The design and validation of an online speaking test for young learners in Uruguay: challenges and innovations

Nahal Khabbazbashi[1], Fumiyo Nakatsuhara[1], Chihiro Inoue[1], Gabriela Kaplan[2], and Anthony Green[1]

[1]Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire, United Kingdom
[2] Plan Ceibal, Uruguay

Abstract

This research presents the development of an online speaking test of English for students at the end of primary and beginning of secondary school education in state schools in Uruguay. Following the success of the Plan Ceibal *one computer-tablet per child* initiative, there was a drive to further utilize technology to improve the language ability of students, particularly in speaking, where the majority of students are at CEFR levels pre-A1 and A1. The national concern over a lack of spoken communicative skills amongst students led to a decision to develop a new speaking test, specifically tailored to local needs.

This paper provides an overview of the speaking test development and validation project designed with the following objectives in mind: to establish, track, and report annually learners' achievements against the Common European Framework of Reference for Languages (CEFR) targeting CEFR levels pre-A1 to A2, to inform teaching and learning, and to promote speaking practice in classrooms. Results of a three-phase mixed-methods study involving small-scale and large-scale trials with learners and examiners as well as a CEFR-linking exercise with expert panelists will be reported. Different sources of evidence will be brought together to build a validity argument for the test. The paper will also focus on some of the challenges involved in assessing young learners and discuss how design decisions, local knowledge and expertise, and technological innovations can be used to address such challenges with implications for other similar test development projects.
(248 words)

Key words

Young learners, speaking assessment, test development and validation, English language learners, CEFR

Introduction

In a globalized world where knowledge of English is "often associated with economic and political power" (Huang, Chang, Niu, & Zhi, 2018:27), the development of communicative competence in English for young learners has been at the centre of educational reform efforts (Bailey, 2017). Johnstone (2009:33) calls the prevalent implementation of early teaching and learning of foreign languages (and English in particular) "the world's biggest policy development in education". There has been a parallel need to develop assessment solutions that meet the accountability requirements of governments, educational institutions, and other stakeholders in ensuring that such educational reform efforts are successful (Bailey, 2017).

This reflects a "policy shift towards evidence-based instruction" (Nikolov & Timpe-Laughlin, 2020: 2). While both national-level assessments and commercial proficiency tests such as the Cambridge Young Learners suite of exams, TOEFL Junior, and Aptis for Teens have responded to this demand, there is little publicly available empirical research conducted on the assessment of young learners (Butler, 2019) with even fewer studies on the assessment of speaking skills and related test validation efforts (Huang, Bailey, Sass, & Chang, 2021). The current study addresses this gap and reports on the design and validation of a national-level online test of English speaking for young learners in the educational context of Uruguay.

In reporting on this project, the focus will be on some of the challenges involved in assessing young learners and on how design decisions, local knowledge and expertise, and technological innovation can be used to address such challenges. The validation methodology and results of validation research conducted on the test will be presented and the paper will conclude by bringing different sources of evidence to build a validity argument for the test, drawing on Weir's (2005) socio-cognitive framework for test development and validation.

Assessing young learners

The term "young learners" generally refers to children from the age of five to the age of 12 or the end of primary education (Alderson & Banerjee, 2001) although Hasselgreen and Caudwell (2016) also include teenagers (12/13 to 17 years) under this umbrella term. In designing tests for young learners, it is important to consider the needs and key characteristics of this population. Butler (2016:360-362) lists several such characteristics. These include:

- the young learners' learning contexts, which may vary on a "language-content" continuum (Inbar-Lourie & Shohamy, 2009:84) that ranges from an emphasis on language to a focus on content
- the age of candidates and related developmental factors such as cognitive, socio-cognitive, linguistic, and affective development trajectories
- individual differences between learners
- the 'centrality of learning' where language is primarily used for meaning making and gaining understanding of different content areas
- the vulnerability of young learners to potentially negative impact of assessment on motivation, confidence, and anxiety levels.

Age-related developmental aspects are of particular interest, as they may constrain how tasks are approached. Field (2018:147) for example, notes how in comparison to adults, young learners may be more limited, among other factors, in their working memory resources, their ability to approach tasks strategically, the world knowledge or schemata they can draw on, and their ability to engage effectively with turn-taking conventions.

Another increasingly important consideration in assessing young learners is the use of digital technologies and learners' familiarity with them. In their state-of-the-art review of the use of digital technologies in language learning, Macaro, Handley, and Walter (2012:1) found the beneficial impact of technology on L2 learning to be "slight and inconclusive" but they generally found learners to display a positive attitude towards technology in their L2 learning. Papp and Walczak (2016) drew on national policies and IT curricula in different primary and secondary education settings to suggest how learning is becoming "increasingly mediated through technology" (p.140) and that this may affect young learners' attitudes and abilities in taking digitally-mediated tests. In their validation study of the computer-based version of the Cambridge English Young Learner suite of exams, Papp and Walczak's (2016) findings included highly positive feedback from candidates, parents, and trial observers towards the computer-based test as a "fun, accessible, and alternative" way of assessing young learners. These positive attitudes might be explained by the advantages of technologies in incorporating animated objects (Butler, 2016), visual stimuli, and elements of gaming that can help with retaining attention and motivation in young learners.

These distinctive characteristics of young learners have influenced the ways in which the language proficiency construct has been defined and operationalised for this population. Nikolov and Timpe-Laughlin (2020) provide a useful review of the constructs and frameworks used in assessing young learners' foreign language abilities and how they have changed over time. Earlier attempts, for example, showed a diversity of approaches and a lack of consensus regarding understandings of proficiency for young learners with features that were "strongly reminiscent of constructs proposed in the context of adult L2 learning and assessment" (p.5). The authors, however, noted certain trends that reflected the needs and characteristics of young learners; for example, a focus on speaking and interaction, use of contexts and settings that were familiar to test takers, and drawing on tasks and materials that were considered age-appropriate and motivating. As the field progressed, so did the need for accountability with the development of different models and frameworks. The CEFR, in particular, has been hugely influential in the teaching, learning, and assessment of young learners' language abilities. The framework, however, was not designed for the needs of this particular cohort and the descriptors, particularly at the higher levels, are not necessarily appropriate for the age and cognitive development levels of younger learners (Little, 2007). It is therefore crucial for descriptors to be adapted for the specific local educational context, the age of the candidature, and their language ability levels (Bailey, 2016), which may require further divisions at lower CEFR levels (for examples of such adaptations see Benigno & de Jong, 2016; Hasselgreen, 2003; Papp & Walczak, 2016; and Szabó, 2018a; 2018b).

To summarise, young learners have unique characteristics that need to be taken into account when designing assessments to ensure their appropriateness for different local contexts, age groups, developmental stages, and language levels.

Young learners and validation research

Validity lies at the heart of educational measurement and is viewed as fundamental to the development and evaluation of tests (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Current approaches to validity emphasise the role of *test validation* as a process of collecting, synthesising, and evaluating evidence to support inferences made on the basis of test scores and results (Cizek, 2011; Kane, 2013). Despite its importance, there are very few published validation studies focusing on the assessment of young learners (Butler, 2019; Papp & Rixon, 2018) with most of this research being in the context of commercial large-scale international assessments (Nikolov & Timple-Laughlin, 2020), such as the Cambridge Young Learner English (YLE) series (Papp & Walczak, 2016; Papp & Rixon, 2018) and the TOEFL Young Student series (So et al., 2015; Cho et al., 2016). One critical aspect of test validation that requires more attention is "collaboration with local users and stakeholders" to ensure that large-scale standardised assessments are appropriate "for the local group of learners" (Nikolov & Timple-Laughlin, 2020: 13).

In this section, examples of published validity research projects that focus specifically on the assessment of speaking in young learners are presented with a view to inform the current study. Recent research by Huang et al. (2021) examined the validity of the *TOEFL Junior*® speaking test by evaluating two sources of empirical evidence: the test's internal structure and its relationship to other assessments drawing on Messick's (1989) validation framework and Kane's interpretation argument framework (Kane, 2013). The methodology involved the administration, scoring, and analysis of the following to 252 Taiwanese adolescent learners of English: the *TOEFL Junior®* speaking test, two additional speaking tests – a communicative monologue test and a communicative interactive test, a survey on English language learning experiences, and self-assessments of language abilities. Results of correlations and confirmatory factor analysis provided evidence in support of the test's internal structure with the test's subtests generally loading on the same theorized factor. Results of correlations between scores on *TOEFL Junior®* speaking test and other speaking measures ranged from r=.55 to r=.73 suggesting strong relationships in the expected directions. Correlations with construct-relevant learning experiences were also considered high though smaller in comparison to the speaking measures, generally falling between r=.30 to r=.49. These findings were taken together as strong support for the construct validity of the *TOEFL Junior®* speaking test although the authors emphasised the need for the collection of other validity evidence such as those related to the consequences of test use. Lee and Winke's (2018) research compared native English-speaking (NS) and non-native English-speaking (NNS) young learners' response processes and attentional foci by tracking their eye movements as they responded to the *TOEFL® Primary*™ computer-mediated speaking test. The authors also pointed to the lack of research on children's perspectives on language assessments (Butler & Zheng, 2014) and addressed it by eliciting children's reactions through interviews and picture drawing. Large score differences were observed between NS and NNS but more unexpectedly, patterns in eye movement varied substantially between the two groups. There were more prolonged fixations by NNSs on task-irrelevant features (such as on-screen timers) whereas their NS counterparts spent more time looking at content features (e.g. pictures and visual stimuli) that would help them prepare for speech production. This led the authors to question the appropriateness of borrowing features such as timers from adult-testing scenarios and emphasised the importance of implementing task design features that were child-appropriate.

