ASHA American Speech-Language-Hearing Association

LSHSS

Research Article

LanguageScreen: The Development, Validation, and Standardization of an Automated Language Assessment App

Charles Hulme,[a,g] Joshua McGrane,[b] Mihaela Duta,[c] Gillian West,[d] Denise Cripps,[e] Abhishek Dasgupta,[c] Sarah Hearne,[a] Rachel Gardner,[a] and Margaret Snowling[e,f]

[a]Department of Education, University of Oxford, United Kingdom [b]Assessment and Evaluation Research Centre, The University of Melbourne, Victoria, Australia [c]Department of Computer Science, University of Oxford, United Kingdom [d]Department of Language and Cognition, University College London, United Kingdom [e]St. John's College, Oxford, United Kingdom [f]Department of Experimental Psychology, University of Oxford, United Kingdom [g]Department of Psychology, Health and Professional Development, Oxford Brookes University, United Kingdom

ARTICLE INFO

Article History:
Received January 12, 2024
Revision received March 16, 2024
Accepted April 3, 2024

Editor-in-Chief: Kelly Farquharson

https://doi.org/10.1044/2024_LSHSS-24-00004

ABSTRACT

Purpose: Oral language skills provide a critical foundation for formal education and especially for the development of children's literacy (reading and spelling) skills. It is therefore important for teachers to be able to assess children's language skills, especially if they are concerned about their learning. We report the development and standardization of a mobile app—LanguageScreen—that can be used by education professionals to assess children's language ability.
Method: The standardization sample included data from approximately 350,000 children aged 3;06 (years;months) to 8;11 who were screened for receptive and expressive language skills using LanguageScreen. Rasch scaling was used to select items of appropriate difficulty on a single unidimensional scale.
Results: LanguageScreen has excellent psychometric properties, including high reliability, good fit to the Rasch model, and minimal differential item functioning across key student groups. Girls outperformed boys, and children with English as an additional language scored less well compared to monolingual English speakers.
Conclusions: LanguageScreen provides an easy-to-use, reliable, child-friendly means of identifying children with language difficulties. Its use in schools may serve to raise teachers' awareness of variations in language skills and their importance for educational practice.

Language skills are fundamental to many aspects of cognitive and psychosocial development and provide a critical foundation for formal education. More specifically, language skills are vital for the development of word reading and reading comprehension (Hjetland et al., 2020; Hulme et al., 2015) as well as numeracy and mathematical skills (Chow & Ekholm, 2019; Hornburg et al., 2018). Language is also crucial for social and emotional development and the ability to make friends and regulate behavior (Conti-Ramsden & Botting, 2004; Norbury et al., 2016; I. T. Petersen et al., 2013; Van Agt et al., 2011). In this light, it is unsurprising that children with language difficulties are at risk of poor educational achievement (Stothard, et al., 1998), as well as longer term psychosocial difficulties (Clegg et al., 2005; Conti-Ramsden et al., 2018; Winstanley et al., 2018) including offending behaviors (Chow et al., 2022). Together, these findings suggest that identifying language difficulties in the early years of schooling is important as an initial step toward intervening to try to prevent a downward spiral of poor education and reduced life chances.

Language difficulties are common, and population estimates suggest that 7% of children enter school with identifiable language disorders (Norbury et al., 2016;

Correspondence to Charles Hulme: charles.hulme@education.ox.ac.uk. ***Disclosure:*** *Charles Hulme, Margaret Snowling, and Gillian West are directors of OxEd and Assessment Ltd, a University of Oxford spinout company founded to distribute LanguageScreen as a commercial product. Mihaela Duta is a shareholder in OxEd and Assessment Ltd. All other authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

Reilly et al., 2010; Tomblin et al., 1997). Whereas some 2%–3% have speech difficulties (Bishop & Hayiou-Thomas, 2008), others have less easily observable problems affecting receptive and expressive language skills. Such language difficulties (previously known as specific language impairment and more recently known as developmental language disorder [DLD]) may go unnoticed in the classroom where a child can often follow teacher instructions by copying others. According to Zhang and Tomblin (2000), referral and eligibility for service delivery are more likely in the case of children with speech than in the case of children with language difficulties, and an issue of concern is that DLDs may only come to light after a child is referred for treatment of a behavioral disorder (Bishop et al., 2016; Kaiser et al., 2022; McGregor, 2020; Paul, 2007; Tannock & Schachar, 1996). These findings provide a strong rationale for screening to allow the early identification of oral language difficulties (Hendricks et al., 2019) as well as for assessing possible causes of poor response to reading intervention (Fuchs et al., 2012; D. B. Petersen & Spencer, 2012). Here, we describe the development and standardization of a mobile app that can be used to assess children's language as an initial step toward delivering appropriate interventions.

Language Screening Tests

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-age 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). Similarly, a review providing guidance for Welsh schools lists screeners suitable for bilingual children as well as for those speaking primarily English or Welsh but only for children below the age of 5 years (Baker et al., 2022).

Language Screening in Preschool

An important aim of preschool language screening is to identify children with language and communication disorders who will benefit from early intervention (e.g., Korpilahti et al., 2016). However, since there are many different trajectories of language development, with some children showing early delays that resolve and others having language difficulties that emerge later, the field remains somewhat divided as to the utility of screening in preschool. Two recent systematic reviews have addressed this issue: Wallace et al. (2015) evaluated the use of screening for speech and language delays in preschool children (aged 5;00 [years;months] and younger) with a view to their use in primary care settings. The sensitivity of many existing measures was not good, ranging from 50% to 94%, with specificity ranging from 45% to 96%. So and To (2022) conducted a meta-analysis of 67 screening tools for children aged 6;00 and younger; these tools used direct language assessment, clinical markers derived from parent observations, or both. Findings from a meta-regression found that only about one third of the tests obtained fair accuracy and 14% obtained good accuracy in detecting language disorder; those based on children's language ability were more sensitive than those that used clinical markers, and screening at above the age of 4 years was more accurate than screening at younger ages.

More recently, Holzinger et al. (2022) have reported the development of a screening instrument for German children in the penultimate year of preschool. The combination of scores from a brief 10-min assessment of receptive and expressive grammar demonstrated excellent accuracy, with good sensitivity and specificity, when predicting language disorder (operationalized by performance at least 1.25 *SD*s below average on two standardized language tests) in a sample of 374 children. These findings suggest that the direct assessment of language may be useful prior to school entry.

Language Screening in School

Arguably, there is more agreement on the importance of language screening for school-age children (Adlof & Hogan, 2019). The language screening tests available for children of school age are typically rating scales for use by teachers or parents. Examples include the Classroom Communication and Learning Checklist (Wiig & Secord, 1994) and the Children's Communication Checklist–Second Edition (Norbury et al., 2004). Additionally, in England, schools are required to use the Language and Literacy scales of the Reception Baseline Assessment completed in the first weeks of school and the Communication and Language scales of the Early Years Foundation Stage Profile completed at the end of the first school year. 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, they are susceptible to expectancy bias, and reliability is reduced further if different assessors are involved.

Due to potential limitations in the accuracy of rating scales, there are advantages in using direct assessments of children's language skills. There are, however, very few such measures that are available for school-age children. One of the few such measures is the WellComm toolkit (https://www.gl-assessment.co.uk/assessments/products/wellcomm/), which combines direct assessments of children's language skills with ratings of communication skills.

