank, F. A. (2014). Technical manual. W*oodcock-Johnson IV*. Rolling Meadows, IL: Riverside Publishing.

Miller, L. J. (Ed.). (2020). Developing norm-referenced standardized tests. Routledge.

Muñoz-López, J., Fresneda, M. D., Mendoza-Lara, E., & Carballo, G. (2008). Psychometric properties of a grammatical comprehension test. RN, 47(1), 21-26. https://doi.org/10.33588/rn.4701.2008037

Portney LG, Watkins MP (2000) Statistical measures of reliability. In: Foundations of clinical research: applications to practice, 2nd edn. Prentice Hall, New Jersey, pp 557–584

Sobers, S. M., Whitehead, H. L., N'Goh, K. N. A., Ball, M. C., Tanoh, F., Akpé, H., & Jasińska, K. K. (2023). Is a Phone-Based Language and Literacy Assessment a Reliable and Valid Measure of Children's Reading Skills in Low-Resource Settings? Reading Research Quarterly, 58(4), 733-754. https://doi.org/10.1002/rrq.511

Wells CS. Item Response Theory. In: Assessing Measurement Invariance for Applied Research. Educational and Psychological Testing in a Global Context. Cambridge University Press; 2021:108-160.

Yeatman, J. D., Tang, K. A., Donnelly, P. M., Yablonski, M., Ramamurthy, M., Karipidis, I. I., ... & Domingue, B. W. (2021). Rapid online assessment of reading ability. Scientific reports, 11(1), 6396.

Yeatman, J. D., Tran, J. E., Burkhardt, A. K., Ma, W. A., Mitchell, J. L., Yablonski, M., ... & Richie-Halford, A. (2024, December). Development and validation of a rapid and precise online sentence reading efficiency assessment. In Frontiers in education (Vol. 9, p. 1494431). Frontiers Media SA.

Appendix

Annex 1 - Additional Reading and References in Deliverable 1

Language Assessment in Deaf and Hard of Hearing (DHH) Children:

Graham, P. J., & Shuler-Krause, E. (2020). Building strong foundations for educational achievement: Language assessments in early childhood education for deaf and hard of hearing children. *Psychology in the Schools*, *57*(3), 418-425. https://doi.org/10.1002/pits.22333

Pizzo, L., & Chilvers, A. (2019). Assessment of Language and Literacy in Children Who Are d/Deaf and Hard of Hearing. *Education Sciences*, *9*(3), 223. https://doi.org/10.3390/educsci9030223

Qi, S., & Mitchell, R. E. (2011). Large-Scale Academic Achievement Testing of Deaf and Hard-of-Hearing Students: Past, Present, and Future. *The Journal of Deaf Studies and Deaf Education*, *17*(1), 1-18. https://doi.org/10.1093/deafed/enr028

Language Assessment in Bilingual Children (with relevance for bimodal bilingual children):

Gathercole, V. C. (2010). Bilingual children: Language and assessment issues for educators. International handbook of psychology in education, 1.

Freeman, M. R., & Schroeder, S. R. (2022). Assessing language skills in bilingual children: Current trends in research and practice. Journal of Child Science, 12(01), e33-e46.

De Lamo White, C., & Jin, L. (2011). Evaluation of speech and language assessment approaches with bilingual children. International Journal of Language & Communication Disorders, 46(6), 613-627.

Language Development in Deaf and Hard of Hearing (DHH) Children:

Nelson, N. W., & Crumpton, T. (2015). Reading, Writing, and Spoken Language Assessment Profiles for Students Who Are Deaf and Hard of Hearing Compared With Students With Language Learning Disabilities. *Topics in Language Disorders*, *35*(2), 157-179. https://doi.org/10.1097/tld.0000000000000055

Lederberg, A. R., Schick, B., & Spencer, P. E. (2013). Language and literacy development of deaf and hard-of-hearing children: Successes and challenges. *Developmental Psychology*, *49*(1), 15-30. https://doi.org/10.1037/a0029558

Annex 2 - Landscape Analysis & Market Research

DECILE: Landscape Analysis & Market Research

Introduction

To underpin our work with DECILE on growth, we conducted a piece of landscape analysis and market research, to look at the design, functionality and business models of eight comparator products in both the Spanish- and English-speaking worlds.

This resource is designed to serve as a stimulus for reflection. Our rapid investigation yielded some insights, ideas and inspiration which could have value for DECILE as the team shape the path forward. The DECILE team could take this work further by investigating featured products in more depth and potentially contacting the most interesting organisations to learn more about their models and their journey to scale.

Overarching reflections

- We found multiple hearing assessment apps but we didn't find a direct English-language equivalent for DECILE.
- There seem to be far more assessments focused on *how much* children can hear (and using word understanding as a route to that insight) rather than on *which aspects of language children have difficulty hearing*.
- The assessments we found generally require 1:1 adult supervision.
- We didn't spot any that required the app to recognise speech, or any that employed AI.
- We didn't spot any apps that adjust accents according to location.
- DECILE could draw valuable ideas and inspiration from the functionality, tech features and business models of comparator products. See below for full details. These ideas include:

Specific product ideas	**Business model ideas**
Adapt the assessment to the age and ability of the child, to ensure it is appropriate for all children.	Consider funding partnerships that allow the product to be used for free by schools.

Draw on insights from mobile game technology to add further gamification into assessments. Trial and adapt for children with Down's Syndrome, who are particularly susceptible to glue ear. Auto-assess the volume levels in a room and recommending a move to somewhere quieter if it's too noisy. Design the app for use with bone conduction headphones, which have been shown to help children with glue ear. Offer ideas for intervention, linked to insights dashboard - blending assessment and pedagogical aims. Use songs and audiobooks for assessment purposes.	If offering DECILE free to schools, consider setting up a formalised research-practice partnership as part of that arrangement. Consider the searchability of the product name: as a common English noun, and a Spanish verbal phrase, is DECILE too generic to be easily searchable online? Create a welcoming website for the non-specialist; avoid making the language and graphics too technical. Use strong, emotive language to make the value of the product clear. Share information on reach and number impacted, as DECILE grows, to communicate credibility.

