Research report

Title: The development and validation of the Short Language Measure (SLaM): a brief measure of general language ability for children in their first year at school

Jessica Matov§#, Fiona Mensah#†, Fallon Cook†, Sheena Reilly ¶ and Richard Dowell§ α

§ Department of Audiology and Speech Pathology, the University of Melbourne, Carlton, VIC, Australia
# Department of Paediatrics, the University of Melbourne, Parkville, VIC, Australia
† Murdoch Children's Research Institute, Royal Children's Hospital, Parkville, VIC, Australia
¶ Menzies Health Institute Queensland, Griffith University, QLD, Australia
α Royal Victorian Eye and Ear Hospital, East Melbourne, VIC, Australia

Author details:
Ms Jessica Matov
Department of Audiology and Speech Pathology
The University of Melbourne, Carlton, VIC, Australia
VIC 3053
Tel: +61 421 643 496 Email: jessica.matov@mcri.edu.au

Dr Fiona Mensah
Clinical Epidemiology & Biostatistics (CEBU)
Murdoch Childrens Research Institute, The Royal Children's Hospital, Flemington Rd, Parkville, VIC 3052
Tel: + 61 3 9345 4741 Email: fiona.mensah@mcri.edu.au

Dr Fallon Cook
Healthy Mothers Healthy Families, Population Health
Murdoch Childrens Research Institute, The Royal Children's Hospital, Flemington Rd, Parkville, VIC 3052
Tel: + 61 3 9345 5484 Email: fallon.cook@mcri.edu.au

Prof Sheena Reilly
Menzies Health Institute Queensland, Griffith University, Parklands Drive, Southport, QLD 4222
Tel: + 61 7 5678 8664 Email: s.reilly@griffith.edu.au

Prof Richard Dowell
Department of Audiology and Speech Pathology
The University of Melbourne, Carlton, VIC, Australia

VIC 3053
Tel: +61 3 9035 5339 Email: rcd@unimelb.edu.au

Key words: assessment, language, language impairment, school-age children, screening

Acknowledgments

We thank participating families. We also thank Dr Colleen Holt for her linguistics input and review of SLaM's test items. Fiona Mensah and Sheena Reilly received Australian National Health and Medical Research Council Early Career and Career Development Fellowships (FM #1037449; #1111160), and Practitioner Fellowships (SR, #491210; #1041892). Jessica Matov was supported through an Australian Government Research Training Program Scholarship. Research at the Murdoch Children's Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program.

Abstract

Background: There is no sufficiently accurate short language measure that could be used by speech-language pathologists, teachers or paraprofessionals to screen young school-aged children to identify those requiring in-depth language evaluations. This may be due to poor development of available measures, which have omitted crucial test development steps. Applying more stringent development procedures could result in a measure with sufficient accuracy.
Aims: To create and validate a short language measure that has acceptable accuracy, validity and reliability, and can be used to identify children who require further assessment and/or referral to speech-language services.
Methods & Procedures: The study consisted of two phases. In Phase one (measure creation), 56 children were assessed with 160 direction-following and sentence-recall test items and a reference measure, the Clinical Evaluation of Language Fundamentals-fourth edition (CELF-4). Items were then examined for their individual characteristics (validity, reliability, difficulty and discrimination) via item analysis and the highest quality items were selected to form the Short Language Measure (SLaM). In Phase two (measure validation), 126 children were assessed with the SLaM and the reference measure (CELF-4) to determine SLaM's accuracy, validity and reliability.
Outcomes & Results: Forty test items were selected to form SLaM in Phase one. Findings from Phase two indicated that SLaM had an accuracy of 94% (sensitivity = 94%, specificity = 93%), validity of 0.89 and reliability of 0.93. These values remained relatively consistent across both phases.
Conclusions & Implications: Results indicated that SLaM has excellent psychometric properties. SLaM can be used to identify children who need further evaluation by a speech-language pathologist.

What this paper adds?
What is already known on this subject?

Prior research suggests that combining a direction-following and a sentence-recall task has sufficient discrimination accuracy and agreement with an omnibus language measure. Trialling a large set of direction-following and sentence-recall test items to select those with the highest individual characteristics could result in an effective short language measure.

What this study adds?

A short language measure (SLaM) was created and validated on two independent samples of children. Items with the highest validities, reliabilities and discrimination capacities were selected to form SLaM. This procedure resulted in a measure with high validity and reliability, that exceeded the criterion for adequate discrimination accuracy.

Clinical implications of this study?

SLaM is an effective measure that can accurately identify children who require detailed evaluations by speech-language pathologists.