This review brings out clearly that currently available language screening tools could be improved. Here, we describe the development of a language assessment

app—LanguageScreen (OxEd and Assessment Ltd, 2022)—that can be used by education professionals. We believe that the critical advantages of LanguageScreen compared to other available language screening measures include the following: (a) It involves direct assessments of children's language skills rather than relying on ratings that may be biased; (b) testing is automated, reducing possible tester bias and increasing reliability; (c) automated scoring and reporting reduces testing time and avoids errors; (d) it is suitable for a wide range of ages, spanning the preschool and school years (from 3;05 to 9;00); (e) the test is easy to use and can be used by adults without any special training; (f) the test has excellent reliability; and (g) the test has been validated against well-standardized measures of language ability that are both more expensive and more difficult to use.

Rationale for the Study

In developing LanguageScreen, we set out to design an objective and reliable test that would be quick and easy to administer and that would provide data that could be used to produce a single score indicative of a child's language proficiency. Accordingly, we decided to follow a Rasch measurement theory approach throughout the piloting and final analysis of the test. The test was designed to be suitable for children from preschool to middle school years and to sample both receptive and expressive language skills.

The development of LanguageScreen was guided by evidence that, at least in the early years of schooling, language skills are well described by a single latent factor. Tomblin and Zhang (2006) found that a unidimensional language factor accounted well for scores from a range of tests assessing receptive vocabulary (picture identification), expressive vocabulary (providing definitions for spoken words), receptive grammatical skills, and expressive language use (grammatical completion and sentence imitation) in 6-year-old children. Similarly, Klem et al. (2015) found that a unidimensional language factor (defined by loadings from sentence repetition, vocabulary knowledge, and grammatical skills) provided an excellent fit to their data from a large sample of 4- to 5-year-old children. The unitary language factor identified by Klem et al. showed a high degree of longitudinal stability, as did a unitary language latent variable identified in a study by West et al. (2021).

The extent to which language ability can be considered a unitary trait will depend upon both the sample and the measures used. A study by the Language and Reading Research Consortium (2015) assessed language ability with a wide range of measures in samples of prekindergarten; kindergarten; and Grades 1, 2, and 3 children. In the prekindergarten sample, a comparison of one-, two-, and three-factor confirmatory factor analysis models indicated that a one-factor model was to be preferred. However, in the later age groups, there was a suggestion of a gradual differentiation of language ability into three factors by Grade 3 (grammar, vocabulary, and discourse), although it should be noted that the three factors correlated highly with each other in all age groups and, as the authors noted, in relation to the Grade 3 findings: "While there may be a statistical preference for a three-dimensional model [. . .] there were also relatively strong correlations between these constructs" (pp. 1960–1961). It is also worth noting that some of the measures of discourse used here (e.g., detecting inconsistencies between different parts of a spoken passage) are not typical measures of oral language comprehension and may involve high levels of attention, thus reducing their correlation with the other measures of language ability used (such as naming pictures or selecting a picture from a set of four to match a spoken word).

Our choice of domains to be included in the LanguageScreen app was based on those used commonly for the assessment and diagnosis of language disorders (e.g., Tomblin & Zhang, 2006). Subscales were constructed to assess expressive vocabulary, receptive vocabulary, grammar (sentence repetition), and listening (narrative) comprehension. The Rasch model, which was used here to guide item selection and subsequent analyses, assumes that items form a unidimensional scale. One critical advantage of a test that conforms to the Rasch model is that the total score is a sufficient statistic (i.e., the total score provides all the information needed to assess the ability of an individual; see Andrich, 2005).

Data collected from assessments of a large sample of preschool- and school-age children also allowed us to test two other hypotheses:

1. Gender differences in language skills: We expected girls to perform better than boys on the test but only to a small extent (see, e.g., Wallentin, 2020).
2. The effects of bilingualism on proficiency in the school language (English): We expected children with English as an additional language (EAL) to have lower scores on an English language assessment, such as LanguageScreen, compared to their monolingual English peers due to lower levels of language exposure (see, e.g., Whiteside & Norbury, 2017).

Method

Design

Four principles guided the development of LanguageScreen: (a) It should be quick and easy to use, (b) it would be implemented in an app that enabled automatic

scoring to reduce errors and burden on testers, (c) the app would run on Android and Apple devices, and (d) the test would be a reliable and valid measure of language skills suitable for use with children between 3 and 9 years of age. Validity would be assessed by reference to well-established standardized measures of language in current use.

Study Dates

A pilot set of the items to be used to assess the four language domains was assembled in 2016. The pilot version of the app, referred to as ATLAS (Automated Test of Language Abilities), was implemented as a mobile app; permission for the research and development of the app was granted by the Central University Research Ethics Committee of the University of Oxford in 2017. Data collection commenced in 2018, and the analyses reported here used data collected following the release of the revised app (renamed as LanguageScreen) in 2020.

Test Development and Content

LanguageScreen was developed to provide education professionals 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. 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. Pictures were selected as being culturally appropriate for the British context. It is acknowledged that adaptations to the test may be required for use in different cultures.

Expressive Vocabulary (EV). 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. When implemented in the app, the child sees a series of stylized colored pictures and is asked to name each one. The assessor presses a button on the screen to indicate whether the response is correct or incorrect. The test contains 24 items ranging in age of acquisition from 22.1 to 140 months (Morrison & Ellis, 2000): bed, castle, ladder, umbrella, bell, glove, sword, drawer, scarecrow, whale, volcano, fence, wheelbarrow, acorn, plug, anchor, stool, handcuffs, parachute, eyelash, envelope, needle, stethoscope, and pliers. Testing discontinues after eight consecutive errors.

Receptive Vocabulary (RV). 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 was paired with a similar-sounding (phonological) distractor, a meaning-related (semantic) distractor, and an unrelated distractor (see Table 1). The selection of distractors was based on confusability with reference to phonology and semantics; distractors were not closely matched for frequency of occurrence with the targets. When implemented in the app, the child hears a word and is asked to touch one of the four stylized colored pictures that corresponds to the word presented. There are 23 items ranging in age of acquisition from 22.1 to 140 months (Morrison & Ellis, 2000). Testing discontinues after eight consecutive errors.

Sentence Repetition (SR). For this test, the child hears a spoken sentence and is asked to repeat it verbatim. 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 (see Table 2). Testing discontinues after five consecutive errors.

Listening Comprehension (LC). 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). The child hears three short stories (without pictorial support), and immediately after hearing each story, they answer questions posed about the content of the story. There are 16 questions that include both literal (factual) and inferential questions. The examiner is presented with acceptable responses for each question on the screen to facilitate scoring. Each question is scored as correct/incorrect (1/0) by tapping buttons on the screen. Testing continues provided the child answers at least one question correctly on the first two passages (the test is discontinued if they answer all questions on the first two passages incorrectly).