Finally, Papp and Walczak (2016), also referred to earlier, used a mixed-methods study examining the comparability of computer-based (CB) and paper-based (PB) versions of the Cambridge Young Learners English series (including the speaking component). Both versions of the test were administered to young learners from different first languages and proficiency levels along with questionnaires eliciting information on different background variables (e.g. country, age, gender, years of instruction, computer preferences, and frequency of computer use). Strong correlations were observed between scores on computer-based and paper-based versions of the tests. Results of regression analyses also suggested number of years of English language instruction – a construct-relevant variable – as the main significant predictor of performance in both PB and CB tests. Taken together these findings provided strong support for the use of both versions of the test in tapping into a similar construct of spoken ability.

This brief review of validation research on young learners highlights the importance of collecting different kinds of validity evidence particularly evidence concerning the appropriateness of test design features, including technology, for the unique characteristics of young learners; the fitness of the test for the local context; and whether assessment models and frameworks have been appropriately adapted to the needs and language levels of the target population. These are areas that will be touched on in the next sections in relation to the test development project and validation research described in this study.

The study

The current study reports on the development and validation of a new online English speaking test at the end of primary and beginning of secondary school education in state schools in Uruguay. This was a collaborative project between two teams – the Plan Ceibal team in Uruguay and a team of language testing researchers in the UK. Plan Ceibal was created in 2007 "as a plan for inclusion and equal opportunities with the aim of supporting Uruguayan educational policies with technology. Since its implementation, every child who enters the public education system in any part of the country is given a computer for personal use with free Internet connection at school. In addition, Plan Ceibal provides programs, educational resources and teacher training courses that transform the ways of teaching and learning" (ceibal.edu.uy).

Following the success of the Plan Ceibal *one computer-tablet per child* initiative, there was a drive to further utilise technology to improve the language ability of students, particularly in speaking, where the majority of students are at levels pre-A1 and A1 of the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001; 2018). The national concern over a lack of spoken communicative skills amongst students led to a decision by Plan Ceibal to develop, in collaboration with researchers in the UK, a new online speaking test specifically tailored to local needs. The following test purposes were identified by Plan Ceibal accordingly: (a) to establish, track, and report annually the learners' achievements against the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001; 2018) targeting CEFR levels pre-A1 to A2, (b) to inform teaching and learning, and (c) and to promote speaking practice in classrooms.

The test's target population includes all learners in the 6th grade of Primary state school education in Uruguay (aged between 11 and 12 years old) and all learners in the 1st grade of Secondary state school education in Uruguay (aged between 12 and 13 years old).

Before explaining the test in more detail, the next section presents an overview of the test development and validation framework that guided the study

Test development and validation framework

The framework guiding this research is Weir's (2005) socio-cognitive framework (SCF) for speaking test development and validation further elaborated in Taylor (ed., 2011). SCF informs test development, research and validation. It has been adopted by many leading testing organisations and researchers worldwide. It combines social, cognitive and evaluative dimensions of language, linking these to the contexts and consequences of test use.

In brief, the framework consists of six components: test taker characteristics, cognitive validity, context validity, scoring validity, consequential validity, and criterion-related validity. *Test taker characteristics* refer to features of the candidates, for example, age, sex, personality, background knowledge, and language learning experience. As already discussed in the literature review, these characteristics can influence the way in which a task is performed. The *cognitive validity* component relates to the mental processes activated by tasks and should reflect an established theory of the mental processes underlying the construct of interest. For a speaking test, for example, cognitive validity evidence should demonstrate that speaking tasks "activate cognitive processes in a test taker's mind similar to those employed in the real-life speaking tasks and situations the test aims to generalise to" (Weir, Vidaković, & Galaczi, 2013). The *context validity* component of the SCF is used as a superordinate category related to tasks and the performance conditions under which they are performed. Within test constraints, Weir (2005:56) suggests that performance conditions should be as similar to authentic language use as possible. *Scoring validity* is a superordinate term used by Weir (2005:22) to refer to all aspects of reliability and in terms of speaking includes considerations of rating criteria, rating scales, and the application of the rating scales by human raters (or machines in the case of automated assessment). Consequential validity in the SCF refers to the impact of tests on institutions and society as well as washback in the classroom or workplace. Lastly, *criterion-related validity,* refers to a consideration of evidence external to the test that lends further support for the meaningfulness of scores or "the extent to which test scores reflect a suitable external criterion of performance or demonstration of the same abilities as are included in the test" (O'Sullivan & Weir 2011:23–24).

By conceptualising validity as "multifaceted", different sources of validity evidence are fitted together in the SCF to present a unified approach to validity. Note that numerous combinations of questions can be posed under different validity components of the SCF, which can in turn be addressed using a variety of approaches in collecting validity evidence. Weir (2005) also emphasises that test validation evidence should not be collected only after a test is fully developed (i.e. *a posteriori* validation) highlighting the importance of *a priori* validation studies during the development of a test to inform the test design. This is a principle followed in this test development and validation project.

Test construct

The construct underlying the speaking test is that of oral language proficiency involving the production of (monologic) speech together with aspects of receptive listening skills. While a direct face-to-face or video-conferencing mode of delivery would have the great advantage of tapping into a broader construct of speaking, it would also pose a substantial pragmatic challenge in terms of resources, planning, and arranging for qualified interlocutors. This is particularly the case for this specific assessment context and the requirement for large-scale administration of the test across primary and secondary schools in Uruguay and in all regions of the country. The Plan Ceibal objective was to be able to simultaneously reach a significant number of students at one time point in order to have a representative sample of student performances in speaking every year. As such, a semi-direct test was considered the only practical solution for reaching the target population and a decision was subsequently made by Plan Ceibal to deliver the speaking test via computer or tablet. This, in turn, had implications for test construct and task design; a semi-direct computerised speaking test by its nature narrows the speaking test construct to one-way production, as it does not allow for (any) interactions either with an examiner or with another candidate. Nevertheless, a semi-direct test was considered a better alternative – in terms of encouraging speaking practice and positive washback in classrooms – than having no speaking test at all. This decision was to some extent supported by research evidence from other assessment contexts that suggest strong comparability between face-to-face and computer-based versions of speaking tests (see for example Isbell & Winke (2019); Thompson, Cox, and Knapp (2016)). Moreover, to further provide validity evidence, a research project is currently in progress by Plan Ceibal to examine the comparability of scores across online and face-to-face modes of delivery.

Validity considerations

Drawing on Weir's (2005) socio-cognitive framework for speaking test development and validation, a range of parameters for cognitive, context, and scoring validity were carefully considered when defining the construct of the test. For example, Field (2018:192-193) considers the semi-direct test format from a cognitive perspective in relation to young learners; a challenging aspect of the format is that it does not allow candidates to ask for clarifications or to check their understanding and thus creating 'an unsupportive task environment for young learners’. On the other hand, this format does allow young learners to have more control over the speech produced, as there will be no interruptions or unexpected turns from interlocutor(s). Nevertheless, a monologue is considered to “impose considerably greater cognitive demands upon a young candidate than an interactive situation” (p.193). In assessment contexts where only semi-direct speaking tasks can be used, both Field (2018) and Hasselgreen and Caudwell (2016) recommend the use of narrative tasks as well as descriptions and very simple explanations. These recommendations were taken into account when selecting tasks and developing test specifications (see next section).

Test description

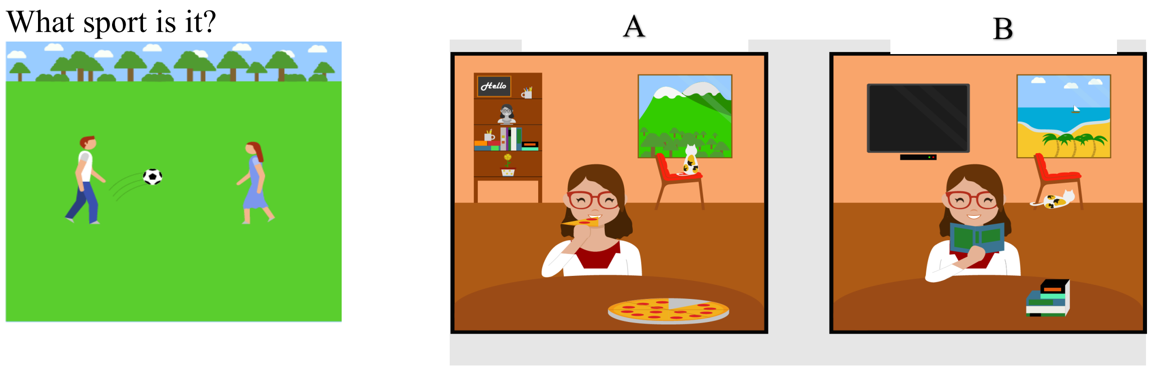
The Plan Ceibal speaking test is an online browser-based test. Test tasks are presented to candidates on the screen and through headphones. Responses are digitally recorded, stored, and sent to examiners for scoring.