Landscape Analysis: A review of five comparable products

Basic info *Product name* *Website* *Location (where based)* *Language(s)*	**Purpose** *What it's for* *How it works* *How it describes itself*	**Size/reach info** *Potential customers* *How many* *How widespread* *Cost*	**Tech features** *Use of AI* *Use of user profiles* *Other effective tech use*	**Insights for DECILE**
Hear Glue Ear - App https://www.cambridgedigitalhealth.co.uk/hear-glue-ear-app Developed in the UK by Cambridge Digital Health: "Specialists in developing apps for health and research" Positive independent evaluation here; lots of additional info.	"A free, award-winning app, designed for children experiencing hearing loss due to glue ear" "A valuable tool to help families manage their child's glue ear at home." Includes: Songs, audiobooks and games that function both as pedagogical tools and ongoing assessment. A regular hearing screen, presented to children as a game (to provide additional data to clinicians, not to be	Free UK focused Customer: families and healthcare practitioners. Unclear, but potentially large customer base: being used to save costs and improve treatment by the National Health Service. 1K+ downloads on Android Store.	Tech features for accessibility: (from this NHS page) The app can be used on a mobile phone or tablet. It can assess the volume levels in a room and recommend moving somewhere quieter if it's too noisy Parents can understand how easily their child can hear speech with single-word speech tests. The child listens to a word at different volumes and presses the correct picture. This helps parents	Accessibility feature ideas Blending of assessment & pedagogical aims Use of songs and audiobooks for assessment purposes.

	used as a robust diagnostic tool). Up to date info for parents. A clinician's portal for providing insights on progress between appointments.		understand if their child cannot hear certain words. The app can be used with bone conduction headphones which have been shown to help children with glue ear. Speech and language therapists can also upload and deliver individual recorded sessions to a particular child's app if needed. Specifically trialled and adapted for children with Down's Syndrome too.	
Language Screen - App https://oxedandassessment.com/languagescreen/ Developed in the UK by Oxford Uni spin-out company who "create educational assessment apps and interventions	"An assessment tool for 3-11 year olds, to track the progress of pupils' oral language skills and identify those who would benefit from interventions or specialist support." 1-1 screening tool (to spot issues that might go unnoticed):	UK focused. Paid-for product - cost unclear (have to sign up for free trial). For use in school by educators. Used with >350k children.	Phone or tablet app - recently added web browser option too. Developing 'Progress Reports' to track individual students' performance over time. Accessibility features: "To ensure the test is	Good reporting functionality. Very quick to administer. Intervention program for teachers to use that focuses on the learning needs of individual students based on NELI assessment

which improve educational outcomes for children." Independent evaluation here - very positive. "LanguageScreen provides an easy-to-use, reliable, child-friendly means of identifying children with language difficulties."	3-4 tests of fundamental language skills - takes <10mins. "The adult will guide the child through the tests, following the instructions provided in the app, and discreetly mark their response correct or incorrect using the controls at the bottom of the screen." Provides immediate reporting to educators (sample here). No pedagogical recommendations for educators.	1k+ downloads on Android Store.	appropriate for all children, the LanguageScreen assessment adapts to the age and ability of the child." App Dashboard	
Sound Scouts - App https://www.soundscouts.com/en-gb/ Developed in Australia. Positive evaluation here (paywall).	"Sound Scouts mission is to help eliminate undetected hearing loss in children. The aim is to provide an accessible hearing test solution for school aged children and adults across Australia & around the world."	Australia focused Paid-for product: Testing credits are purchased in-app for £22.99 (subject to changes from Store provider). Each credit allows a Player / Client (person being tested)	Focus on "mobile game technology".	Very specific assessment, being run as a business. Centrality of gamification to the assessment.

	"Designed to make testing easy by incorporating the science of a hearing test into a fun mobile game." Checking for: 1.Conductive hearing loss (middle/outer ear issues) 2.Sensorineural hearing loss (inner ear issues) 3.Difficulties listening in noise (potentially caused by processing issues, attention or language issues including EAL/D) 1-1 adult and child (age 4-18). 6-8 mins to cover 3 tests. Auto-generated report.	access to four test sessions. Schools, clinics and organisations can access bulk purchasing discounts with a minimum pack size of 10. 10k+ downloads on Android Store.		
Rapid Online Assessment of Reading (ROAR) Developed in USA - Stanford University Highly detailed ROAR overview doc here - outlining deliberate	"investigate the factors contributing to reading difficulties including dyslexia. By developing and rigorously validating automated assessment tools that enable large-scale data collection, and bridge	K-12 students (validated with 20,000, across 20 US states) Free - but as part of a formalised research-practice partnership. Schools request access on a	Adaptive. Validation, reliability, and item response theory work can be a useful model for Decile.	Measures sentence and morphological processing as part of the suite of assessments. The Spanish language version is in use in Colombia.

'open science' research-practice ideology Teachers' Administration Guide Family Guide (context etc) Recommendations for supporting students who need extra support	research and practice, we can help researchers and educators understand and support the diversity of learners" Currently working to validate other assessments: morphology, working memory, rapid visual processing, picture vocab, written vocab. Ongoing iterative process (see graphic of cycle below).	Partner Interest Form. Whole class can be assessed at once (one child per computer, with headphones) - children complete independently in 12-15 mins		Similarly developed in university lab, and then scaled to schools. This model is free. Example of a particularly research-centric approach to scaling.
EarlyBird Education A reading assessment - developed and validated at Boston Children's Hospital in partnership with faculty at the Florida Center for Reading Research.	Screener for dyslexia "The EarlyBird program uses a game-based assessment to identify potential reading challenges like dyslexia, then gives schools and families the tools they need to intervene." For use by teachers and parents - with the parental version spun out and rebranded as Sprout Labs.	Two researchers teamed up with a business executive to launch EarlyBird as a commercial venture in 2020. Acquired by Imagine Learning in 2024, a digital curriculum provider with a broad reach. Used "by schools in 20 states, by tens of thousands of students and their teachers."	Gamified app Online dashboard area - focus shifting from 'identification to intervention' Dashboard written overview Dashboard walkthrough video	Similarly developed in a university lab, and then scaled to reach families and schools. May be a useful model for scaling approaches. Teacher dashboards, professional development workshops, evidence-based lesson plans, etc. Now a commercial product.