Administration

The app and website, to which data are uploaded, are designed to be highly secure. To use the app for assessments, the user first creates an account and enters the details of the children to be assessed (name, gender, date of birth). Once an account is created and details of the children are uploaded, the user can download a set of QR codes for the children to be assessed. The user then downloads the app to an Android or Apple tablet or phone. To begin an assessment, the examiner scans the QR code to identify the child, and the assessment begins. The instructions make clear that children's responses

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

Item	Target	Phonological distractor	Semantic distractor	Unrelated distractor
1	balloon	baboon	airship	watch
2	telephone	xylophone	computer	bear
3	yo-yo	dodo	shuttlecock	heart
4	shell	shed	crab	trumpet
5	microphone	microscope	speaker	basket
6	sheep	shop	goat	slipper
7	drum	drawer	tambourine	car
8	thumb	plum	leg	purse
9	van	fan	truck	camera
10	barrel	bottle	chest	spoon
11	caravan	carrot	wagon	knot
12	suitcase	bookcase	trunk	helmet
13	mountain	fountain	volcano	shop
14	racoon	moon	fox	biscuit
15	jug	slug	vase	bin
16	chain	train	padlock	knife
17	sledge	hedge	skis	whistle
18	medal	model	trophy	singer
19	spanner	spatula	screwdriver	boat
20	scales	sail	clock	cup
21	arrow	wheelbarrow	dart	bowl
22	toad	toe	lizard	clown
23	flask	flag	kettle	ladybird

should be scored for accuracy discounting dialectical variation. An assessment can be paused if necessary and then restarted at the point of pausing by rescanning the child's QR code. The app stores no personally identifiable information about the child being tested.

The four subscales take roughly 10 min to administer, and data are automatically uploaded to a secure server where the child's test data are linked to their personal details (including the child's name, date of birth, and date of testing). A report of the scores for each child can then be downloaded from the user's account. The report provides lists of the children who have been assessed, ranked by overall language standard scores for each year group, along with instructions on how to interpret scores.

Table 2. Items in the Sentence Repetition subtest of LanguageScreen.

Item	Sentence structure
1	Birds fly.
2	Babies cry a lot.
3	Joe likes dogs.
4	I help mum.
5	We go to school on the bus.
6	My red scarf is nice and warm.
7	The field is full of flowers.
8	My grandad loves chocolate cake.
9	The ducks always swim to get the bread.
10	The teacher promised the boy a sticker.
11	Sally gave a birthday present to her friend.
12	Mummy baked the children an apple pie.
13	A boy gave the girl a ride on his bike.
14	Cats love to chase mice just for fun.

Participants

The LanguageScreen app was supplied to approximately 10,000 schools as part of a COVID-19 catch-up scheme in English preschools and primary schools. Screening was carried out by teachers and their assistants in these schools, who did not receive training to use the app. Schools were asked to test all children in each classroom that was screened. Data were available from 8,273 schools containing 348,944 children for the present analyses, indicating that schools tested approximately 42 children on average (where a typical class size is in the region of 25, but many schools had only one class per year group). All pupils up to the age of 9 years were eligible for testing. In practice, most children assessed were in the first year of formal schooling (referred to as *reception* in England, with pupils entering reception at the age of 4.5 years). Of the sample, 168,931 were identified as female and 178,907 were identified as male (a total of

1,106 were identified as either "unknown" or "other" in terms of gender).

Statistical Model and Analysis Plan

The study was designed to be analyzed using the Rasch model to determine item characteristics and was conceived within a theoretical framework that views language as a unitary (latent) trait (e.g., Tomblin & Zhang, 2006). LanguageScreen consists of 77 dichotomously scored items from the four subscales. The items in each subscale are presented in order of difficulty (easiest to hardest) as determined by earlier pilot phases where the items were administered to smaller samples and evaluated using the Rasch model for fit and refined or omitted where appropriate. Below, we present data concerning the psychometric properties of LanguageScreen based on a very large standardization sample.

Descriptive statistics were computed using Stata 17.0 (StataCorp, 2021). Classical reliability statistics were estimated using the lavaan and semTools packages in R and RStudio (Jorgensen et al., 2022; R Core Team, 2022; Rosseel, 2012; RStudio Team, 2020). The Rasch model analysis was conducted using the mirt package in R (Chalmers, 2012), and the differential item functioning (DIF) analyses were conducted using base R functions and the DescTools package (Signorell et al., 2019).

Cronbach's alpha (α) coefficient and McDonald's omega hierarchical (ω_h) coefficient were used to evaluate total score reliability, and the latter was calculated according to the procedure suggested by Flora (2020). First, a confirmatory bifactor analysis was applied where all items loaded on a general factor as well as a specific factor for their respective subscale. This confirmatory model showed excellent fit to the test data (Comparative Fit Index = .95, Tucker–Lewis index = .94, root-mean-square error of approximation [RMSEA] = .02) and was used to calculate the omega hierarchical coefficient according to Green and Yang's (2009) formulation. Unlike the alpha coefficient, the omega hierarchical coefficient provides a reliability estimate for the variance accounted for by just the general factor and thus provides evidence of the degree of unidimensionality across the items (Revelle & Zinbarg, 2009).

The item response data were analyzed using the Rasch model to evaluate the reliability and sufficiency of the total test score; the item difficulties and their fit to the model; and the invariance of the assessment across age, gender, and EAL status (Andrich, 2005). The Rasch model was chosen because LanguageScreen was developed to provide a total score that gives a reliable measure of a unidimensional language construct. LanguageScreen was developed, piloted, and refined in accordance with Rasch measurement theory to establish the sufficiency, reliability, and validity of this total score (Andrich, 2018). Reliability was evaluated in terms of the person separation reliability (PSR) statistic, which is analogous to the alpha coefficient, and is an estimate of the ratio of true variance to observed variance. Overall model fit was evaluated in terms of the RMSEA (cutoff value of .06) and standardized root-mean-square residual (SRMSR; cutoff value of .08) values, and item fit was evaluated using the infit mean-square residual statistic, with critical values of less than 0.8 and greater than 1.2, as well as by graphical inspection of the item characteristic curves.

The invariance of the assessment was evaluated in terms of DIF by age, gender, and EAL status using a logistic regression approach. For the latter two variables, only those who identified as male and female and those who identified as EAL or non-EAL were included in the analysis. This approach to estimating DIF has been broadly applied and was chosen here as it allows for both continuous (age in months) and categorical (gender, EAL status) predictors, and it enables the investigation of both uniform DIF, which indicates differences in the items' difficulty across groups, and nonuniform DIF, which indicates differences in the items' discrimination across groups (Swaminathan & Rogers, 1990). This approach involves estimating three logistic regression models for each item: (a) a base model that only includes the ability estimate as a predictor, which, in this case, was the ability estimate from the Rasch analysis; (b) a model that includes both the ability estimate and the group factor as predictors, which is used to evaluate uniform DIF through comparison with the base model; and (c) a model that includes the ability estimate, the group factor, and their interaction as predictors, which is used to evaluate nonuniform DIF through comparison to the second model. Given the extremely large sample size, trivial differences in item difficulties between the groups will be statistically significant. Thus, the magnitude of each item's uniform and nonuniform DIF was evaluated in terms of differences in Nagelkerke's (1991) pseudo-R^2 effect size measure across the models, and these pseudo-R^2 differences were further categorized using Jodoin and Gierl's (2001) recommendations into three categories: A = negligible, B = moderate, and C = large.

Results

Descriptive Statistics

Table 3 shows the means (and standard deviations) for scores on each subscale of the LanguageScreen test as a function of age, with the sample divided into 6-month age bands, spanning ages 3;06–3;11 (42–47 months) to

Table 3. Means and standard deviations of subtest scores by 6-month age band (including sample size in each age band).