The test comprises four parts (see Table 1 for an overview of each test part: target level, number of items, target language functions, and timing). Before the start of the test, candidates watch an instructional video in Spanish and go through a technical audio check. In Part 1 (*Question & Answer*), candidates are required to answer five questions about themselves (e.g., What is your favourite colour?). In Part 2 (*Name it*), candidates are shown five images to identify and name. These may include familiar objects, seasons, sports, etc. In Part 3 (*Talk about familiar topics*), candidates are asked to talk about a familiar topic (e.g., your best friend) and are provided with three prompts/cues to help them produce short monologues. In Part 4 (*Spot the differences*), candidates compare two pictures to identify differences and describe them. Based on trial results, an example response was added to this part to provide scaffolding (e.g., In A, there is a pizza on the table, and in B there are books on the table. Based on input from teachers, these four tasks are intended to simulate the kinds of conversation that a young learner can have at school with their teachers or peers . In all parts, candidates can opt to listen to the questions once (single play) or twice (double play). Field (2018) argues that at lower levels of proficiency, double play may be desirable as it reduces the time pressure imposed on the listener when processing speech in real time. He also adds that double play is no longer necessarily inauthentic, as many of our real-life listening opportunities via digital technologies allow for repeated listening. Double play is also associated with lower levels of anxiety for candidates (Holzknecht, 2019).

Selected task screenshots from Parts 2 and 5 are provided in Figure 1.

Table 1 Speaking test structure

Part	Level	No. of items	Language function	Timing
Test instruction video & audio check (not assessed)				
Part 1: Q & A	Pre-A1/ A1	5	Provide personal information	3 mins
Part 2: Name it	Pre-A1/ A1	5	Name objects using pictures	3 mins
Part 3: Talk about familiar topics	A1/ A2	3	Describe routines, familiar people, past experiences	4 mins
Part 4: Spot the differences	A2	1	Compare/contrast, describe differences	2 mins 30 secs



Part 2 (*Name it*) Part 4 (*Spot the differences*)

Figure 1 Selected task screen shots

Validity considerations

In developing test specifications and designing tasks, one of the most important validity considerations was ensuring that the test was appropriate for the young learner target population, that the test elicited learners' best performances, and that there were synergies between the assessments and classroom practices in the local context. The principles guiding test development were informed by the review of the literature and recommendations for task design (Hasselgreen & Caudwell, 2016; Field, 2018; Papp & Rixon, 2018; McKay, 2006) as well as a comprehensive review of local teaching materials, curricula, and speech samples provided by Plan Ceibal. Below is a summary of these guiding principles:

- **Tasks:** use narratives, descriptions, and simple explanations; these are frequently-used genres in primary classrooms and are likely to elicit young learners' best performances. Use task features and technologies that are motivating, unthreatening, fun, and game-like with humorous/entertaining pictures where possible to reduce any test-related anxiety and help elicit best performance.
- **Ordering of tasks:** order tasks in terms of cognitive load and difficulty from easy – requiring one-word responses (Pre-A1) – to more challenging requiring the production of a short stretch of speech to describe, narrate, and compare/contrast (A2). This helps not to demotivate lower level candidates when they first start the test.
- **Domain:** use topics in a personal domain related to everyday situations, events, and activities, social relationships with peers and friends, opinions, feelings, and experiences; these are appropriate topics/themes to be included for young learners in assessment contexts. Use areas of immediate experience for task settings; the younger the learners, the less world knowledge they have to draw on and therefore it is crucial to provide them with settings they are familiar with.
- **Language functions:** use language functions that are considered important in the local teaching context and align closely with the curricula.
- **Scaffolding:** provide sufficient support to complete tasks for example with the use of visuals and prompts; young learners need support to put them at ease, reduce any test-related anxiety, and help elicit best performance. Examples of scaffolding provided in the test include (a) an instructional video in Spanish by a young, friendly-looking teacher to familiarise candidates with the different platform functionalities; (b) use of

Spanish for instructions for the different test parts; and (c) allowing candidates the option to repeat instructions and questions at least once.

- **Input language:** use teacher judgments and tools such as English Vocabulary Profile and Grammar Profile to ensure language of the input is accessible to learners at the target output level.
- **Timing**: limit **test length** to a maximum of 13 minutes to accommodate young learners' shorter attention spans
- **Piloting and stakeholder feedback**: implement feedback sessions and trials to inform revisions and modifications to the test.

A set of tasks were drafted based on these principles and subsequently went through a series of revisions and modifications following iterative discussions with the project teams, feedback from key stakeholders (e.g., raters, teachers, Plan Ceibal staff including Quality Managers responsible for improving English Language Teaching products and services), and the results of two trials.

Scoring

The operational speaking test is scored remotely by human examiners. Examiners are asked to use a four-band analytic rating scale that ranges from *below Pre-A1* to *A2 and above* and to award scores based on three criteria: *Task Fulfilment*, *Grammar and Vocabulary*, and *Delivery* (see Appendix A for the final version of the rating scale). They are asked to consider each of these independently of the other two.

Validity considerations

There were two stipulations outlined by Plan Ceibal for scoring: firstly, that test results would be reported on the CEFR from pre-A1 to A2; and secondly, that these CEFR levels would be further broken down into seven sub-levels (A0, A0+, A1-, A1, A1+, A2-, A2) in order to provide learners with fine-grained diagnostic information for feedback purposes. Based on the review of the literature, our main focus was to ensure (a) the adequate linking of the test results to the CEFR, (b) the appropriateness of the rating scale and descriptors for the target ability levels, and (c) the functioning of the rating scale in reliably distinguishing between candidates. The validation research (see next section) was therefore designed with these considerations in mind.

The next section outlines the validation research methodology aimed at collecting empirical validity evidence from a variety of sources.

Speaking test validation

Design

A three-phase mixed-methods research design was adopted for this study (Phase 1: small-scale *a priori* validation study, Phase 2: large-scale *a priori* validation study, Phase 3: the CEFR linking exercise) to collect the following sources of validity evidence:

- Candidate feedback on different aspects of the test as a way of establishing test appropriateness for the target population (*focus on context validity*)
- Information on rating scale functionality and psychometric properties of the speaking test to ensure dependability of scores (*focus on scoring validity*)
- Results from a standard-setting exercise to confirm synergies between a locally-developed test and an international test (*focus on criterion-related validity*)

Details of each phase of data collection, participants, the instruments used, and main findings from each phase that fed into subsequent phases are provided below.

Phase 1: small-scale a priori validation study

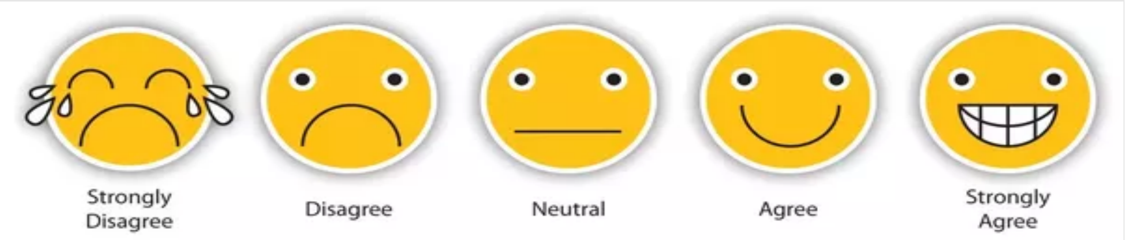
The first phase of research was a small-scale *a priori* validation study conducted at an early stage of test development (Weir, 2005). The aim of this phase was to (a) evaluate various contextual parameters of the test (e.g. test functionality in technical terms, clarity of instructions for candidates, appropriateness of tasks and suggested timings for the target candidature) and (b) collect samples of speech to inform the drafting of an empirically-based rating scale.

Participants

Student participants. A sample of 50 students from three schools participated in this phase of the study. Sample selection was informed by teacher evaluation of students to ensure a range of language ability levels were represented.
Rater participants. Three experienced Plan Ceibal raters also participated in this phase. Raters were selected based on their familiarity with the CEFR and other international young learner speaking tests.

Instruments

Speaking test. The preliminary version of the speaking test was used in this phase.
Candidate Questionnaire. A candidate questionnaire was designed to (a) elicit student attitudes towards the speaking test, (b) identify any problems – either regarding tasks or technical issues – and (c) collect background information. Smiley faces were originally used for a five-point Likert-type items as follows:



Following feedback from student observers and Plan Ceibal teachers, the response options were amended to be simple adjectives (in Spanish) rather than representing the degree of agreement to each statement. The English translation of the final version of the questionnaire (used in Phase 2) can be accessed in Appendix B. The questionnaires were available online.