	Presenting the assessment as part of a whole suite of ongoing support and resources. Along with insight, schools receive “Just-in-time evidence-based lesson plans targeting specific skills appropriate for all teachers with any level of experience.”	Cost to schools not advertised - instead schools submit details to arrange a sales call. Paid: $149 per month for at home use. Package for families: -Access to the Pip School instructional app -EarlyBird Assessment 3 times per year -One-on-one parent/caregiver meetings with a Reading Specialist -Monthly progress monitoring check-ins with a Reading Specialist (with your child) -Weekly progress reports sent to your inbox -Access to Parent Resources, which includes resources and at home activities -Exclusive access to Sprout Labs Program events		Example of a particularly commercial approach to scaling - and of a particularly intervention-focused product.

Landscape Analysis: Images for interest or inspiration

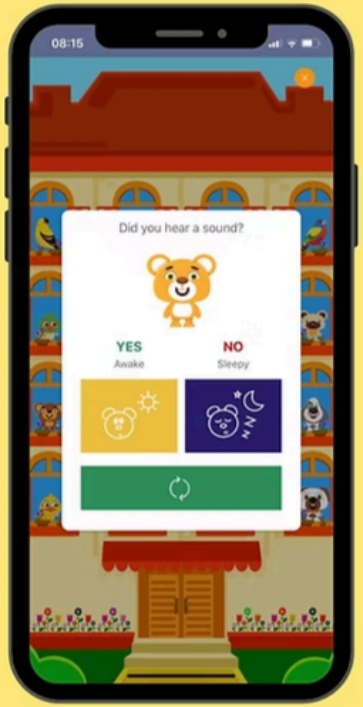
Hear Glue Ear

Use of songs and audiobooks within assessments:

LEAP

The hearing screen* is presented as a fun game that a child can play regularly. It will provide an insight into how the child's hearing is fluctuating over time.

The results show if there are areas within normal hearing range where a child may not be hearing well, and are presented with examples of sounds that might be difficult (e.g whispered conversation level).
Parents and clinicians will be able to see how stable a child's hearing is, and whether appointments are required more regularly.



Language Screen

Assessment overviews:

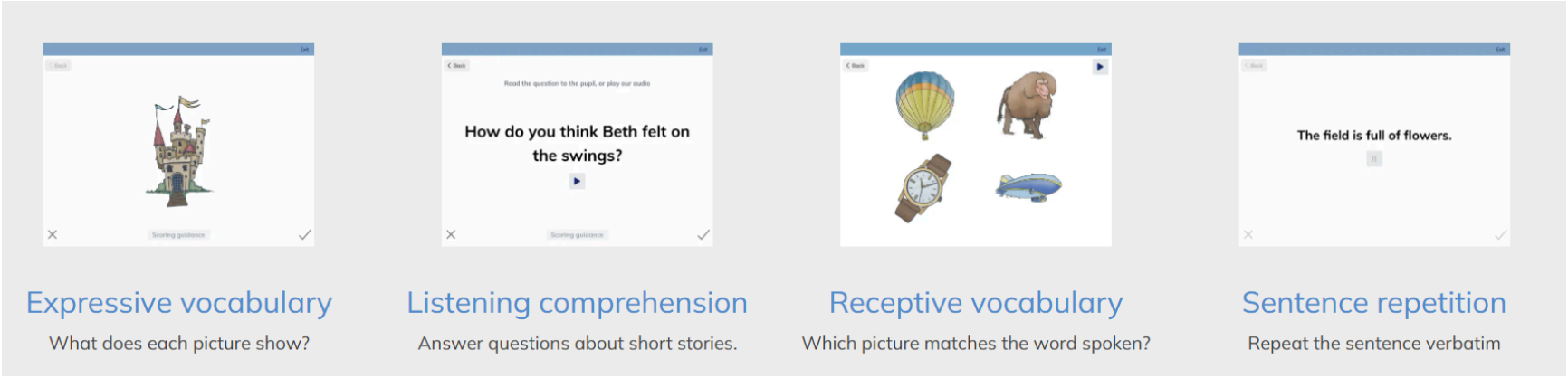


Table 1. Items and distractors in the Receptive Vocabulary subtest of LanguageScreen.

Item	Target	Phonological distractor	Semantic distractor	Unrelated distractor
1	balloon	baboon	airship	watch
2	telephone	xylophone	computer	bear
3	yo-yo	dodo	shuttlecock	heart
4	shell	shed	crab	trumpet
5	microphone	microscope	speaker	basket

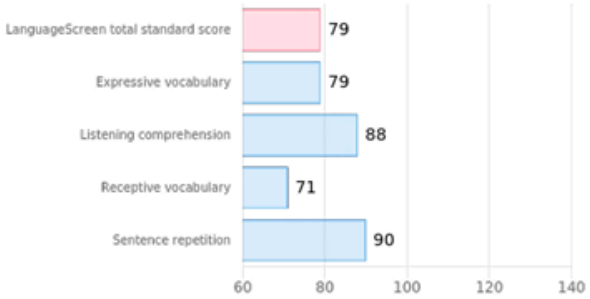
Dashboard for individual:

LanguageScreen assessment results for Pearl Alexander

1A, Year 1, 5yr 2mo

Date/time assessed	Age at assessment	Assessment language	Expressive vocabulary	Listening comprehension	Receptive vocabulary	Sentence repetition	LanguageScreen total standard score	LanguageScreen total percentile rank
02/10/2023 10:22	5yr 1mo	English (UK)	79 (RS: 7)	88 (RS: 6)	71 (RS: 9)	90 (RS: 7)	79	9

LanguageScreen Profile for Pearl Alexander



Pearl's LanguageScreen Total standard score is 79. This equates to a percentile rank of 9, a score that places Pearl within the bottom 9% of children in their age group.

Pearl's language skills are a clear cause for concern; Pearl should be given language intervention if possible.

The graph shows the LanguageScreen Total standard score as well as the differences between standard scores on the subtests obtained by Pearl. These differences may reflect many factors, including lapses of attention, and should not be overinterpreted.

Pearl's highest scores is for Sentence Repetition, a test that assesses the ability to repeat sentences and that is sensitive to grammatical difficulties. Pearl's lowest scores is for Receptive Vocabulary, a test that taps vocabulary understanding as assessed by the ability to match spoken words to pictures.