Age (months)	*n*	Expressive Vocabulary		Receptive Vocabulary		Sentence Repetition		Listening Comprehension		Language total score	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
42–47	1,965	9.68	4.92	13.50	4.36	5.47	3.81	4.32	3.86	32.96	14.48
48–53	55,306	11.69	4.51	15.63	3.84	7.29	3.72	6.55	4.02	41.16	13.57
54–59	157,642	12.66	4.56	16.47	3.74	8.14	3.67	7.72	4.06	45.00	13.53
60–65	115,400	13.72	4.57	17.33	3.62	9.04	3.53	8.88	3.94	48.96	13.22
66–71	11,720	13.95	4.73	17.57	3.73	9.12	3.54	9.35	3.98	49.99	13.55
72–77	3,403	14.57	4.63	18.08	3.63	9.30	3.49	9.95	3.72	51.89	12.95
78–83	1,173	14.63	4.95	18.09	3.97	9.09	3.61	9.69	3.92	51.50	14.16
84–89	567	15.51	5.00	18.78	3.71	9.42	3.57	10.20	3.74	53.92	13.72
90–95	382	16.15	4.66	19.57	3.38	10.25	3.57	10.73	3.80	56.70	13.29
96–101	759	17.36	4.59	20.17	3.24	11.39	3.07	12.04	3.23	60.97	11.92
102–107	626	17.75	4.24	20.41	3.05	11.64	2.74	12.24	3.11	62.04	11.00

8;06–8;11 (102–107 months). Sample sizes are markedly uneven across the different age groups, with much larger samples in some of the younger age groups.

Figure 1 shows a violin plot of total raw score as a function of age group. A gradual increase in raw scores with age can be seen, with particularly steep increases across the bottom four 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.

Figure 2 shows LanguageScreen total raw scores as a function of age group and gender, along with 95% confidence intervals (CIs; please note that some of the CIs are very small due to the large sample size and so are not visible on the graph). Overall, as expected, there is a small but highly significant advantage for girls compared to boys, d = 0.14, 95% CI [0.132, 0.146], F(347, 815) = 89.06, p = .0001, which does not vary as a function of age, F(347, 815) = 1.40, p = .17.

Figure 3 shows LanguageScreen total raw scores as a function of age group and EAL status, along with 95% CIs (please note that, again, some of the CIs are not visible on the graph due to the large sample size). Overall, children who identified as EAL have much lower scores, d = 1.03, 95% CI [−1.024, −1.042], F(316, 938) = 2,170.71, p = .0001; this effect is smaller in older age groups, F(316, 918) = 17.36, p = .0001. Although this might be expected given that older children will, on

Figure 1. Violin plot showing the distribution of LanguageScreen total scores as a function of age group.

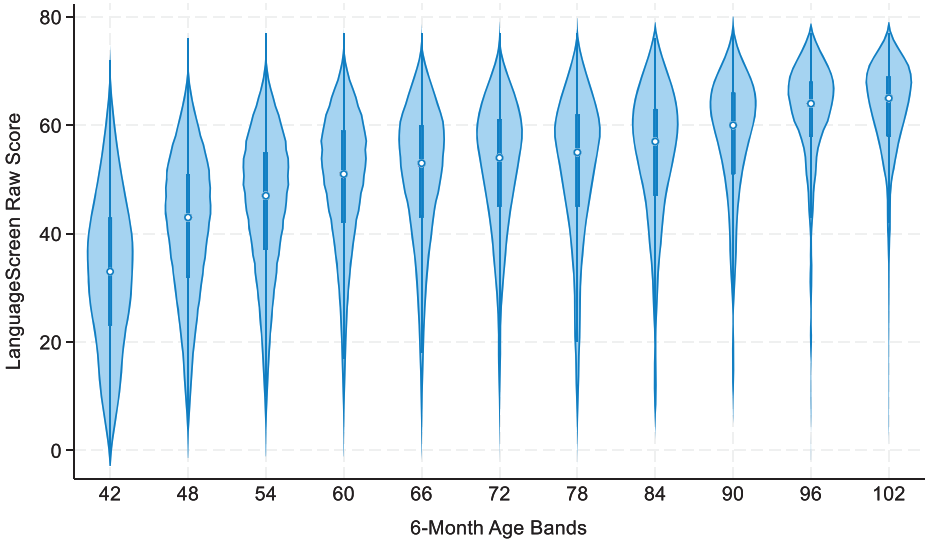
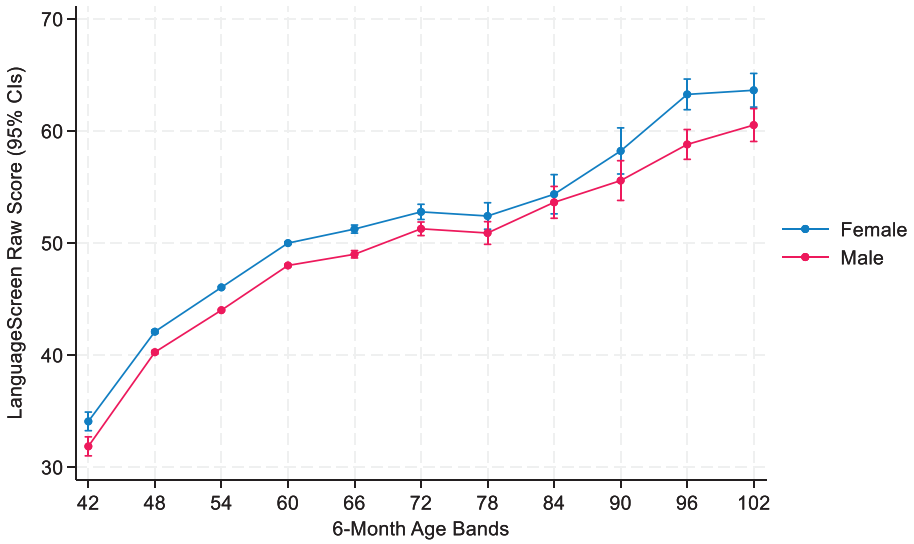


Figure 2. Mean LanguageScreen total scores as a function of age group and gender with 95% confidence intervals (CIs).

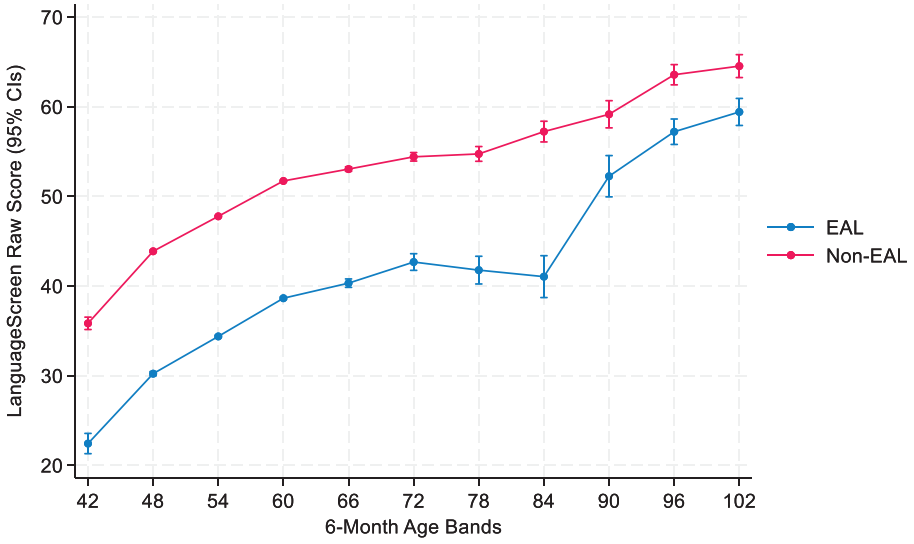


average, have been in English classrooms longer, caution is required in interpreting the finding, given the data are cross-sectional rather than longitudinal.

