LanguageScreen: The Development, Validation, and Standardization of an Automated Language Assessment App (Hulme et al., 2024): A Summary

Read Full Paper

Why was LanguageScreen developed?

Several tools are available for language screening and assessment in English (e.g., Law et al., 1998), although it is notable that most focus on preschool-aged children. A U.S. review lists some 24 screening tests, but few are suitable for children above 5 years of age, and usually, these need to be administered by a trained professional (Berkman et al., 2015).

The language screening tests available for children of school age are typically rating scales for use by teachers or parents. Provided teachers are trained in their use, rating scales can provide valid metrics (e.g., Duff & Clarke, 2011; Seager & Abbot-Smith, 2017); however, such scales:

- involve a degree of subjectivity
- are susceptible to expectancy bias
- show reduced reliability if different assessors are involved.

LanguageScreen was developed to provide education professionals with a quick and accurate way of assessing children's language skills, with a particular emphasis on identifying children who would likely benefit from language support.

What are the advantages of LanguageScreen over existing assessments?

The critical advantages of LanguageScreen compared to other available language screening measures include the following:

- It involves direct assessments of children's language skills rather than relying on ratings that may be biased.
- Testing is automated, reducing possible tester bias and increasing reliability.
- Automated scoring and reporting reduces testing time and avoids errors.
- It is suitable for a wide range of ages, spanning the preschool and school years (from 3;05 to 12;00).
- The test is easy to use and can be administered by adults without any special training.
- The test has excellent reliability.
- The test has been validated against well-standardized measures of language ability that are both more expensive and more difficult to use.

How were the LanguageScreen subtests chosen?

The choice of subtests was based on those used commonly for the assessment and diagnosis of language disorders (e.g. Tomblin and Zhang, 2006).

How were the items within subtests chosen?

Initial selection of items was guided by linguistic and psycholinguistic factors. Subsequently, based on extensive pilot data, items were retained or replaced to ensure good coverage of the range of ability targeted by the test.

Expressive Vocabulary

The starting point for this test was a graded set of 20 items for naming (from Snowling et al., 1988), supplemented by items chosen from "age of acquisition" tables (Ellis & Morrison, 1998). Pictures of the items that were considered unambiguous were arranged in order of difficulty for piloting.

The test contains 24 items arranged in order of difficulty. The child must name each item aloud. Items vary in age of acquisition from c. 2 – 11.5 years (Morrison & Ellis, 2000).

Receptive Vocabulary

The child sees a set of 4 pictures on the screen and must point to the picture that matches a word spoken by the App. There are 23 target items ranging in age of acquisition from c. 2 – 11.5 years (Morrison & Ellis, 2000).

The choice of target items for the receptive vocabulary test followed the same process as for the expressive language test. Following the work of Snowling et al. (1988), each target is paired with:

- a similar-sounding (phonological) distractor
- a meaning-related (semantic) distractor
- an unrelated distractor.

Sentence Repetition

The child hears a sequence of 14 spoken sentences and has to repeat each sentence verbatim. The sentences vary in length and complexity and are arranged in order of difficulty.

Twenty-two items were piloted, being chosen to reflect a range of sentence structures from an experimental sentence repetition test; these were arranged in order of difficulty according to data from 260 children assessed at the ages of 6 and 8 years participating in the Wellcome Language and Reading Project (Snowling et al., 2019). Accuracy was scored following each item (correct/incorrect), and a single error made by the child rendered that item incorrect. Following item analyses, 14 items were chosen for use in the app.

Listening Comprehension

The child hears a sequence of short, spoken passages, and each passage is followed by a sequence of spoken questions.

The listening comprehension test is an adapted version of one used in an evaluation of the Nuffield Early Language Intervention program (Fricke et al., 2013). There are 16 questions that include both literal (factual) and inferential questions.

How was LanguageScreen standardised and validated?

Standardisation sample

Approximately 350,000 children were included in the standardisation sample, spanning ages 3;06 to 8;11.

Results

There was a gradual increase in raw scores with age, with particularly steep increases across the youngest age groups (42–60 months). The test is relatively free from ceiling effects, and even in the oldest age group, just two out of 759 children obtained the maximum score of 77. It is notable that in each age band, there is a significant tail representing children with language difficulties.

Reliability analysis

The LanguageScreen total score gives the best estimate of a child's language level. LanguageScreen showed good-to-excellent reliability for the total scale (α = .92, Rasch Person Separation Reliability .94).

Validation

As part of a randomised controlled trial, 1,156 children who were the five lowest scoring children in each classroom (20% of all children assessed at pretest) were reassessed by speech and language therapists using well-standardized measures of language ability (Clinical Evaluation of Language Fundamentals [CELF] Expressive Vocabulary and Recalling Sentences [Semel et al., 2006], Renfrew Action Picture Test Information and Grammar [Renfrew, 2003]). These assessments by the speech and language therapists took approximately 40 min per child.

The total raw score from the speech and language therapist language assessment scores correlated highly with the LanguageScreen total raw scores (r = .74). Using latent variable analyses to exclude measurement error, a LanguageScreen factor correlated almost perfectly (*r* = .95) with a latent variable defined by the 4 individually administered language measures (West et al., 2021).

In addition, LanguageScreen showed comparable gains in scores in the children who had received intervention to the gains shown on the tests individually administered by professionals.

To summarise, LanguageScreen correlates well with much longer, well-standardized tests of language ability and is sensitive to improvements in language skills brought about by intervention.

Is LanguageScreen a diagnostic tool?

The ease of use of LanguageScreen puts a reliable assessment of children's oral language directly into the hands of educators either to screen the whole class, as an initial step in assessing a child's special educational needs, or for monitoring children's language development over time.

It is important to emphasize, however, that a school-based assessment tool cannot replace professional input from skilled speech and language professionals. Instead, LanguageScreen might be used to foster collaborative practice between educators and other professional services as they work together to provide effective support for children with language difficulties.

For example, a specialist therapist could scrutinize screening data with a teacher and, together, make decisions about identifying children for further language assessment or to put in place appropriate interventions. Along similar lines, a school (educational) psychologist could make use of such data when consulted regarding a child's emotional or behavioural difficulties as a check on possible underlying causes.