Introduction

Approximately 10% of children begin school without the necessary oral language skills required for learning (Norbury et al. 2016). Language skills are essential for educational attainment as they provide a foundation for literacy development and learning. Children experiencing language difficulties have a greater risk of low academic achievement and subsequently, have poorer vocational opportunities (Johnson, Beitchman & Brownlie 2010). Ensuring the adequate provision of services for children with language difficulties in their first year at school may prevent children from falling behind in their learning and literacy development.

The provision of services for children with language difficulties relies on accurate detection and referral to speech-language services. However, studies indicate high rates of misclassification and under-identification of language difficulties in school-aged children (Bishop & McDonald 2009, Broomfield & Dodd 2004, Curran et al. 2015, Poll, Betz & Miller 2010, Tomblin et al. 1997). Bishop and McDonald (2009) reported that over half of nine-year-olds identified with language difficulties, via the use of a range of measures, had not previously been referred to services despite experiencing literacy difficulties.

The high rates of inaccurate language referrals and under-identification of language difficulties in school-aged children could be explained by a gap in teacher knowledge. Teachers are the primary referrers of school-aged children to speech-language services (Curran et al. 2015). Yet, a large proportion (88%) rate their knowledge of language difficulties as 'limited' or 'very limited' (Sadler 2005) and 76% acknowledged the need for additional training and support to identify children with poor language skills (Mroz 2006). Hence, it may be speculated that the inaccurate detection of language difficulties may be mitigated by adequate teacher training. However, studies indicate that educating teachers about language difficulties and training them to identify children via communication checklists still results in unacceptably poor detection rates of children who score below expected on standardised language measures (Antoniazzi, Snow & Dickson-Swift 2010, Jessup et al. 2008).

A solution to the inaccurate detection of language difficulties, could be to screen the language abilities of children via the administration of a short language measure. A short language measure (requiring 10 minutes or less to deliver) may be administered by speech-language pathologists, teachers or paraprofessionals (trained staff), to indicate children requiring in-depth language evaluations. Such a measure could be used across schools to screen children starting their formal education to facilitate the timely identification and support of those with language difficulties. Screening children at the beginning of school would enable focused targeting of language difficulties within the school system, at the crucial time when oral language skills become essential for learning and literacy development. Yet, the feasibility and benefits of using a measure in this way would depend on its psychometric properties.

The psychometric properties of an assessment relate to a measure's reliability and validity (Kline 2013). Reliability refers to the extent to which an assessment gives consistent results. In this paper, reliability is considered to be the degree to which a measure's items produce similar scores (internal consistency). Validity refers to the extent to which an assessment measures what it intends to measure. As the intention of a short language measure is to assess language ability, validity is considered in terms of a measure's relationship to, and agreement with an omnibus language measure (concurrent validity).

Another important aspect of validity is a measure's accuracy (Betz, Eickhoff & Sullivan 2013). Accuracy refers to the percentage of children correctly classified by a measure as having language difficulties or normal-range language skills. The accuracy of a measure can be evaluated by its sensitivity and specificity rates. To be clinically useful, a short language measure must correctly identify children with language difficulties (sensitivity) and children with age-appropriate language skills (specificity). Any short measure must meet acceptable accuracy standards, as misclassifying children can lead to a lack of appropriate support or the misuse of resources. Plante and Vance (1994) recommended that language assessments should have an accuracy (sensitivity and specificity) of 90% or more. Hence, this criterion was adopted for the current study.

Various short language measures that could be used to assess young school-aged children have been developed (Dockrell & Marshall 2015, Law et al. 2000). However, use of these measures results in high misclassification rates as their accuracy, sensitivity and/or specificity values are below the acceptable standard. For example, the Grammar and Phonology Screening (GAPS) test was developed as a short language measure to be administered by speech-language pathologists or paraprofessionals (Gardner et al. 2006). A subsequent validation study indicated high specificity (96%) but substantially lower

sensitivity (20%) in comparison to diagnosis made using a standardised language measure (Nash, Leavett & Childs 2011).

The insufficient accuracy of existing measures may be attributed to their included language tasks. Most short language measures only assess single language tasks. These include sentence-recall (Seeff-Gabriel, Chiat & Dodd 2010, Sturner et al. 1993), direction-following (Cole & Fewell 1983) and nonword repetition tasks (Dollaghan & Campbell 1998, Gathercole et al. 1994). Yet, single language tasks are at best, moderately accurate at distinguishing the presence and absence of language difficulties (Pawlowska 2014). Other measures combine two or three tasks (e.g. Fluharty 1973, Gardner et al. 2006) but do not include a task combination empirically found to be sufficiently accurate.