Expressive Vocabulary (EV) assesses the ability to name pictures; a measure of vocabulary knowledge (24 items).
Listening Comprehension (LC) assesses the ability to understand spoken stories; a measure of both literal and inferential language comprehension and expressive language skills (3 stories, and a total of 16 questions).
Receptive Vocabulary (RV) assesses the ability to match spoken words to pictures; a measure of vocabulary understanding (23 items).
Sentence Repetition (SR) assesses the ability to repeat sentences; a measure of language comprehension and production skills (14 items).

Raw Scores (RS) are simply the number of correct answers on each subtest. Raw scores are hard to interpret because different subtests have different number of questions, and the expected score on each subtest will increase as a child gets older.
The pattern of **standard scores** for each child across the different LanguageScreen subtests shows their profile (what subtests they found easier and what subtests they found harder). However, differences between subtests may reflect many factors, including lapses of attention, and should not be overinterpreted.

Dashboard for group:

Current Year 1 summary

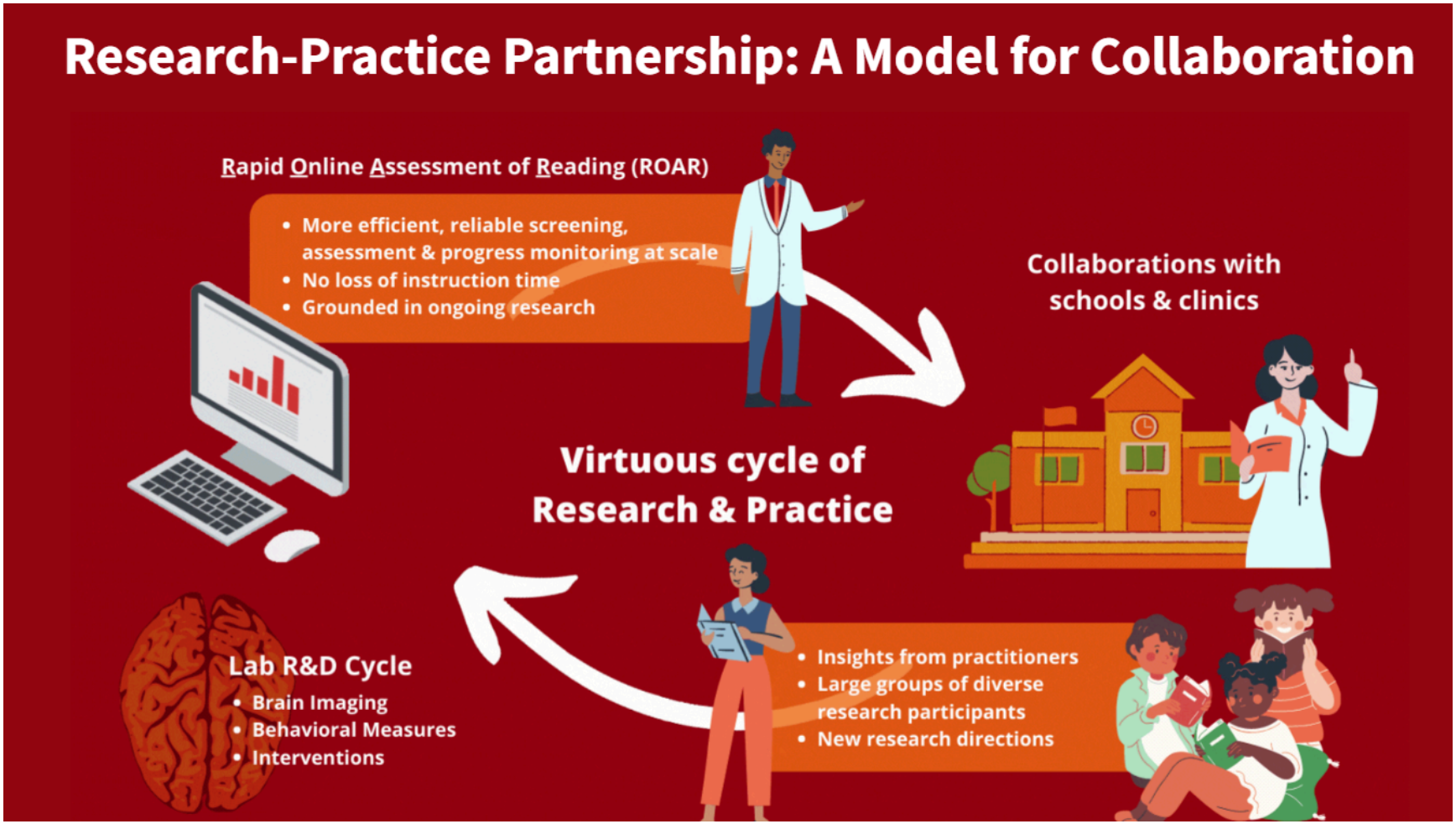
First names	Surname	DoB	Current class	Assessment date	Age at assessment	Standard score (average 100)	Percentile	Description	Alert	Latest intervention	Completion date
Tessa	Vega	22/12/2017	1A	02/10/2023	5yr 9mo	65	1	clear concerns			
Richard	Whitaker	08/09/2018	1A	02/10/2023	5yr 0mo	65	1	clear concerns			
Rubie	Anderson	15/07/2018	1A	02/10/2023	5yr 2mo	75	5	clear concerns			
Pearl	Alexander	08/08/2018	1A	02/10/2023	5yr 1mo	79	9	clear concerns			
Bruce	Bowers	22/02/2018	1A	02/10/2023	5yr 7mo	83	14	slight concerns			
Walter	Phillips	15/02/2018	1A	02/10/2023	5yr 7mo	86	19	slight concerns			
Eddy	Riley	15/06/2018	1A	02/10/2023	5yr 3mo	86	19	slight concerns			
Sharise	Ballard	08/04/2018	1A	02/10/2023	5yr 5mo	88	22	slight concerns			
Franklin	Bentley	22/06/2018	1A	02/10/2023	5yr 3mo	92	32	no concerns			
Lucia	Perry	15/12/2017	1A	02/10/2023	5yr 9mo	101	55	no concerns			
Arron	Nelson	08/02/2018	1A	02/10/2023	5yr 7mo	104	61	no concerns			
Jacques	Ochoa	22/09/2018	1A	02/10/2023	5yr 0mo	104	62	no concerns			
Livia	Cameron	22/04/2018	1A	02/10/2023	5yr 5mo	104	62	no concerns			
Ayla	Ortega	08/01/2018	1A	02/10/2023	5yr 8mo	105	65	no concerns			
Jennifer	Owen	08/07/2018	1A	02/10/2023	5yr 2mo	108	72	no concerns			
Kayla	Mullins	22/01/2018	1A	02/10/2023	5yr 8mo	110	75	no concerns			
Carla	Clark	15/01/2018	1A	02/10/2023	5yr 8mo	113	81	no concerns			