Data collection and analysis

Data collection took place in schools. Participants were invited to a quiet room with laptops/computers to take the speaking test. There was a teacher observer present at every data collection session to explain the procedures, set up the test, and help students with any questions or technical issues. Once the speaking test was completed, students completed the feedback questionnaire online.
Student speaking responses were extracted from the platform and three raters were invited to award holistic impressionistic judgments of CEFR ability from A0 (Pre-A1) to A2 broken down into seven levels. No descriptors were available at this point and raters relied on their knowledge of the CEFR and their experiences with rating other young learner exams. Note that rater training guidelines were developed at a later stage and once a workable rating scale was drafted.
The questionnaire was analysed using descriptive statistics and any open comments were summarised into themes. The speaking score data was analysed using Many-Facet Rasch

Measurement (MFRM) with the programme *Facets* (Linacre, 2018a). MFRM provides a technical solution to the well-documented rater effect in performance assessment by allowing different facets of the testing situation to be measured independently and then mapped onto a common linear scale measured in “logits”. Importantly, candidates’ ability measures from the analysis are estimated independently of the particular rater or task assigned to them, with their raw scores adjusted for the effects of the facets of performance. The resulting candidate fair-average mark is a more objective estimate of the candidate’s ability.

Findings (Phase 1)

Student questionnaires. Findings from the student questionnaire results (n=50 in total) generally suggested a positive attitude towards the speaking test with the majority of participants finding the instructions clear (n=46), the test fun (n=31), and the test ok (n=21) or easy (n=17). The most problematic part of the test was found to be Part 3 (a story-telling task with pictures). This was corroborated by observers who noticed students (even at higher ability levels) struggling with this task.

Speaking score data. Findings from the analysis of speaking score data showed that most candidates were at A0 and A1 levels, despite Plan Ceibal efforts to include candidates with stronger speaking abilities at the participant recruitment stage. MRFM analyses also indicated that the candidates could be reliably separated into approximately four statistically distinct levels (H=4.44; *R*=0.90) and that the categories A0 and A0+ were not easily distinguishable by raters.

The following major decisions were subsequently made on the basis of Phase 1 results and discussions with the local test development team:

- Replace the *story-telling task with pictures* with *a news-telling task with prompts* (now Part 3 in Table 1). Although picture narrative tasks are considered an appropriate genre for young learners (Hasselgreen & Caudwell, 2016; Papp & Rixon, 2018), Phase 1 results indicated the task to be too demanding for the target learners. A short news-telling task was instead selected as more suitable alternative that better aligned the assessment task with learning tasks in the curriculum.
- Reduce the number of CEFR sub-levels for reporting purposes to five levels: Pre-A1, A1-, A1, A1+ and A2.
- A decision to draft descriptors based on analysis of performance features observed in the trial on three criteria of *Task fulfilment, Language,* and *Delivery* and a review of other established descriptors for young learners (e.g., Cambridge YLE, CEFR-J, Trinity ISE Foundation/I, CEFR-related descriptors available at: https://www.coe.int/en/web/common-european-framework-reference-languages/bank-of-supplementary-descriptors).

Phase 2: large-scale a priori validation study

The second phase of the *a priori* validation research was of a confirmatory nature and was designed with the following aims: (a) to check the test functionality in technical terms for administrations to larger groups of candidates, (b) to get candidate feedback regarding the speaking test and the revised Part 3, (c) to get rater feedback regarding appropriateness of the draft scale, (d) to develop rater training guidelines, and (e) to select speaking performances for Phase 3 of the study (the CEFR standard setting process).

Participants

Student participants. A sample of 139 students participated in this phase of the validation study. Table 2 shows breakdown of the candidates by age with the majority of candidates aged 11 (52%) and 12 (36%) respectively which is representative of the age composition of the target population. Students were selected from twelve schools (six primary schools and six secondary/vocational schools) in three different districts in Uruguay: Montevideo, Canelones, and Rivera. Steps were taken to ensure students from diverse socio-economic backgrounds (divided into quintiles in Uruguay) were represented in this sample.

Table 2 Breakdown of Phase 2 participants by age

Age	Number	Percentage (%)
10	2	1%
11	72	52%
12	49	36%
13	13	9%
14	2	1%
Total	**139**	**100%**

Rater participants. A group of eight experienced raters – all Quality Managers (QMs) from Plan Ceibal – participated in Phase 2.

Instruments

Speaking test. The revised version of the speaking test (following modifications from Phase 1) was used in Phase 2.

Candidate Questionnaire. A revised version of the candidate questionnaire (following feedback from Phase 1) was used in Phase 2.

Rating scale. A revised version of the rating scale (analytic scale with three criteria and maximum possible score of 9) was used for rating purposes.

Rater Questionnaire. A rater questionnaire was designed to elicit rater feedback on the draft rating scale, its usability, the different criteria and descriptors, as well as observations regarding the test tasks and the extent to which they elicited rateable samples of speech. The questionnaire included a section for open comments (see Appendix C for the questionnaire items).

Data collection and analysis

Student data collection followed similar steps to Phase 1 and involved the administration of the speaking test followed by the online questionnaire.

A rating matrix was devised to facilitate rating of speech samples by raters and create a link between raters to meet the requirements of MFRM. All students were rated by at least two raters. Additionally, five speech samples representing different levels of proficiency were rated by all raters to ensure raters' exposure to a range of ability levels as well as linking scores and raters across the data set: a requirement for MFRM analyses. Once ratings were assigned, raters completed the rater questionnaire.

Descriptive statistics were used to summarise the student and rater questionnaire results and any open comments were summarised and grouped into themes. The speaking score data was analysed via MFRM.

Findings (Phase 2)

Speaking score data. The score data were analysed with MFRM using a three-facet model (candidate, rater, score). In the first round of analyses, all score points (max score of 9) were used as assigned by raters. Results, however, suggested that some score categories were not reliably assigned by raters with evidence of disordered thresholds.

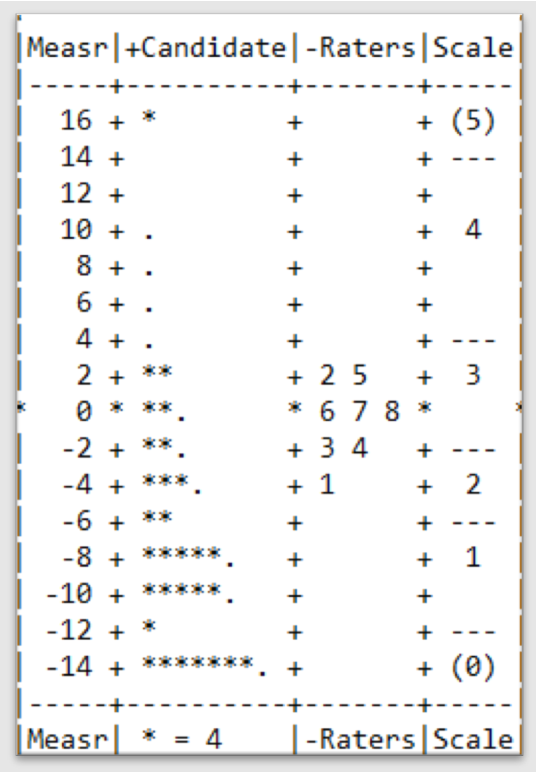
For this reason and on the basis of evidence from the category statistics, the problematic score points were collapsed and another MFRM analysis was run. With the remedial action taken, the rating scale was shown to be functioning appropriately.

The three facets were mapped onto a common interval-level scale (the logit scale) and results are visually represented in the Wright map in Figure 2. This map illustrates in graphical form the calibration for all candidates, raters, and the speaking scale with the logit scale serving as a single frame of reference for interpreting the results of the analyses.

A quick view of the map suggests that the speaking test has been successful in eliciting a range of ability levels from pre-A1 to A2 (spanning a range of 35.14 logits). The strong majority of candidates, however, are at Pre-A1 and A1 levels with only a few candidates displaying stronger ability at A1+ or A2 levels. The candidate group-level statistics (see Table 3) show that the trial candidates can be reliably divided into a minimum of 4.21 ability strata (*R*=0.89).

Candidate fit statistics were considered next. Fit statistics "enable the diagnosis of aberrant observations and idiosyncratic elements" (Linacre, 2018b:14) within each facet. Specifically relevant are the infit and outfit mean statistics, which can indicate misfit. They have an expected value of 1 and a range from zero to infinity where "the higher the . . . mean-square index, the more variability we can expect" in the rating patterns (Myford & Wolfe, 2000, p. 15). Here we only report infit, as it less sensitive to outliers compared to outfit and because it is broadly viewed as more important when evaluating the fitness of the data to the model (Eckes, 2009; Myford & Wolfe, 2004). Values below 1 are considered to be "overfitting" the model and too predictable, while values above 1 are considered to be "underfitting" and too unpredictable (Linacre, 2018b) with the latter generally raising more cause for concern (Eckes, 2009; Linacre, 2018b). In line with Linacre (2018b), the current study adopted lower and upper control limits of 0.5 to 1.5 for the infit mean square index.