Classical Reliability Analyses

LanguageScreen showed good-to-excellent reliability for the total scale (α = .92, ω_h = .75). The omega hierarchical value indicates that 75% of the variance of the composite score is attributable to variance on the bifactor model's general factor. The finding that a single general factor accounts for such a high proportion of the variance in the composite score supports the application of the unidimensional Rasch model to the data. Classical reliabilities (alphas) for the separate subscales were adequate to good (EV = .79, RV = .74, SR = .79, LC = .80). The test–retest reliability of the LanguageScreen total score was evaluated using a sample of children in reception class in the United Kingdom (i.e., the first year of compulsory education). Test–retest reliability was good (r = .78, N = 9,778). The average lag between the first and second assessments here was 6.8 months, making this figure arguably a conservative measure of test–retest reliability.

Figure 3. Mean LanguageScreen total scores as a function of age group and English as an additional language (EAL) status with 95% confidence intervals (CIs).



Rasch Analyses

The data showed a reasonable-to-good overall fit to the Rasch model (SRMSR = .09, RMSEA = .03) and further confirmed the excellent reliability of the LanguageScreen total score (PSR = .94). Figure 4 shows a Wright map giving the difficulty estimates for each of the items across the four subscales expressed in logit units. Lower values represent easier items. As can be seen, each subscale had a broad range of difficulty estimates for the different items, reflecting our intention when constructing the test. LC was the most difficult subscale (M = 0.01, SD = 0.75, PSR = .82), followed by EV (M = −0.40, SD = 1.91, PSR = .87) and SR (M = −0.66, SD = 1.38, PSR = .85) subscales. RV was the easiest subscale (M = −1.45, SD = 1.35, PSR = .79).

Individual Item Fit Estimates

Most items had a good fit to the Rasch model, with only five items (rv3, rv13, rv7, rv6, and rv9 in order of magnitude of misfit) displaying discrimination substantially below the average discrimination of all items (which means they did not discriminate between different language ability levels as much as model expectation) and 11 items (ev4, ev1, sr3, ev2, ev6, ev3, sr2, sr5, sr6, sr7, and sr9) showing discrimination substantially above the average discrimination of all items (which means they showed greater discrimination between ability levels than model expectation).

DIF

DIF (or item bias; Lord, 1980) refers to possible differences between groups in the proportion of individuals of a given level of ability who answer a given item correctly. Items that give different success rates for two or more groups at the same ability level are said to display DIF (Holland & Wainer, 1993). DIF is undesirable within the Rasch framework as it would indicate that some items are biased for some groups of individuals; we therefore conducted analyses to assess the extent to which items in LanguageScreen displayed DIF. Table 4 shows the frequency of different effect sizes for DIF as a function of age, gender, and EAL status. These measures indicate the extent to which different items show differential difficulty for these grouping variables. For age, no items showed more than a negligible effect for both uniform (constant across different ability levels) and nonuniform (varying as a function of ability) DIF. For gender, six items showed moderate uniform-DIF effects, of which four (ev7, ev9, ev18, and ev19) were significantly easier for boys and two (ev17 and ev22) were significantly easier for girls, and one item (ev24) showed a large uniform-DIF effect, indicating it was substantially easier for the boys. No items were found to have more than a negligible nonuniform-DIF

Figure 4. Wright map showing the distribution of ability estimates (left panel) relative to the item difficulty estimates across the four subscales (four right panels) on the Rasch model logit scale. EV = Expressive Vocabulary; RV = Receptive Vocabulary; SR = Sentence Repetition; LC = Listening Comprehension.

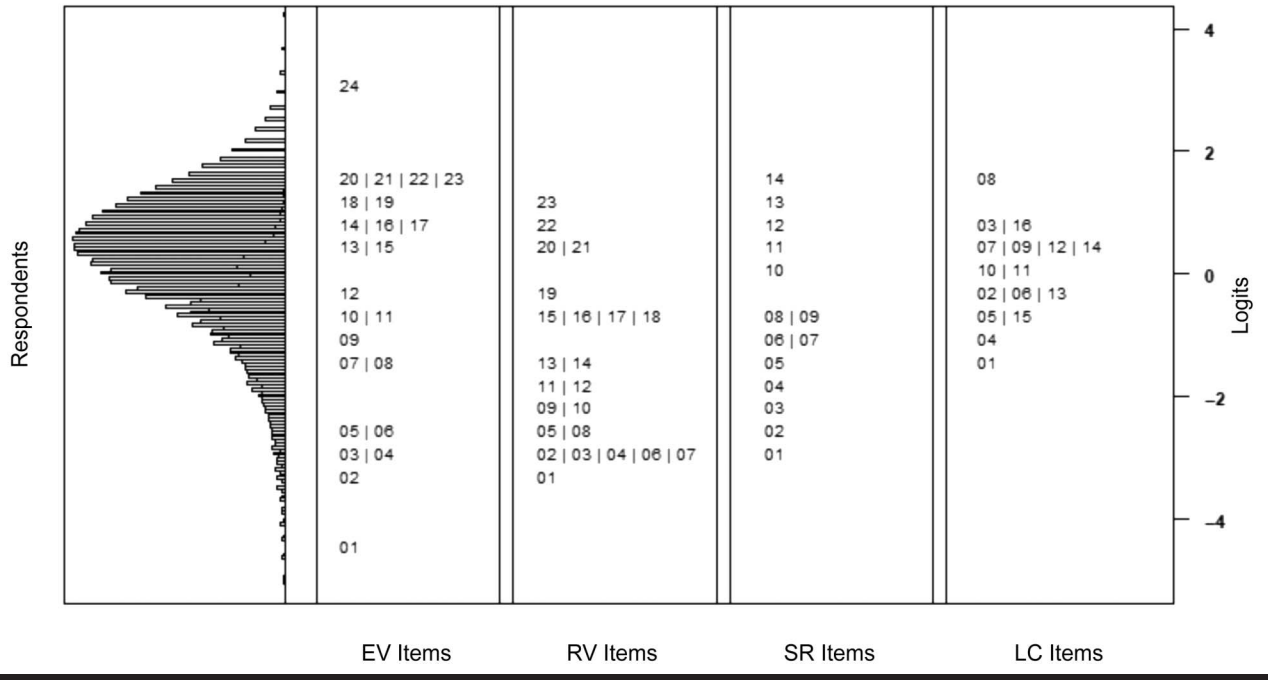


Table 4. Frequency and percentage of items found to have negligible (A), moderate (B), and large (C) differential item functioning for the age, gender, and English as an additional language (EAL) grouping variables.