Sound Scouts

Child-friendly design and graphics:

ROAR

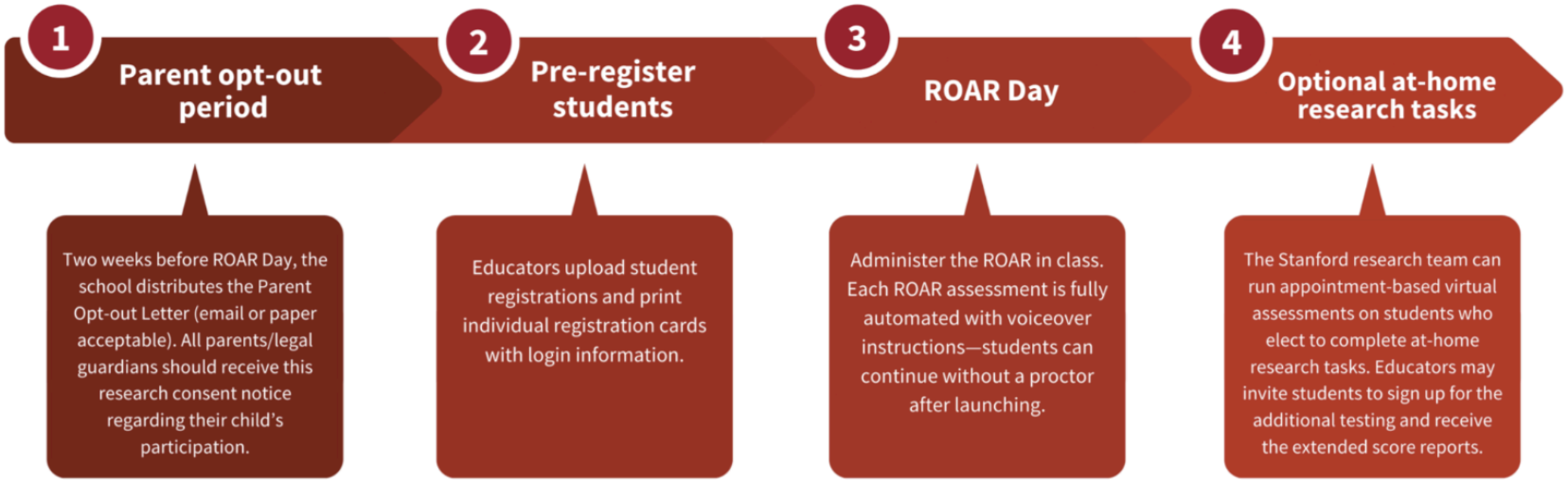
Iterative research-practice cycle:



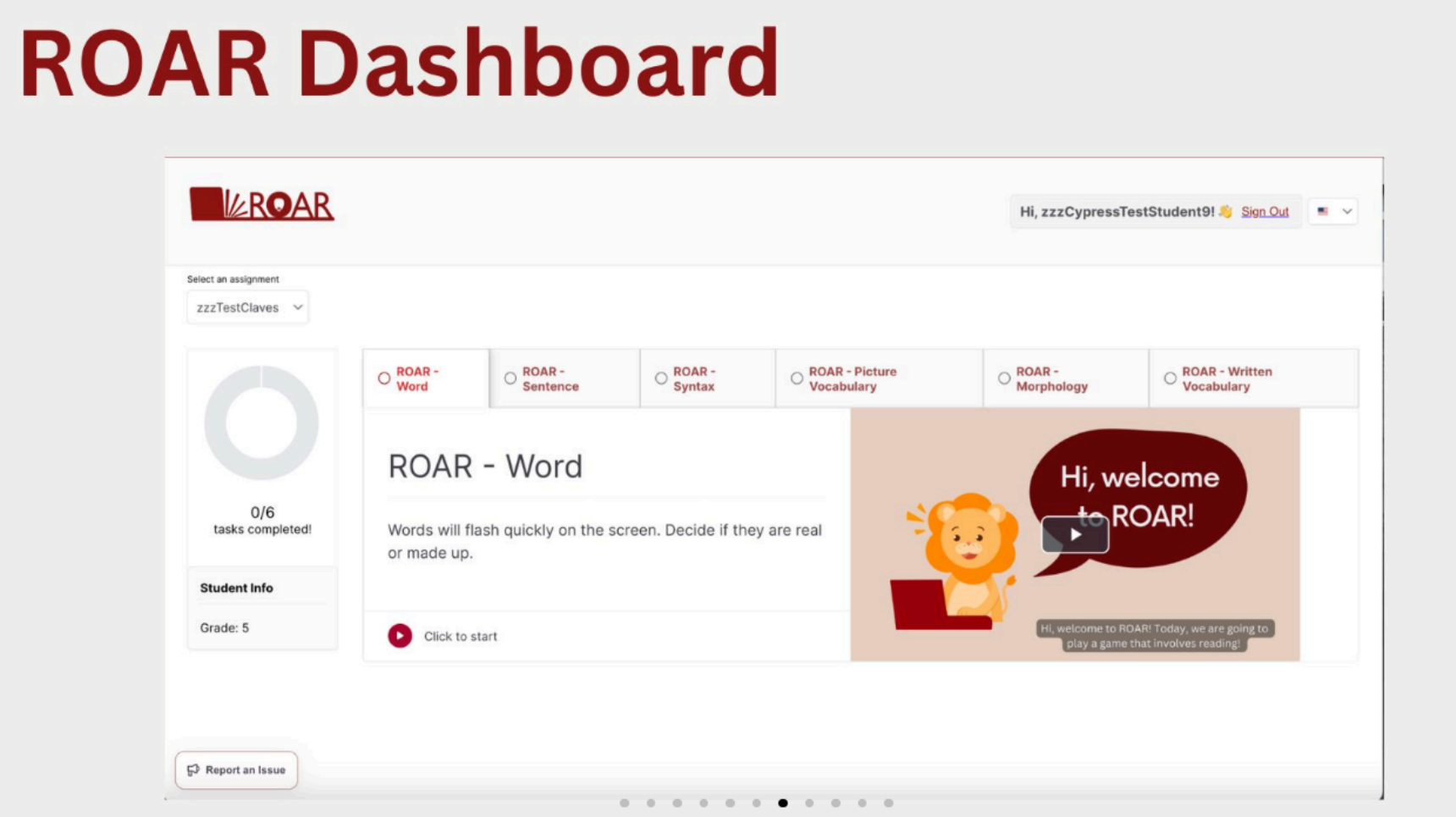
Overview of process for schools to administer ROAR:

Administering ROAR

Timeline for In-Class Administration



Dashboard layout:



Early Bird

Landing page, with buttons for families and educators

Reading. Few things matter more.

The EarlyBird program uses a game-based assessment to identify potential reading challenges like dyslexia, then gives schools and families the tools they need to intervene.

In Schools | At Home

Emotive messaging:

Dashboard:

STUDENT INFO		PREDICTIVE PROFILE	PHONEMIC AWARENESS		PHONICS			FLUENCY		VOCABULARY		COMPREHENSION
Name	Grade	Potential for Word Reading	(%ile)		Score / %ile	(%ile)		(%ile)		(%ile)		(%ile)
			Deletion	Nonword Repetition	Letter Sound	Nonword Reading	Nonword Spelling	Word Reading	Letter RAN	Expressive Vocabulary	Word Matching	Follow Directions
1, Student	1	98% High	81st	70th	81st	79th	82nd	83rd	91st	46th	90th	82nd
2, Student	1	3% Low	17th	32nd	30th	20th	17th	12th	22nd	74th	99th	5th

Market Research: A review of three EdTech Companies that have gone to scale

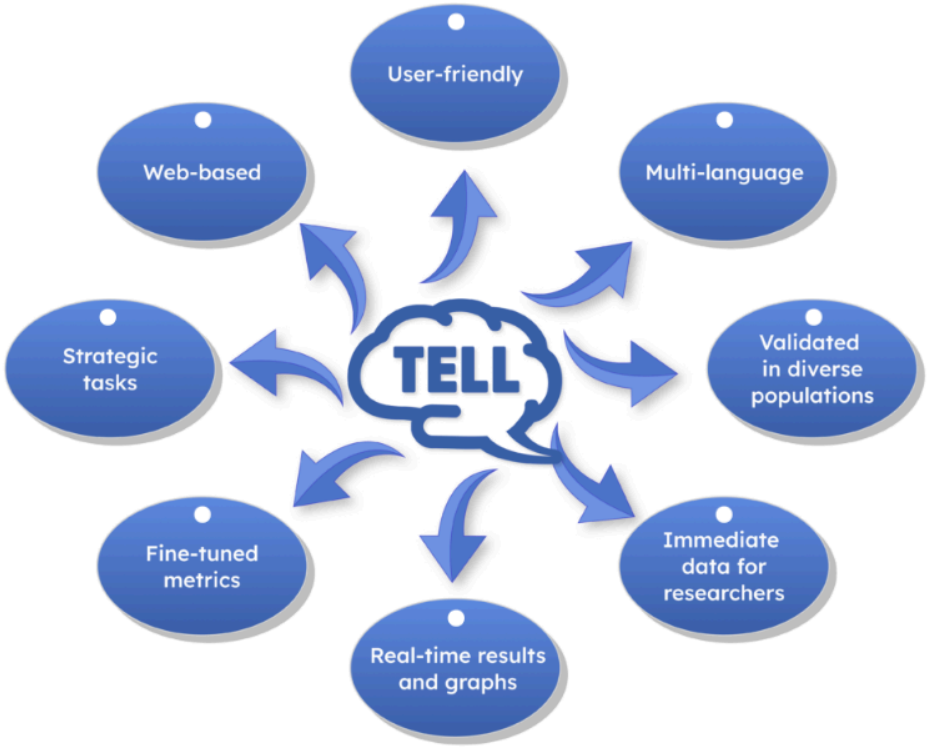
Basic info *Product name* *Website* *Location (where based)* *Language(s)*	**Purpose** *What it's for* *How it works* *How it describes itself*	**Size/reach info** *Potential customers* *How many* *How widespread* *Cost*	**Commercial info** *Product-market fit (why people pay for this)* *What enables scale (e.g. tech features)* *What helped them scale*	**Insights for DECILE**
TELL App "Speech and language science for enhanced neurocognitive assessment"	"TELL's metrics can detect diverse brain diseases, capture their severity, and predict underlying neurobiological disruptions" Graphics below outline processes for clinical and educational settings.	Reach and cost for users are unclear. Available for use by registered centres.	Has a large team, high-profile backers and seems to be used in multiple contexts. But the website feels clunky and has limited info for potential clients. There isn't an obvious off the shelf product. The choice of a generic, difficult to search name is surprising.	Interesting contrast - looks to be technically sophisticated but as a commercial enterprise the comms and presentation feel weak. TELL has a dauntingly scientific-feeling website - not hugely welcoming for the non-specialist. Useful comparison point when creating DECILE's external facing comms.
Propuesta Dale! From Argentina - used across 12 provinces	Supporting prerequisites of literacy Mix of digital and analogue elements.	10k+ downloads on Google Play Store Has been used by 5000 teachers in 2300 schools in Argentina.	Being adopted at the province level - e.g. in Mendoza Website feels outdated and poorly maintained -	Another example of poor presentation that might be masking a strong product. Would be interesting to

	Professional development offer (regular courses on how to use them with students).		last blog posted in 2021.	learn how they negotiated their spread across Argentine provinces.
Glifing https://www.glifing.com/ From Spain but also has tests for Rioplatense Spanish for assessing reading.	Accelerating reading and memory skills through online games, based on an initial evaluation. Pitched at schools, families and other user groups.	Used by more than 750 education centres and more than 200k people in total - globally across 20 countries.	Free trial offered on submission of details. Prices not transparent; have to enter details first. Page of testimonials and logos of institutions that use it. Much more commercial feeling site - steering you to enter details; frequent pop-ups.	Clearly and assertively communicates its reach.