Results in Table 3 show an average infit mean square value of M=0.80; SD=1.41 for the candidates with only 14 candidates out of 139 (approximately 10%) displaying underfit with infit values larger than 1.5.



```
|Measr|+Candidate|-Raters|Scale|
|-----+---------+-------+-----|
|  16 + *       +       + (5) |
|  14 +         +       + --- |
|  12 +         +       +     |
|  10 + .       +       +  4  |
|   8 + .       +       +     |
|   6 + .       +       +     |
|   4 + .       +       + --- |
|   2 + **      + 2 5   +  3  |
*   0 * **.     * 6 7 8 *     *
|  -2 + **.     + 3 4   + --- |
|  -4 + ***.    + 1     +  2  |
|  -6 + **      +       + --- |
|  -8 + *****.  +       +  1  |
| -10 + *****.  +       +     |
| -12 + *       +       + --- |
| -14 + *******. +      + (0) |
|-----+---------+-------+-----|
|Measr| * = 4   |-Raters|Scale|
```

Figure 2 Facets Wright map

Table 3 Candidate measurement results (summary)

Candidates *(n=139)*	Fair-M Average	Measure (Logits)	Model Error	Infit Mean Square (M)
Mean	1.47	-6.17	2.04	0.80
SD	1.26	7.28	1.21	1.41
RMSE 2.37 Adj (True) S.D. 6.89 Strata 4.21 Reliability .89				
Fixed (all same) chi-square: 2005.5 d.f.: 138 significance (probability): .00				

The rater facet was considered next. The Wright map in Figure 2 suggests a narrow range of rater severity compared to the distribution of candidate ability. This is confirmed in the rater measurement results in **Error! Reference source not found.** with the raters displaying a severity range of 6.02 logits from the most lenient rater (Rater 1=-3.08) to the most severe rater (Rater 5=2.94). In terms of fit, raters displayed an average infit mean square value of M=0.99; SD=0.27 with none of the raters exhibiting misfit (see **Error! Reference source not found.**).

Table 4 Rater measurement results

Rater	**Observed Average**	**Fair-M Average**	**Measure**	**Model S.E.**	**Infit Mean Square**
1	1.33	2.42	-3.08	0.47	1.36
4	1.5	2.02	-1.45	0.43	1.12
3	1.87	1.98	-1.30	0.42	1.09
6	1.46	1.66	0.09	0.44	1.28

7	1.66	1.51	0.62	0.43	0.77
8	1.66	1.47	0.77	0.43	0.52
2	1.46	1.31	1.40	0.42	0.72
5	1.21	1.07	2.94	0.44	1.08
Mean (N=8)	1.52	1.68	0	0.44	0.99
SD	0.19	0.41	1.76	0.02	0.27

RMSE 0.44 Adj (True) S.D. 1.71 Strata 3.92 Reliability .94

Fixed (all same) chi-square: 124.2 d.f.:7 significance (probability): .00

The overall fit of the data to the Rasch model was also evaluated by examining the table of unexpected responses. Model fit can be considered acceptable when the percentage of standardised residuals which fall outside of the range of −2.00 to 2.00 are 5% or less. Of the total number of valid responses (n=305), the table of unexpected responses only included one response with standardised residual outside of the acceptable range (0.32%<5%). Taken together, these results suggest a satisfactory fit of the data to the Rasch model and support the use of the speaking test in distinguishing between candidates from different ability levels (even within the relatively narrow range of CEFR levels observed among test takers). It was concluded that the level of the test was generally appropriate for its target population although the last two parts may prove somewhat difficult for pre-A1 learners. While the A2 category (i.e. Level 5 in Figure 2) was rarely used, a decision was made by Plan Ceibal to keep it so that the test could also capture the ability of the limited number of stronger candidates.

Rater questionnaires. The results of the rater questionnaire (n=8 in total) showed that the majority of raters found the revised version of the rating scale *somewhat useful* (n=7) with the descriptors distinguishing well between levels (n=6). Raters believed that the speaking test elicited rateable samples of speech *to some extent* (n=7) and *definitely* (n=1). They also believed audio quality to be generally clear for rating purposes. The two criteria of *language* and *delivery* were generally found to be useful although opinions on the usefulness of the *task fulfilment* criterion were more mixed. Linking descriptors to performances was found to be *somewhat difficult* by six of the raters with two finding it *easy*. Note that no rater training was offered at this point and raters' open comments suggested that benchmark performances, increased familiarity with the rating scale, and rater training would greatly enhance the rating process. Raters also provided suggestions for wording of descriptors and raised questions regarding the alignment of the maximum score on the rating scale with A2 descriptors on the CEFR. This feedback was considered in making further revisions to the rating scale in Phase 3 and it also informed the development of new rater training guidelines.

Student questionnaires. Findings from the surveys (n=139 in total) suggested a positive attitude towards the revised version of the speaking test. The majority of respondents found the instructions to be *clear* (95%), the speaking test to be *fun* (60%) or *OK* (37%), and the sound check to be *easy* (61%) or *OK* (28%). When asked about their favourite part of the test, parts 2 and 1 were selected by 43% and 30% respectively. The most difficult parts, on the other hand, were found to be Part 4 (59%) and Part 3 (30%). This is not surprising, as the score results suggested the majority of participants to be at the pre-A1 and lower A1 levels and the final two test parts were designed for higher ability levels (A1+ and A2). Nevertheless, the test was found to be *OK* (54%) or *easy* (30%) by the majority of respondents.

Students' open comments were coded by the Plan Ceibal team as either negative, neutral, positive, or irrelevant. Of the open comments provided (once irrelevant/no responses were removed), 63% were positive and 26% negative. Illustrative examples of open comments are included verbatim alongside translations provided by Plan Ceibal in Appendix D.

Even though students reported finding the speaking test challenging or difficult, they also found it to be fun and interesting, providing opportunities for practice. Some of the negative comments related to issues with audio quality and/or issues with listening ability. To minimise the (negative) effects of audio quality on performance, a decision was made by the project teams to provide written input as well as audio input in the final version of the speaking test and to evaluate the effectiveness of this decision in future validation studies.

Phase 3: the CEFR linking exercise

The aim of Phase 3 was to link the newly developed speaking test to an international framework (the CEFR). Here we present the main approach taken for the linking exercise and report on the results of each main stage focusing in particular on the challenges of defining and differentiating micro levels in CEFR Pre-A1 – A2.

Main approach

This linking phase was guided by the approach recommended in the *Manual for Relating Language Examinations to the CEFR* (Council of Europe, 2009) (hereafter referred to as the Council of Europe Manual), in which a diverse panel of experts met, discussed and built consensus as to how the content and level of the test related to the CEFR.

Because of the logistical challenges involved in setting up a meeting of researchers based in the United Kingdom with test developers and local experts working in Uruguay, an asynchronous "twin-panel" approach was taken (e.g. Green, 2012; Brunfaut & Harding, 2014). The panel meetings in both countries were coordinated by two UK researchers, who have extensive experience of conducting and participating in standard setting studies.

The Uruguay panel comprised:

- A member from the Administración Nacional de Educación Pública (Department of Language Policy of the National Administration of Education)
- Two members from Plan Ceibal with a range of responsibilities within the Plan Ceibal programme
- A member from Segundas Lenguas with a range of responsibilities in the Segundas Lenguas programme
- Three experts with knowledge of different areas of Applied Linguistics and language education.

The UK panel comprised:

- Four members of staff from a university-based language testing research centre, all specialising in language assessment. All of them have previously been involved in CEFR linking projects for other tests.
- Four experts with expertise in language assessment and knowledge of different areas of Applied Linguistics.

Eight of the 15 participants had taken part in CEFR linking exercises before (two in the Uruguay panel, six of the UK panel). Seven panellists held PhDs and a further five had Masters degrees in language assessment, language education or related fields.

The procedures for the linking exercise followed the five conceptual stages stipulated by Council of Europe Manual: *Familiarisation*, *Specification*, *Standardisation/Benchmarking, Standard Setting* and *Validation*. In the next sections, we will briefly provide an overview of each stage and present relevant results.

Familiarisation

As a pre-requisite for any effective linking to the CEFR and an essential preliminary to both Specification and Standardisation/Benchmarking stages (Council of Europe, 2009), *Familiarisation* involves training activities to ensure that participants in the linking process have a detailed knowledge of the CEFR, the levels, and the illustrative scales. This stage included self-access activities for all panellists and for the staff involved in completing the *Specification* forms prior to the panel meetings. The online self-access activities included reflecting on one's familiarity level with the CEFR, reading the relevant CEFR scales, observing and judging illustrative performances, and sorting exercises of descriptors from the CEFR. Some descriptors on the use of vocabulary and grammar were also included from the DIALANG project (Alderson, 2006) to provide more elaborated descriptors relevant to the target CEFR levels.

Fourteen of the 15 panellists completed this online activity (the remaining one reported that they had experienced technical difficulties using the online forms). One self-reported "poor", six "basic" and seven "good" knowledge of the CEFR at this point.

Overall, 63.2% of the 756 level judgements (i.e. 14 panellists assigning a CEFR level to 54 descriptors) were correct. 94.8% of judgements were within one band of the established level, which demonstrated a good degree of familiarity with the CEFR among the panellists. In addition, during the panel meetings in both countries, panellists carried out a descriptor-sorting exercise in groups to consolidate their understanding. Individual results were not recorded, but these judgements were all accurate to within one band.

Specification

Specification involves profiling the coverage of the test (content and task types) in relation to the categories in Chapters 4 (*Language use and the language learner*) and 5 (The *user/learner's competences*) in the original CEFR (Council of Europe, 2001). Both Uruguay and UK panels followed three steps, as explained below.

Profiling test-takers: Defining the three levels within A1. The panellists firstly worked in groups of two or three to come up with a definition of a learner at each of the following CEFR levels: Pre-A1, A1-, A1, A1+ and A2. It was necessary to create/specify the A1- and A1+ levels for the Uruguayan test-taker population, despite the absence of these levels in the CEFR. The key words that represent the three A1 levels are summarised in Table 5.