	A		B		C	
Grouping	***n***	**%**	***n***	**%**	***n***	**%**
Age	77	100.0	0	0.0	0	0.0
Gender	70	90.9	6	7.8	1	1.3
EAL	75	97.4	2	2.6	0	0.0

effect for gender. For EAL status, two items (ev4 and rv11) showed moderate uniform-DIF effects, with both being significantly easier for non-EAL respondents, and no items were found to have more than a negligible nonuniform-DIF effect. The finding that a minimum of 90% of items show negligible uniform-DIF effects and no items were found to have more than negligible nonuniform-DIF effects for each of the grouping variables implies that the effect of DIF on the total score can be expected to be minimal and provides further support for the claim that the total score on LanguageScreen gives an accurate indication of a participant's ability level irrespective of their background characteristics.

Discussion

This study was conducted to assess the psychometric properties of LanguageScreen, a test designed to be used by education professionals to assess children's language ability. We have presented data on the reliability of the test when administered by school staff who assessed children without any special training. Using Rasch scaling, we have established that the test's total score is a sufficient statistic that gives a reliable estimate of a child's language ability (Andrich, 2005). The test is quick to administer (approximately 10 min).

Research Findings

The development of LanguageScreen was guided by the theoretical assumption that language skills, to a first approximation, are well described as a unitary trait, and this led us to use the Rasch model to guide item selection and analyze the data. The Rasch model provided a good fit to data from the four subscales (EV, RV, SR, and LC), indicating that a single score can be used to measure overall language ability. Because the data fitted the Rasch model well, the total score is a sufficient statistic, meaning that it gives all the information needed to infer a person's ability. We believe that the LanguageScreen total score, which we have shown here to be highly reliable, provides a useful starting point for characterizing a child's language level and for monitoring the growth in language skills following intervention (cf. West et al., 2021). In addition, we have shown using data from LanguageScreen that, as expected, there are small but significant differences in language ability between boys and girls and that children who speak English as a second or additional language are less proficient in English than those from monolingual English-speaking homes.

Validation

The data analyzed here included pretest data from a large randomized controlled trial (West et al., 2021) that evaluated the effects of a language intervention. LanguageScreen was used in that study to select children who would be considered suitable to receive language intervention. In that study, 5,719 children in Reception classes in 193 schools were assessed using LanguageScreen by school staff. From that sample, 1,156 children who were the five lowest scoring children in each classroom (20% of all children assessed) were then 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 score from the speech and language therapist language assessment scores correlated highly with the LanguageScreen total scores (r = .74). It should be noted that this correlation is subject to restriction of range (since only the bottom 20% of the sample in terms of LanguageScreen scores were reassessed with the individually administered tests). In addition, in this study, LanguageScreen was readministered to all children after the intervention was completed. LanguageScreen showed comparable gains in scores in the children who had received intervention to the gains shown on the tests individually administered by professionals. These findings provide strong support for the validity of LanguageScreen. The test correlates well with much longer, well-standardized tests of language ability and is sensitive to improvements in language skills brought about by intervention.

Clinical and Educational Implications

The main aim of this research was to provide teachers and other education professionals with a quick

and reliable method of assessing children's language ability. LanguageScreen provides an objective measure of language ability in a test that takes approximately 10 min. The data from LanguageScreen are automatically uploaded to a secure website, which generates a report detailing the scores of each child (see OxEd and Assessment Ltd, 2022).

The report contains a list of children ranked according to their language scores by year group. The reports use a "traffic lights" system to flag children whose language skills are a cause for concern: Green (a standard score of 90 or above) indicates no concerns, amber (a standard score between 82 and 89) indicates possible concerns and that a child may benefit from additional language support, and red (a standard score of 81 or below) indicates clear concerns and that a child definitely requires support in developing their language skills.

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 behavioral difficulties as a check on possible underlying causes.

Limitations

The present study was guided by the theoretical assumption that performance across the four subtests in LanguageScreen reflects a single unitary trait. We used the Rasch model to assess the scale, since if this model holds, the total score from the test is a sufficient statistic.

Evidence reviewed earlier suggests that language skills in young children are well captured by a single latent variable, and the findings reported here from LanguageScreen are consistent with those earlier findings. We would not, however, want to make strong claims about the dimensionality of language from these findings, since our development of LanguageScreen was guided by a unidimensional model, and items were selected to fit that model.

The present study was conducted in English schools, with assessments carried out by untrained school staff (typically teaching assistants [teacher aides]) or teachers. It is clear that school staff can use the screening tool effectively as validated by the high test–retest reliability of the assessments and the correlations reported with individual standardized tests given by professionally trained therapists. Moreover, although we believe the findings to be robust and readily generalizable to other English-speaking communities, it is acknowledged that any such adaptation would need to take account of local linguistic and cultural factors to ensure contextual appropriacy and to avoid the misidentification of the needs of individual children (e.g., Bialystok et al., 2010; Goodrich et al., 2023).

It should be noted that we chose not to report evidence on the specificity/sensitivity of LanguageScreen as a diagnostic test of DLD. This is because we see language disorders, such as reading disorders, as dimensional (Snowling & Hulme, 2021, 2024). In this light, it makes little sense to categorize children into binary categories (normal vs. impaired) and predict such a binary outcome from a continuous variable such as LanguageScreen. It should be noted, however, that a major determinant of the specificity and sensitivity of a test is its reliability (Edwards et al., 2022), which, in the case of LanguageScreen, is high ($\alpha = .92$).

Summary and Conclusions

We have reported data from a very large sample of children assessed with LanguageScreen, a new app-based language assessment. LanguageScreen has excellent psychometric properties and provides education professionals with a quick and reliable assessment of children's language skills. From a practical perspective, these findings provide educators and researchers with a very useful tool.

Data Availability Statement

Due to privacy and ethical concerns, neither the data nor the source of the data can be made available.

Acknowledgments

This study was supported by funding from the Heather van der Lely Foundation, awarded to Charles Hulme, Gillian West, and Maggie Snowling, for the first standardization of LanguageScreen in 2018. The authors would like to acknowledge all of the schools and education professionals who have used LanguageScreen, as well as all the children

who have taken part in screening. They would also like to thank Enxhi Sharxhi for administrative assistance.

References

Adlof, S. M., & Hogan, T. P. (2019). If we don't look, we won't see: Measuring language development to inform literacy instruction. *Policy Insights from the Behavioral and Brain Sciences, 6*(2), 210–217. https://doi.org/10.1177/2372732219839075

Andrich, D. (2005). The Rasch model explained. In S. Alagumalai, D. D. Durtis, & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 308–328). Springer-Kluwer. https://doi.org/10.1007/1-4020-3076-2_3

Andrich, D. (2018). Advances in social measurement: A Rasch measurement theory. In F. Guillemin, A. Leplège, S. Briançon, E. Spitz, & J. Coste (Eds.), *Perceived health and adaptation in chronic disease* (pp. 66–91). Routledge.