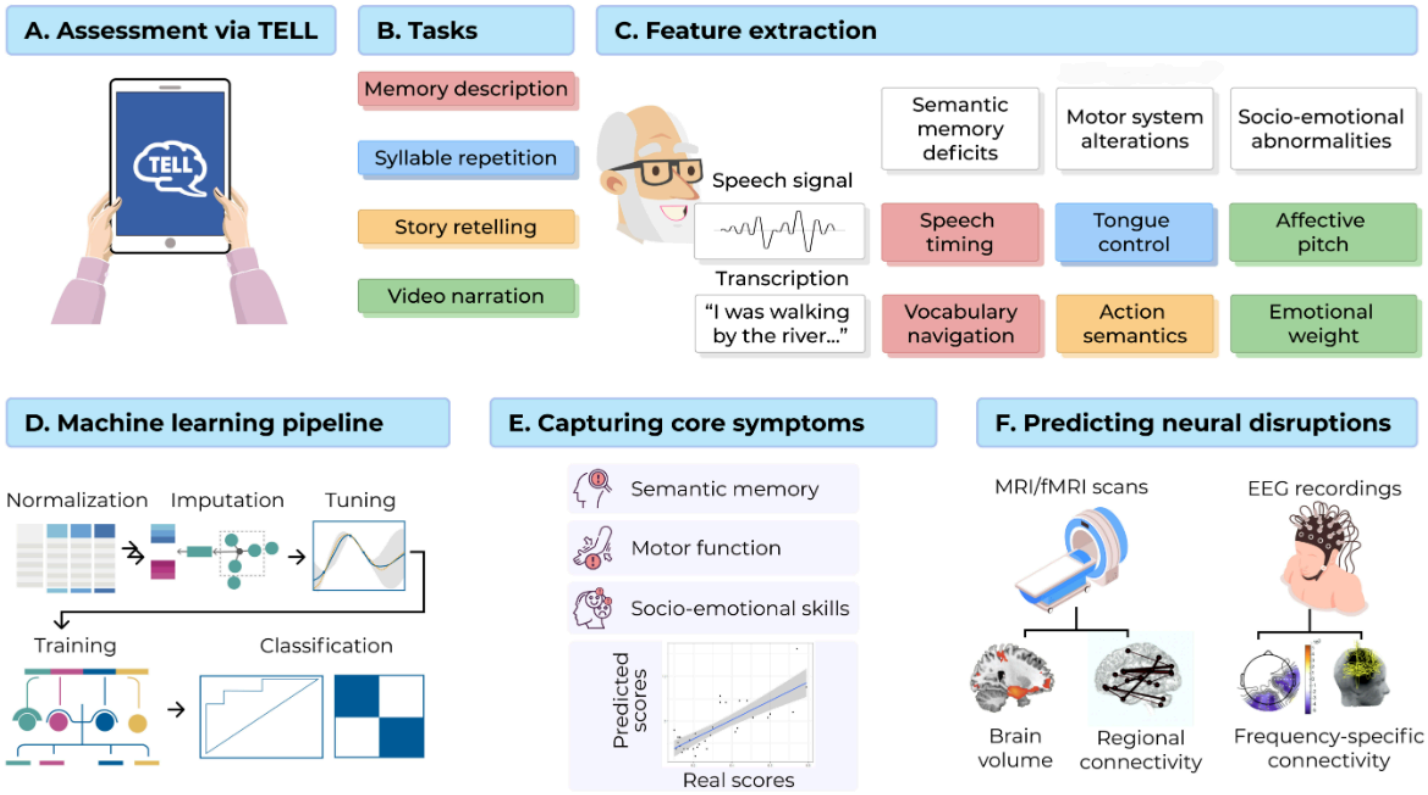
Market Research: Images for interest or inspiration