Table 5 Suggested key words to characterise A1-, A1, and A1+ levels

Level	Key words and phrases by panellists
A1+	Attempts more complex structures (but still within A1 range); communicates information about some personal and familiar matters; few pauses in unexpected places; L1 generally affects intelligibility; mostly intelligible but requires listener effort; links between ideas but not necessarily formally linked
A1	Simple, mainly isolated words, phrases and sentences; interact in a simple way; personal domain; very familiar topics; can link words using 'and', 'then'; much pausing; very strong L1 influence
A1-	Short isolated phrases about self, people and places, but with extremely limited range; frequent pauses and errors; utterances with omissions/reductions; errors in anything beyond formulaic expressions

The panellists found this task challenging, pointing to the difficulties of finding descriptors that adequately distinguished between three levels in one band. This was later confirmed in the results of the MFRM analyses leading to a decision and recommendation to reduce the three A1 levels to two within a band i.e. A1 and A1+. A suggestion was also put forward by panellists to establish a “below pre-A1” category in addition to “Not Rateable” where no speaking attempt was made. This was to include performances where some form of utterance (either in L1, very few isolated words or unintelligible words) was observed but did not meet the pre-A1 descriptors. Panellists reported that the distinction between Not Rateable, Below pre-A1, and Pre-A1 were clear enough and as such, Below pre-A1 was added for later exercises (see Appendix A).

Matching the test content to the CEFR. The panellists continued working in groups, and each group reviewed the *Specification* documents that had been prepared beforehand by members of Plan Ceibal and the language testing research centre in the UK. There was also a short explanation and demonstration of the online speaking test. The panellists then decided which test parts best matched specific CEFR levels and scales. Results are summarised in Table 6. The intended CEFR levels at the test development phase were generally confirmed by the panellists.

Table 6 Test Contents Mapped to the CEFR

Test part	Intended CEFR levels	CEFR levels assigned by panellists	Relevant CEFR scales
1	Pre-A1/A1	Pre-A1	○ Coherence and cohesion ○ Flexibility ○ General linguistic range ○ Grammatical accuracy ○ Information exchange ○ Overall spoken production ○ Spoken fluency ○ Sustained monologue ○ Propositional precision
2	Pre-A1/A1	Pre-A1/A1	
3	A1/A2	A1	
4	A2	A1/A2	

Matching the rating scale to the CEFR. Panellists were next asked to examine the rating scale, compare it to descriptors in the CEFR, and note those levels of the CEFR that corresponded most closely to each level category of the test (anonymised as X, Y, and Z). Results are presented in Table 7. As it can be seen, the CEFR level assignments by panellists match the intended levels in the scale.

Table 7 CEFR alignment indicated by panellists

Criteria	Task Fulfilment	Language	Delivery
Relevant CEFR scales	○ *Conversation* ○ *Information exchange* ○ *Overall spoken interaction* ○ *Overall spoken production* ○ *Propositional precision* ○ *Sustained monologue* ○ *Sociolinguistic appropriateness* ○ *Spoken fluency*	○ *Coherence and cohesion* ○ *General linguistic range* ○ *Grammatical accuracy* ○ *Overall spoken production* ○ *Overall spoken interaction* ○ *Sustained monologue* ○ *Vocabulary range* ○ *Vocabulary control*	○ *General linguistic range* ○ *Overall spoken production* ○ *Phonological control* ○ *Spoken fluency* ○ *Sustained monologue*
Level X	A2	A2	A2
Level Y	A1	A1	A1
Level Z	Pre-A1	Pre-A1	Pre-A1

Standardisation/Benchmarking

Standardisation was conducted to further build a common understanding of the CEFR levels, using 8 previously standardised CEFR illustrative samples from the Centre international d'études pédagogiques (CiEP: 2008) and CEFR-J[1]. Illustrative performances from CEFR-J were particularly relevant because of their focus on the lower levels. *Benchmarking* followed on from *Standardisation* and involved collective scoring of three speaking test performances on relevant CEFR scales.

Standard Setting

The panellists then used relevant CEFR scales to award CEFR levels including "plus" levels – to capture progress within levels, as recommended in Council of Europe (2001) – to 30 performance samples from the new Uruguay speaking test. These speech samples were selected from the results of the Phase 2 MFRM analyses. MFRM allows for an estimation of candidate ability independent of the particular rater assigned to them with the raw scores adjusted for any rater effects. The resulting candidate fair-average mark is a more objective estimate of the candidate ability. These 30 performances were selected based on (a) fair-averages close to desired score points and (b) acceptable fit statistics between 0.5 and 1.5 (see Table 8).

Table 8 Performance selection for CEFR panels

Estimated CEFR (based on Fair-M Average)	Number
Not Rateable (NR)	2
Below pre-A1	0
pre-A1	6
A1-	6
A1	6
A1+	6
A2 (and A2 borderline)	4
A2+	0
Total	**30**

The CEFR levels awarded by the panellists, both in Uruguay and in UK, were collated, converted to numerical values (i.e. Not Rateable = 1, Below Pre-A1 = 2, Pre-A1 = 3, …A2+ = 8) and analysed using MFRM (see Figure 3 for the Wright map). Initial analysis indicated that the 8 levels were not clearly distinguished by panellists. Therefore, some categories were collapsed. The 'Not Rateable' category was merged with 'Below Pre-A1', the 'A1-'category with 'A1', and 'A2+' with 'A2'.

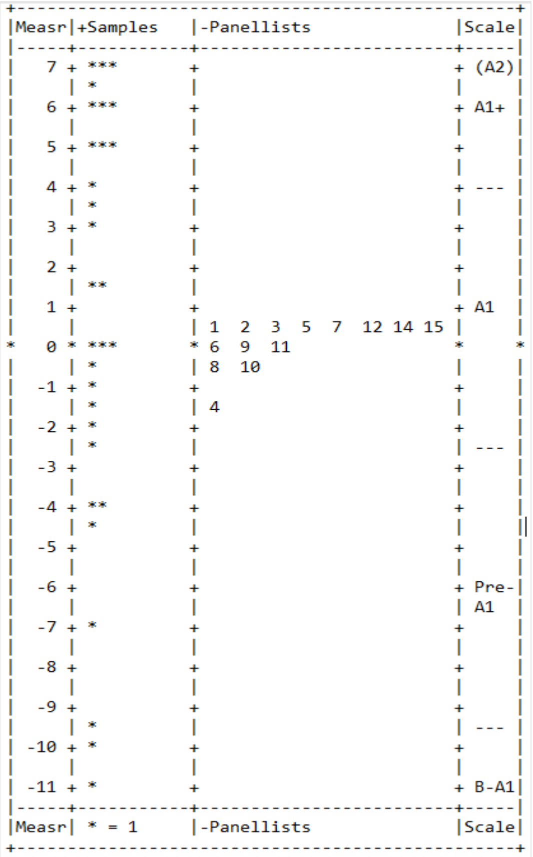
The results in Table 9 show that the majority of panellists displayed infit mean square values in the acceptable range of 0.5 to 1.5 (Linacre, 2018b). Panellist 12 is the only one displaying underfit (value>1.5) with Panellists 8, 2, and 1 showing overfit (values <0.5). None of them were removed from the analysis, prioritising the reliability by preserving as many data points as possible. One panellist (not listed the table) was removed from the analysis given (a) their significantly higher level of severity compared to other panellists (logit measure of 3.75, while the second most severe panellist had a logit measure of 0.3) and (b) high infit mean square value of 2.71 which indicated large inconsistencies in judgements.

Table 9 Panellist measurement report

Panellist ID	Measure	Fair-M	S.E	Infit MnSq
4	-1.84	3.04	0.42	0.98

[1] The Japanese adaptation of the CEFR. http://www.cefr-j.org/index.html

8	-0.9	2.96	0.84	0.21
10	-0.76	2.94	0.43	0.95
11	-0.38	2.9	0.43	0.71
9	-0.38	2.9	0.43	1.39
6	-0.38	2.9	0.43	0.61
15	0	2.85	0.44	1.28
7	0	2.85	0.44	1.09
5	0	2.85	0.44	0.74
2	0	2.85	0.44	0.42
14	0.19	2.82	0.44	1.4
3	0.19	2.82	0.44	1.38
1	0.19	2.82	0.44	0.4
12	0.31	2.80	0.63	1.6



```
+---------------------------------------------------+
|Measr|+Samples    |-Panellists               |Scale|
|-----+------------+--------------------------+-----|
|   7 + ***        +                          + (A2)|
|     | *          |                          |     |
|   6 + ***        +                          + A1+ |
|     |            |                          |     |
|   5 + ***        +                          +     |
|     |            |                          |     |
|   4 + *          +                          + --- |
|     | *          |                          |     |
|   3 + *          +                          +     |
|     |            |                          |     |
|   2 +            +                          +     |
|     | **         |                          |     |
|   1 +            +                          + A1  |
|     |            | 1  2  3  5  7  12 14 15  |     |
*   0 * ***        * 6  9  11                 *     *
|     | *          | 8  10                    |     |
|  -1 + *          +                          +     |
|     | *          | 4                        |     |
|  -2 + *          +                          +     |
|     | *          |                          | --- |
|  -3 +            +                          +     |
|     |            |                          |     |
|  -4 + **         +                          +     |
|     | *          |                          |     |
|  -5 +            +                          +     |
|     |            |                          |     |
|  -6 +            +                          + Pre-|
|     |            |                          | A1  |
|  -7 + *          +                          +     |
|     |            |                          |     |
|  -8 +            +                          +     |
|     |            |                          |     |
|  -9 +            +                          +     |
|     | *          |                          | --- |
| -10 + *          +                          +     |
|     |            |                          |     |
| -11 + *          +                          + B-A1|
|-----+------------+--------------------------+-----|
|Measr| * = 1      |-Panellists               |Scale|
+---------------------------------------------------+
```

Figure 3 Wright map of spoken performance samples rated by the panellists

The rater measurement report showed that there was only 0.24 logit's difference between the most lenient panellist (Panellist 4) with a fair-average score of 3.04 and most severe (Panellist 12) with 2.80. Moreover, the separation indices suggested very similar severity levels (H=1.09; R=0.24) for panellists, which taken together suggests the success of the *Familiarisation* process in ensuring acceptable levels of agreement between panellists. Lastly, the category probability curves (Figure 4) showed that the suggested number of categories are functioning well, supporting the merging of categories into five levels. The rating scale was subsequently revised in light of these findings taking into account the suggestions by the panellists. It is this final version of the scale that can be accessed in Appendix A.