Baker, S., Harding, S. A., Holme, C., Lewis, R., Seifert, M., & Wren, Y. (2022). *Review of early language screening suitable for children in Wales from birth to 5 years*. Welsh Government. https://www.gov.wales/review-early-language-screening-suitable-children-wales-birth-5-years-summary-html#section-102498

Berkman, N. D., Wallace, I., Watson, L., Coyne-Beasley, T., Cullen, K., Wood, C., & Lohr, K. N. (2015). *Screening for speech and language delays and disorders in children age 5 years or younger: A systematic review for the U.S. Preventive Services Task Force*. Agency for Healthcare Research and Quality. https://www.ncbi.nlm.nih.gov/books/NBK305674/

Bialystok, E., Luk, G., Peets, K. F., & Yang, S. (2010). Receptive vocabulary differences in monolingual and bilingual children. *Bilingualism: Language and Cognition, 13*(4), 525–531. https://doi.org/10.1017/S1366728909990423

Bishop, D. V. M., & Hayiou-Thomas, M. E. (2008). Heritability of specific language impairment depends on diagnostic criteria. *Genes, Brain and Behavior, 7*(3), 365–372. https://doi.org/10.1111/j.1601-183X.2007.00360.x

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & CATALISE Consortium. (2016). CATALISE: A multinational and multidisciplinary Delphi consensus study. Identifying language impairments in children. *PLOS ONE, 11*(7), Article e0158753. https://doi.org/10.1371/journal.pone.0158753

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chow, J. C., & Ekholm, E. (2019). Language domains differentially predict mathematics performance in young children. *Early Childhood Research Quarterly, 46,* 179–186. https://doi.org/10.1016/j.ecresq.2018.02.011

Chow, J. C., Wallace, E. S., Senter, R., Kumm, S., & Mason, C. Q. (2022). A systematic review and meta-analysis of the language skills of youth offenders. *Journal of Speech, Language, and Hearing Research, 65*(3), 1166–1182. https://doi.org/10.1044/2021_JSLHR-20-00308

Clegg, J., Hollis, C., Mawhood, L., & Rutter, M. (2005). Developmental language disorders—A follow-up in later adult life. Cognitive, language and psychosocial outcomes. *The Journal of Child Psychology and Psychiatry, 46*(2), 128–149. https://doi.org/10.1111/j.1469-7610.2004.00342.x

Conti-Ramsden, G., & Botting, N. (2004). Social difficulties and victimization in children with SLI at 11 years of age. *Journal of Speech, Language, and Hearing Research, 47*(1), 145–161. https://doi.org/10.1044/1092-4388(2004/013)

Conti-Ramsden, G., Durkin, K., Toseeb, U., Botting, N., & Pickles, A. (2018). Education and employment outcomes of young adults with a history of developmental language disorder. *International Journal of Language & Communication Disorders, 53*(2), 237–255. https://doi.org/10.1111/1460-6984.12338

Duff, F. J., & Clarke, P. J. (2011). Practitioner review: Reading disorders: What are the effective interventions and how should they be implemented and evaluated? *The Journal of Child Psychology and Psychiatry, 52*(1), 3–12. https://doi.org/10.1111/j.1469-7610.2010.02310.x

Edwards, A. A., van Dijk, W., White, C. M., & Schatschneider, C. (2022). Screening screeners: Calculating classification indices using correlations and cut-points. *Annals of Dyslexia, 72*(3), 445–460. https://doi.org/10.1007/s11881-022-00261-5

Ellis, A. W., & Morrison, C. M. (1998). Real age-of-acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(2), 515–523. https://doi.org/10.1037//0278-7393.24.2.515

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science, 3*(4), 484–501. https://doi.org/10.1177/2515245920951747

Fricke, S., Bowyer-Crane, C., Haley, A. J., Hulme, C., & Snowling, M. J. (2013). Efficacy of language intervention in the early years. *The Journal of Child Psychology and Psychiatry, 54*(3), 280–290. https://doi.org/10.1111/jcpp.12010

Fuchs, D., Fuchs, L. S., & Compton, D. L. (2012). Smart RTI: A next-generation approach to multilevel prevention. *Exceptional Children, 78*(3), 263–279. https://doi.org/10.1177/001440291207800301

Goodrich, J. M., Fitton, L., Chan, J., & Davis, C. J. (2023). Assessing oral language when screening multilingual children for learning disabilities in reading. *Intervention in School and Clinic, 58*(3), 164–172. https://doi.org/10.1177/10534512221081264

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*(1), 155–167. https://doi.org/10.1007/S11336-008-9099-3

Hendricks, A. E., Adlof, S. M., Alonzo, C. N., Fox, A. B., & Hogan, T. P. (2019). Identifying children at risk for developmental language disorder using a brief, whole-classroom screen. *Journal of Speech, Language, and Hearing Research, 62*(4), 896–908. https://doi.org/10.1044/2018_JSLHR-L-18-0093

Hjetland, H. N., Brinchmann, E. I., Scherer, R., Hulme, C., & Melby-Lervåg, M. (2020). Preschool pathways to reading comprehension: A systematic meta-analytic review. *Educational Research Review, 30,* Article 100323. https://doi.org/10.1016/j.edurev.2020.100323

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Lawrence Erlbaum.

Holzinger, D., Weber, C., & Diendorfer, B. (2022). Development and validation of a language screening for implementation in pre-school settings. *Frontiers in Public Health, 10,* Article 866598. https://doi.org/10.3389/fpubh.2022.866598

Hornburg, C. B., Schmitt, S. A., & Purpura, D. J. (2018). Relations between preschoolers' mathematical language understanding and specific numeracy skills. *Journal of Experimental Child Psychology, 176,* 84–100. https://doi.org/10.1016/j.jecp.2018.07.005

Hulme, C., Nash, H. M., Gooch, D., Lervåg, A., & Snowling, M. J. (2015). The foundations of literacy development in children at familial risk of dyslexia. *Psychological Science, 26*(12), 1877–1886. https://doi.org/10.1177/0956797615603702

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329–349. https://doi.org/10.1207/S15324818AME1404_2

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). *semTools: Useful tools for structural equation modeling* (R Package Version 0.5-6). https://CRAN.R-project.org/package=semTools

Kaiser, A. P., Chow, J. C., & Cunningham, J. E. (2022). A case for early language and behavior screening: Implications for policy and child development. *Policy Insights from the Behavioral and Brain Sciences, 9*(1), 120–128. https://doi.org/10.1177/23727322211068886

Klem, M., Melby-Lervåg, M., Hagtvet, B., Lyster, S.-A. H., Gustafsson, J.-E., & Hulme, C. (2015). Sentence repetition is a measure of children's language skills rather than working memory limitations. *Developmental Science, 18*(1), 146–154. https://doi.org/10.1111/desc.12202

Korpilahti, P., Kaljonen, A., & Jansson-Verkasalo, E. (2016). Population-based screening for language delay: Let's talk STEPS study. *Psychology, 07*(02), 205–214. https://doi.org/10.4236/psych.2016.72023

Language and Reading Research Consortium. (2015). The dimensionality of language ability in young children. *Child Development, 86*(6), 1948–1965. https://doi.org/10.1111/cdev.12450

Law, J., Boyle, J., Harris, F., Harkness, A., & Nye, C. (1998). Screening for speech and language delay: A systematic review of the literature. *Health Technology Assessment, 2*(9). https://doi.org/10.3310/hta2090

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.