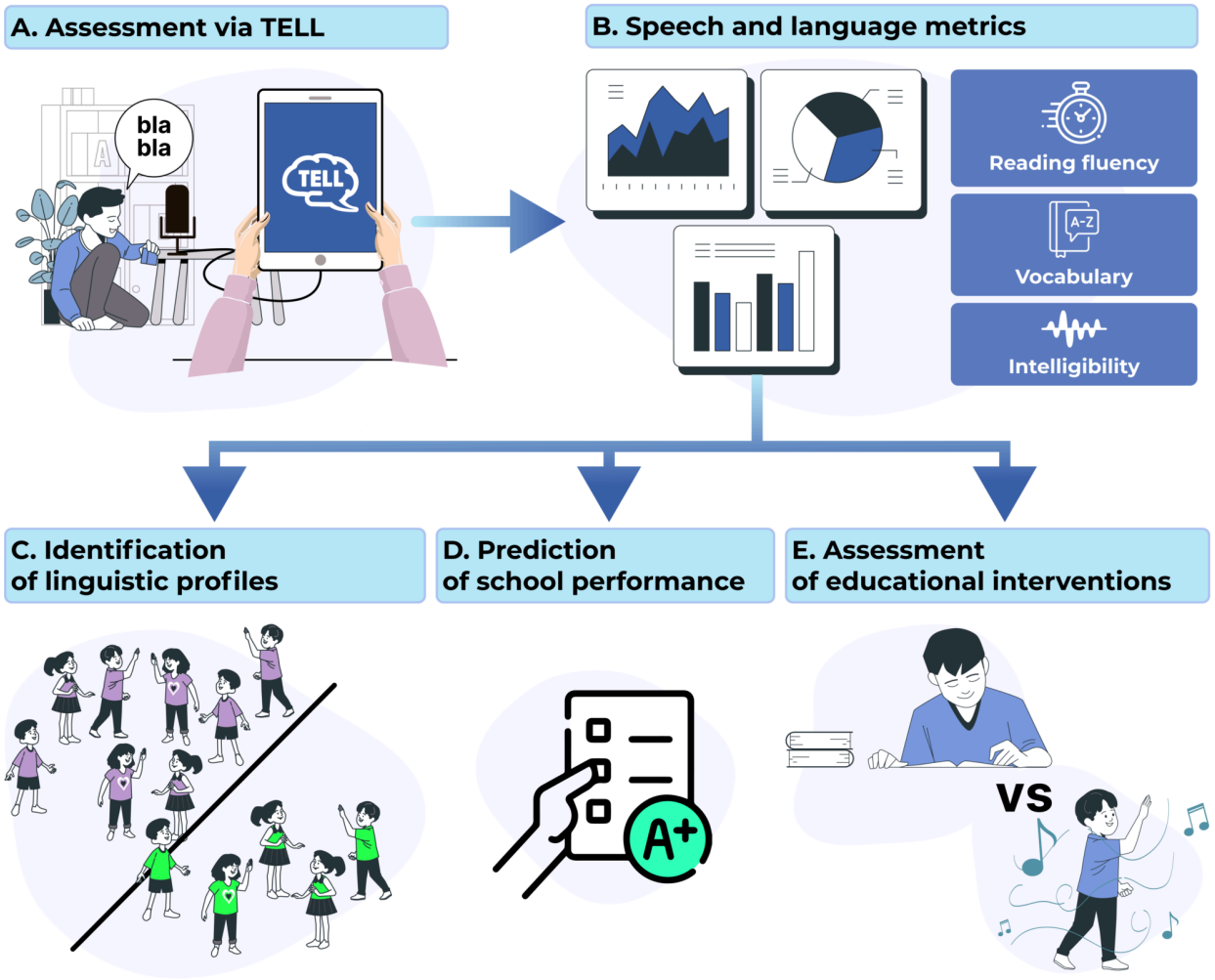
TELL

Overview

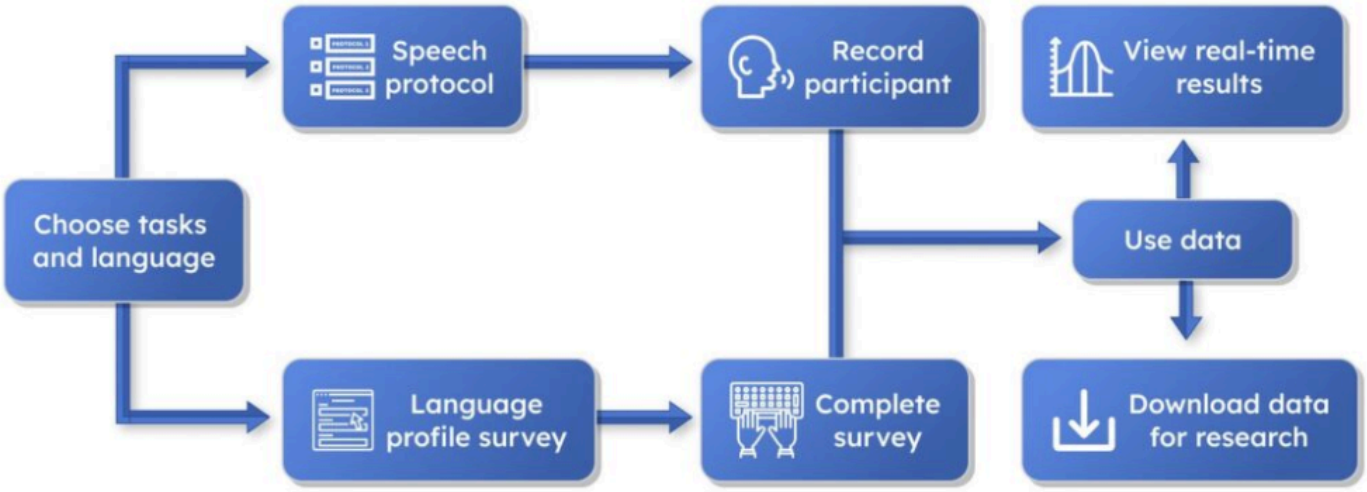


LEAP


A. Assessment via TELL
TELL
B. Tasks
Memory description
Syllable repetition
Story retelling
Video narration
C. Feature extraction
Speech signal
Transcription
"I was walking by the river..."
Semantic memory deficits
Motor system alterations
Socio-emotional abnormalities
Speech timing
Tongue control
Affective pitch
Vocabulary navigation
Action semantics
Emotional weight
D. Machine learning pipeline
Normalization
Imputation
Tuning
Training
Classification
E. Capturing core symptoms
Semantic memory
Motor function
Socio-emotional skills
Predicted scores
Real scores
F. Predicting neural disruptions
MRI/fMRI scans
EEG recordings
Brain volume
Regional connectivity
Frequency-specific connectivity


A. Assessment via TELL
B. Speech and language metrics
bla
bla
TELL
Reading fluency
Vocabulary
Intelligibility
C. Identification
of linguistic profiles
D. Prediction
of school performance
E. Assessment
of educational interventions
A+
VS

LEAP


Choose tasks and language
Speech protocol
Language profile survey
Record participant
Complete survey
View real-time results
Use data
Download data for research

Glifing

Reach:

Ideas for further investigation

You might also want to look at the following lab-developed tools:

BalanceAI, a lab-developed tool from University of Toronto that has gone to scale. BalanceAI supports and assesses foundational literacy skills, and provides scaffolded intervention in areas of metacognition and self-regulated learning. It provides teacher guides on how to interpret scores as well as how teachers should intervene and strategies of how to support learners.

DIBELS, a curriculum-based measure (specific to the US). It may be worthwhile to consider if there is alignment between language skills in curriculum and DECILE.

VLP (Virginia Literacy Partnerships), at the University of Virginia: several screeners and other tools in a large-scale gov-university partnership. This could be a model to emulate for DECILE's long term scaling.

And it might be interesting to look at the design of:

Apps for developing early speech

- Speech Blubs
- Articulation Station

Apps for testing the softest sound you can hear

- HearWHO
- Hearing Numbers

AI-enabled apps for early reading

- AIRA (Google Cloud) https://github.com/GoogleCloudPlatform/aira
- Microsoft Reading Coach https://coach.microsoft.com/en-gb - uses speech recognition