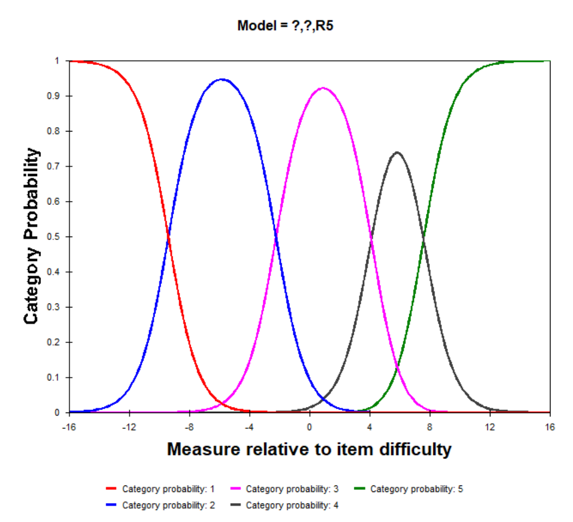


Figure 4 Category probability curves for five levels

Discussion and Conclusions

Towards building a validity argument

In this section differences sources of validity evidence are brought together to support the use of the Plan Ceibal speaking test for its intended purposes.

Accommodating test taker characteristics

The principles guiding test design were informed by a review of the literature on young learners and their characteristics as well as local teaching materials and curricula. This laid the foundation for contextual and cognitive parameters in the test. Many of the revisions and modifications to the test were informed by iterative discussions between the UK researchers and key personnel in Plan Ceibal, insights from experienced teachers and language testing practitioners in Uruguay, and feedback from a representative sample of the target population. Taken together, these can serve as evidence supporting the appropriateness of the speaking test for the target candidature within their local context.

Context validity evidence

Evidence for the context validity of the test is derived from the results of the student questionnaires in Phases 1 and 2 of the research. Findings generally suggested a positive attitude towards the test and different contextual parameters (e.g. timings, clarity of instructions, and technical aspects). Student feedback as well as results of MFRM analyses of score data also pointed to aspects of the test that were too difficult or cognitively too challenging for the target population. These were subsequently addressed through revisions and modifications to test specifications.

Scoring validity evidence

Evidence for the scoring validity of the test is derived from the results of the MFRM analyses of score data in different phases of the project. The candidate measurement results from Phase 2 showed that the speaking test was able to divide candidates into at least four distinct ability strata despite the narrow ability range of the participants (Pre-A1 to A2). Results from rater questionnaires pointed to some limitations of the draft rating scale but the majority of raters found the rating scale to be somewhat useful with descriptors distinguishing well between levels.

Criterion-related validity evidence

Criterion-related validity evidence for the test derives from the results of the CEFR linking exercise. The intended CEFR levels at the test development phase were generally confirmed by the expert panellists. The CEFR level assignments by panellists also matched the intended levels in the rating scale. These results support a link between a locally-developed test and an international framework.

Conclusion and future directions

The test development and validation research presented in this paper has provided preliminary validity evidence in support of the Plan Ceibal online speaking test for its intended purposes. At every stage of the project, research evidence, available technologies, and local and international expertise were brought together to build a test that would be appropriate for the unique characteristics of the target population, aligned to an international framework, and in harmony with its local context. In so doing, a model of collaboration between local and international partners was also established, facilitating the sharing of complementary expertise and knowledge.

Validation is an ongoing process and not all aspects of validity were examined in this study. Indeed, after the project reported here, Plan Ceibal has conducted a feasibility study with 469 students across 56 schools. They are also currently undertaking a comparative study between this new online speaking test and a face-to-face test of the same test design. Other possible avenues for future research include:

- An investigation of test washback in classrooms
- A large-scale survey of key stakeholders once the test is operationalised at the national level
- An investigation into students' response processes as they take the test using methodologies such as eye-tracking

Future research into the *consequential validity* of the test is of paramount importance in order to evaluate the extent to which intended consequences of this new test are realised. As in the first two lines of future research listed above, it is essential to report and track any changes observed in classroom practices in relation to teaching and learning of speaking skills. It is equally important to identify any changes in the perceptions of students, teachers, parents, material writers, and other stakeholders of the English education system in Uruguay towards the learning and assessment of spoken skills. Finally, while this project addressed *cognitive validity* considerations through test design, various scaffolding features, and recommendations from the Plan Ceibal team, empirical validation of cognitive validity was outside the scope of the research. Gauging the cognitive validity of the test through empirical research – investigating the actual processes of spoken production of learners during the test and seeking the optimal cognitive demands in each test task – is therefore needed to further consolidate the validity argument of this new test.

Despite the need for future research and further refinements of the test, this project has taken a first step towards promoting speaking skills in primary and secondary education in Uruguay. While the test was developed to suit a specific local context, we believe that the challenges associated with designing and delivering an online test for young learners and differentiating micro levels within a narrow range of language ability (e.g. pre-A1 to A2) may well apply to other educational contexts. We hope that this paper will contribute to and encourage dialogue among language testing researchers, local stakeholders, and practitioners on the development of spoken assessments that meet both local requirements and international standards.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA

Alderson, J.C. (2006). Diagnosing foreign language proficiency: The interface between learning and assessment. London/NY: Continuum.

Alderson, J. C., & Banerjee, J. (2001). Language testing and assessment (Part I) *Language Teaching 34*(4): 213–236.

Bailey, A. L. (2017). Progressions of a new language: Characterizing explanation development for assessment with young language learners. *Annual Review of Applied Linguistics, 37*, 241–263.

Benigno, V., & de Jong, J. (2016). A CEFR-based inventory of YL descriptors: Principles and challenges. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 43–64). Heidelberg, Germany: Springer.

Brunfaut, T., & Harding, L. (2014). *Linking the GEPT listening test to the Common European Framework of Reference.* LTTC-GEPT Research Reports *RG-05.* Taipei, Taiwan: Language Training and Testing Centre.

Butler, Y. G. (2016). Assessing young learners. In Tsagari, D &Banerjee, J. (Eds). *Handbook of second language assessment* (pp. 359–376). De Gruyter Mouton.

Butler, Y. G. (2019). Assessment of young English learners in instructional settings. In X.Gao (Ed) *Second Handbook of English Language Teaching*. (pp. 477–496). Springer.

Butler, Y. G., & Zeng, W. (2014). Young foreign language learners' interactions during task-based paired assessments. *Language Assessment Quarterly, 11*(1), 45–75.

Cho, Y., Ginsburgh, M., Morgan, R., Moulder, B., Xi, X., & Hauck, M. C. (2016). *Designing the TOEFL® Primary™ Tests*. (Research Memorandum No. RM-16-02). Princeton, NJ: Educational Testing Service.

Cizek, G J (2011) *Reconceptualizing validity and the place of consequences*, paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA, 9–11 April 2011.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*, Strasbourg: Council of Europe.

Council of Europe (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual,* Strasbourg: Council of Europe.

Council of Europe (2018) *Common European Framework of Reference for Languages: Learning, teaching, assessment (Companion volume with new descriptors)*, Strasbourg: Council of Europe.

Field, J. (2018). The cognitive validity of tests of listening and speaking designed for young learners. In S. Papp & S. Rixon (Eds.). *Examining young learners: Research and practice in assessing the English of school-age learners.* Studies in Language Testing vol. 47 (pp. 128–200). Cambridge: UCLES/Cambridge University Press.

Green, A. B. (2012). *Conducting CEFR Benchmarking for the Versant English Placement Test: Final Project Report*. NCS Pearson.

Hasselgreen, A., Caudwell, & G. (2016). *Assessing the language of young learners*. British council monographs on modern language testing. Sheffield/Bristol, UK: Equinox

Hasselgreen, A. (2003). *Bergen 'Can Do' project*. Strasbourg, France: Council of Europe. Retrieved from http://archive.ecml.at/documents/pub221E2003_Hasselgreen.pdf (accessed 15 November 2021)

Holzknecht, F. (2019). *Double play in listening assessment*. Unpublished PhD Thesis, Lancaster University (United Kingdom).