McGregor, K. K. (2020). How we fail children with developmental language disorder. *Language, Speech, and Hearing Services in Schools, 51*(4), 981–992. https://doi.org/10.1044/2020_LSHSS-20-00003

Morrison, C. M., & Ellis, A. W. (2000). Real age of acquisition effects in word naming and lexical decision. *British Journal of Psychology, 91*(2), 167–180. https://doi.org/10.1348/000712600161763

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691–692. https://doi.org/10.1093/biomet/78.3.691

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., & Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: Evidence from a population study. *The Journal of Child Psychology and Psychiatry, 57*(11), 1247–1257. https://doi.org/10.1111/jcpp.12573

Norbury, C. F., Nash, M., Baird, G., & Bishop, D. V. M. (2004). Using a parental checklist to identify diagnostic groups in children with communication impairment: A validation of the Children's Communication Checklist–2. *International Journal of Language & Communication Disorders, 39*(3), 345–364. https://doi.org/10.1080/13682820410001654883

OxEd and Assessment Ltd. (2022). *LanguageScreen*. https://oxedandassessment.com/languagescreen/

Paul, R. (2007). *Language disorders from infancy through adolescence: Assessment & intervention* (Vol. 324). Elsevier.

Petersen, D. B., & Spencer, T. D. (2012). The Narrative Language Measures: Tools for language screening, progress monitoring, and intervention planning. *Perspectives on Language Learning and Education, 19*(4), 119–129. https://doi.org/10.1044/lle19.4.119

Petersen, I. T., Bates, J. E., D'Onofrio, B. M., Coyne, C. A., Lansford, J. E., Dodge, K. A., Pettit, G. S., & Van Hulle, C. A. (2013). Language ability predicts the development of behavior problems in children. *Journal of Abnormal Psychology, 122*(2), 542–557. https://doi.org/10.1037/a0031963

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reilly, S., Wake, M., Ukoumunne, O. C., Bavin, E., Prior, M., Cini, E., Conway, L., Eadie, P., & Bretherton, L. (2010). Predicting language outcomes at 4 years of age: Findings from Early Language in Victoria Study. *Pediatrics, 126*(6), e1530–e1537. https://doi.org/10.1542/peds.2010-0254

Renfrew, C. (2003). *Action Picture Test*. Speechmark Publishing.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*(1), 145–154. https://doi.org/10.1007/s11336-008-9102-z

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

RStudio Team. (2020). *RStudio: Integrated development for R*. https://www.rstudio.com/

Seager, E., & Abbot-Smith, K. (2017). Can early years professionals determine which preschoolers have comprehension delays? A comparison of two screening tools. *Child Language Teaching and Therapy, 33*(1), 67–79. https://doi.org/10.1177/0265659016650977

Semel, E., Wiig, E., & Secord, W. (2006). *Clinical Evaluation of Language Fundamentals Preschool–Second Edition UK*. Pearson Assessment.

Signorell, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T., Arppe, A., Baddeley, A., Barton, K., Bolker, B., & Borchers, H. W. (2019). *DescTools: Tools for descriptive statistics* (R Package Version 0.99).

Snowling, M., & Hulme, C. (2024). Do we really need a new definition of dyslexia? A commentary. *Annals of Dyslexia, 74*, 124–128. https://doi.org/10.1007/s11881-024-00305-y

Snowling, M., van Wagtendonk, B., & Stafford, C. (1988). Object-naming deficits in developmental dyslexia. *Journal of Research in Reading, 11*(2), 67–85. https://doi.org/10.1111/j.1467-9817.1988.tb00152.x

Snowling, M. J., & Hulme, C. (2021). Annual Research Review: Reading disorders revisited–the critical importance of oral language. *Journal of Child Psychology and Psychiatry, 62*(5), 635–653. https://doi.org/10.1111/jcpp.13324

Snowling, M. J., Nash, H. M., Gooch, D. C., Hayiou-Thomas, M. E., Hulme, C., & Wellcome Language and Reading Project Team. (2019). Developmental outcomes for children at high risk of dyslexia and children with developmental language disorder. *Child Development, 90*(5), e548–e564. https://doi.org/10.1111/cdev.13216

So, K. K. H., & To, C. K. S. (2022). Systematic review and meta-analysis of screening tools for language disorder. *Frontiers in Pediatrics, 10,* Article 801220. https://doi.org/10.3389/fped.2022.801220

StataCorp. (2021). Stata statistical software: Release 17.

Stothard, S. E., Snowling, M. J., Bishop, D. V. M., Chipchase, B. B., & Kaplan, C. A. (1998). Language-impaired preschoolers: A follow-up into adolescence. *Journal of Speech, Language, and Hearing Research, 41*(2), 407–418. https://doi.org/10.1044/jslhr.4102.407

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Tannock, R., & Schachar, R. (1996). Executive dysfunction as an underlying mechanism of behavior and language problems in attention deficit hyperactivity disorder. In J. H. Beitchman, N. J. Cohen, M. M. Konstantareas, & R. Tannock (Eds.),

Language, learning, and behavior disorders: Developmental, biological, and clinical perspectives (Vol. 38, pp. 128–155). Cambridge University Press.

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research, 40*(6), 1245–1260. https://doi.org/10.1044/jslhr.4006.1245

Tomblin, J. B., & Zhang, X. (2006). The dimensionality of language ability in school-age children. *Journal of Speech, Language, and Hearing Research, 49*(6), 1193–1208. https://doi.org/10.1044/1092-4388(2006/086)

Van Agt, H., Verhoeven, L., Van Den Brink, G., & De Koning, H. (2011). The impact on socio-emotional development and quality of life of language impairment in 8-year-old children. *Developmental Medicine & Child Neurology, 53*(1), 81–88. https://doi.org/10.1111/j.1469-8749.2010.03794.x

Wallace, I. F., Berkman, N. D., Watson, L. R., Coyne-Beasley, T., Wood, C. T., Cullen, K., & Lohr, K. N. (2015). Screening for speech and language delay in children 5 years old and younger: A systematic review. *Pediatrics, 136*(2), e448–e462. https://doi.org/10.1542/peds.2014-3889

Wallentin, M. (2020). Gender differences in language are small but matter for disorders. *Handbook of Clinical Neurology, 175,* 81–102. https://doi.org/10.1016/B978-0-444-64123-6.00007-2

West, G., Snowling, M. J., Lervåg, A., Buchanan-Worster, E., Duta, M., Hall, A., McLachlan, H., & Hulme, C. (2021). Early language screening and intervention can be delivered successfully at scale: Evidence from a cluster randomized controlled trial. *The Journal of Child Psychology and Psychiatry, 62*(12), 1425–1434. https://doi.org/10.1111/jcpp.13415

Whiteside, K. E., & Norbury, C. F. (2017). The persistence and functional impact of English language difficulties experienced by children learning English as an additional language and monolingual peers. *Journal of Speech, Language, and Hearing Research, 60*(7), 2014–2030. https://doi.org/10.1044/2017_JSLHR-L-16-0318

Wiig, E. H., & Secord, W. A. (1994). Classroom and communication checklist. In W. A. Secord, E. H. Wiig, S. H. Coolahan, & D. H. Pallante (Eds.), *Team-based problem solving* (pp. 37–39). Riverside Publishing.

Winstanley, M., Durkin, K., Webb, R. T., & Conti-Ramsden, G. (2018). Financial capability and functional financial literacy in young adults with developmental language disorder. *Autism & Developmental Language Impairments, 3*. https://doi.org/10.1177/2396941518794500

Zhang, X., & Tomblin, J. B. (2000). The association of intervention receipt with speech-language profiles and social-demographic variables. *American Journal of Speech-Language Pathology, 9*(4), 345–357. https://doi.org/10.1044/1058-0360.0904.345