Huang, B. H., Bailey, A. L., Sass, D. A., & Shawn Chang, Y. (2021). An investigation of the validity of a speaking assessment for adolescent English language learners. *Language Testing, 38*(3), 401–428.

Huang, B., Chang, Y.H.S., Niu, L., & Zhi, M. (2018). Examining the effects of socio-economic status and language input on adolescent English learners' speech production outcomes. *System, 73*, 27–36.

Inbar-Lourie, O., & Shohamy, E. (2009). Assessing young language learners: What is the construct? In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 83–96). Berlin, Germany: Mouton de Gruyter.

Isbell, D., & Winke, P. (2019). ACTFL Oral proficiency interview–computer (OPIc). *Language Testing*, *36* (3), 467–477.

Johnstone, R. (2009). An early start: What are the key conditions for generalized success? In J. Enever, J. Moon, & U. Raman (Eds.), *Young learner English language policy and implementation: international perspectives* (pp. 31–42). Reading, UK: Garnet Education Publishing.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Lee, S., & Winke, P. (2018). Young learners' response processes when taking computerized tasks for speaking assessment. *Language Testing, 35*(2), 239–269.

Linacre, J. M. (2018a). Facets Rasch measurement computer program (version 3.81) [computer software]

Linacre, J. M. (2018b). *A user's guide to Facets: Rasch model computer programs*

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal, 91*(4), 645–655.

Macaro, E., Handley, Z., & Walter, C. (2012). A systematic review of CALL in English as a second language: Focus on primary and secondary education. *Language Teaching, 45*(1), 1–43.

McKay, P. (2006). *Assessing young language learners*. Cambridge, UK: Cambridge University Press

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York: American Council on Education, Macmillan

Nikolov, M., & Timpe-Laughlin, V. (2021). Assessing young learners' foreign language abilities. *Language Teaching, 54*(1), 1–37.

Papp, S., & Rixon, S. (2018). *Examining young learners: Research and practice in assessing the English of school-age learners.* Studies in Language Testing vol. 47. Cambridge: UCLES/Cambridge University Press.

Papp, S., & Walczak, A. (2016). The development and validation of a computer-based test of English for young learners: Cambridge English young learners. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 139–190). Heidelberg, Germany: Springer.

So, Y., Wolf, M. K., Hauck, M. C., Mollaun, P., Rybinski, P., Tumposky, D., & Wang, J. (2015). *TOEFL® Junior Design Framework.* TOEFL Junior Research Report No. 02. ETS Research Report, No. RR-15-13. Princeton, NJ: Educational Testing Service.

Szabó, T. (2018a). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Vol. 1: Ages 7–10: Collated representative samples of descriptors of language competences developed for young learners.* Retrieved from https://rm.coe.int/collated-representative-samples-descriptors-young-learners-volume-1-ag/16808b1688 (accessed 15 November 2021)

Szabó, T. (2018b). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Collated representative samples of descriptors of language competences developed for young learners aged 11–15 years*. Retrieved from https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680697fc9 (accessed 15 November 2021)

Taylor, L. (Ed.) (2011) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Cambridge: Cambridge University Press.

Thompson, G. L., Cox, T. L., & Knapp, N. (2016). Comparing the OPI and the OPIc: The effect of test method on oral proficiency scores and student preference. *Foreign Language Annals*, *49*(1), 75–92.

Weir, C. J. (2005). *Language testing and validation*. Basingstoke: Palgrave Macmillan.

Weir, C. J., Vidaković, I., & Galaczi, E. D. (2013). *Measured constructs: A history of Cambridge English Examinations, 1913-2012.* Cambridge: Cambridge University Press.

Appendix A: Operational Rating Scale

Level	Task Fulfilment	Grammar & Vocabulary	Delivery (Intelligibility & Flow of Speech)
A2 and above (Score of 9)	• Responds to **all** test parts. • Meaning is communicated **successfully**. (3)	• Uses **a range** of familiar words and some unfamiliar words may also be attempted. • Uses **a range** of simple structures with some limited attempt at more complex structures. • Errors persist. (3)	• **Almost all** words/phrases are intelligible. • **Keeps going** for a stretch of speech though **pauses persist**. (3)
A1 (Score of 6)	• Attempts at responding to **most (or all)** test parts. • Meaning is communicated with **a degree of success**. (2)	• Uses **familiar words** and **simple phrases** • Uses basic structures in **a learnt repertoire**. • Errors persist. (2)	• **Some** words/phrases are intelligible. • Produces words/phrases with **some** flow but **pauses are frequent.** (2)
Pre-A1 (Score of 3)	• Attempts at responding to **Parts 1 and 2. Limited/no attempt** at responding to **Parts 3 and 4**. • **Limited success** in communicating meaning. (1)	• Uses some **familiar** isolated words (e.g. football, name, animal). • Limited or no grammar use unless as part of a memorised formulaic expression (e.g. my name is…) • **Frequent** errors. (1)	• **A few** words/phrases are intelligible. • Produces some memorised words/phrases with **limited** flow but **filled/silent pauses** are the **default** state. (1)
Below Pre-A1 (0)	• No language in English is produced; primarily filled pauses or L1 **OR** • Very few isolated words and/or formulaic expressions may be produced but these are rare and may be unintelligible.		
TE	• Technical Error → re-take the test		

Note: Instructions for rating

- We suggest an **analytic approach** where scores are assigned to the three different categories independently. This approach facilitates the scoring of candidates who display different profiles e.g. a candidate that can keep going (under A2 descriptor for *Delivery*) but uses basic structures (under A1 descriptor for *Grammar and Vocabulary*).
- If there are technical problems where audio is not recorded or inaudible, a mechanism should be in place to identify these cases and allow re-sits (e.g. Technical Error coded as **TE**). This should be distinguished from a score of **0** as described in the table.
- The rating scale covers a score range from **0 (below Pre-A1)** to **9 (A2)**. The scoring system facilitates distinguishing between candidates for reporting purposes. For each descriptor, a score is provided in parentheses. A conversion table with recommendations for CEFR reporting is provided below.
- In order to qualify meeting CEFR levels **A1** and **A2**, the principle adopted is for candidates meet a **minimum** of two sets of descriptors within that specific CEFR level.

Score Range	Suggested CEFR level for reporting
0-1	Below Pre-A1
2-4	Pre-A1

5-6	A1
7	A1+
8-9	A2

Appendix B: Candidate Questionnaire

Q1 - How did you like the test? [Fun, OK, Boring]
Q2 - The sound check was… [Easy, OK, Difficult]
Q3 - The instructions in Spanish were… [Clear, Not clear]
Q4 - My favourite part was… [Part 1, Part 2, Part 3, Part 4]
Q5 - The most difficult part was… [Part 1, Part 2, Part 3, Part 4]
Q6 - The test in general was… [Easy, OK, Difficult]
Q7 - Do you want to tell us something more about the test? [   ]
Q8 - How old are you? [10, 11, 12, 13, 14]
Q9 - Do you like speaking English? Why? [   ]

Appendix C: Rater questionnaire

1. How did you find the rating scale overall?

   □not useful □somewhat useful □useful □very useful

2. How did you find the task fulfilment criterion?

   □not useful □somewhat useful □useful □very useful

3. How did you find the language criterion?

   □not useful □somewhat useful □useful □very useful

4. How did you find the delivery criterion?

   □not useful □somewhat useful □useful □very useful

5. Did you encounter jagged profile students i.e. those much stronger in one criterion (e.g. delivery) but weaker in others (e.g. language) who made rating difficult?

   □no, not at all □rarely □sometimes □all the time

6. Do you think the descriptors for each score point distinguished well between levels e.g. A1 from A1+ and so on?

   □No, not at all □Yes, somewhat □Yes, very well

7. How easy was it to link descriptors to candidate performances?

   □very difficult □somewhat difficult □easy □very easy

8. How did you find the quality of the audio recordings for rating?

   □not clear at all □somewhat clear □clear □very clear

9. Do you think the test elicited enough ratable samples of speech?

   □no, not at all, □yes, to some extent □yes, definitely

10. Any further comments? (Please specify in English)

Appendix D: Illustrative examples of candidate open comments

Positive comments:

Comments (verbatim)	*Approximate translation*
Me parecio muy facil pero tambien un poco dificil a veces, pero me gusto mucho	I found it very easy but also a little difficult sometimes, but I liked it a lot.
estubo entretenida	entertaining
Me gusto bastante	liked it a lot
estuvo buena en el caso que energese la lectura y la oralidad en ingles	fine because you practice reading and speaking English
fue dificil en algunas partes, pero me diverti	some parts were difficult but I had fun
Fue bastante intuitivo y facil, me gusto bastante	quite intuitive and easy, I quite liked it

Negative comments:

Comments (verbatim)	*Approximate translation*
Hola , a mi me parecio algo dificil	a little difficult for me
no la entendi	I didn't understand it
Tendran que megorar el audio	audio needs improvement
la brueba fue buena pero habian cosas que no sabia decir en ingels y me paresia muy dificil	the test was good but there were things I didn't know, it was very difficult
Lo que decia bob no se entendia	didn't understand Bob
lo que dijo alicia no fue tan entendible	what Alicia says is not that clear