The insufficient accuracy of available measures may also be due to inadequate development. The properties of assessments are related to the quality of their individual test items (Kline 2013). A critical step in measure development is individual item evaluation and selection. Individual items can be examined for their validity, reliability, difficulty and discrimination capacity via item analysis. When constructing an assessment, it is imperative to trial a large set of test items, in order to discard items that do not meet minimally acceptable standards and select items that make the greatest contribution to an assessment's validity and reliability (Crocker & Algina 1986). This process can maximise an assessment's psychometric properties, while reducing its length, and may result in a short language measure with sufficiently high accuracy, validity and reliability.

Short language measures that could be used to assess young school-aged children have often been developed without particular attention to the properties of their individual test items (Cole & Fewell 1983, Dollaghan & Campbell 1998, Fluharty 1973, Gardner et al. 2006, Gathercole et al. 1994, Seeff-Gabriel, Chiat & Dodd 2010, Sturner et al. 1993). They

have been constructed via writing a set of test items followed by measure standardisation and validation to evaluate a measure's overall psychometric properties. No published study aiming to develop a short language measure for young school-aged children has trialled a large set of test items in order to select those with the highest individual characteristics.

The Short Language Measure (SLaM) Creation Project was initiated in 2016 (Murdoch Children's Research Institute/ The University of Melbourne) to attempt to produce a sufficiently accurate short language measure for children in their first year at school. Preliminary work explored the components (constructs) of language and identified the most effective combination of tasks to assess language by analysing results from a large-population longitudinal study (Matov et al. 2018). Participants (n=995 at age 5 and n=1217 at age 7) were assessed with eight language tasks from commercially-available measures for speech-language pathologists. The agreement and accuracy of task combinations were investigated in comparison to the results from an omnibus language measure (the Clinical Evaluation of Language Fundamentals-4th edition) (Semel, Wiig & Secord 2006). The results indicated that one main component of language was assessed by all tasks, indicating that a short language measure should assess general language ability by combining scores from included tasks. Combining a direction-following and a sentence-recall task resulted in the greatest agreement with the omnibus measure and exceeded the criterion for good discriminant accuracy (sensitivity=94%, specificity=91%, accuracy=91%). These findings warrant the production of a novel measure, comprising a direction-following and a sentence-recall task, with a simple marking system so that it may be administered by teachers and paraprofessionals, as well as speech-language pathologists.

The current study

The current paper presents the subsequent phases of the Short Language Measure Creation Project. The purpose of this study was to create and validate a short language measure for children in their first year at school that (1) can accurately distinguish children who require further assessment, (2) requires around 10 minutes or less to administer and, (3) is easy to administer, mark and interpret. The specific study aims were to evaluate the individual item characteristics of a large sample of direction-following and sentence-recall test items, select the highest quality test items to form the Short Language Measure (SLaM), and evaluate the properties of the SLaM in an independent sample to determine whether it is suitable for clinical use. A sub-aim of the study was to investigate the accuracy of parent and teacher concern to identify children with low language ability to draw conclusions about the benefits of the SLaM's use to support language referrals in comparison to referrals based on parents or teacher concern.

Method

The short language measure (SLaM) was developed and validated across two phases, following the preliminary study that determined the measure's content (Matov et al. 2018). The SLaM was created during Phase one and validated during Phase two, using two independent samples. Refer to Figure 1 which details Phase one and two participant recruitment. In Phase one, 160 test items were written and trialed in 56 participants, who were also assessed with a reference measure, the Clinical Evaluation of Language Fundamentals-fourth edition (CELF-4). Participants were selected from an initial representative sample of 73 recruited children to obtain a final sample with comparable proportions of low and normal-range language abilities. Comparable proportions were desired to ensure that each test item had a sufficient representation of responses across varying language ability levels (Kline 2013). The language abilities of participants were

estimated during recruitment via parent and teacher report and were later determined by their performance on the reference measure. Traditional item analysis was conducted to review the individual characteristics of test items and 40 of the highest quality items were selected to form SLaM. In Phase two, a different representative sample of 126 children was assessed with SLaM and the reference measure (CELF-4) to evaluate SLaM's validity, reliability and accuracy.

Insert Figure 1.

The Short Language Measure (SLaM)

The SLaM is a short measure designed to briefly evaluate a child's English language ability in their first year of school and determine whether they require further evaluation by a speech-language pathologist. It requires approximately 10 minutes to administer and score and can be delivered by a speech-language pathologist or education professional. The assessment is composed of 40 test items, made up of two language tasks. The direction-following items ask children to point to pictures on a stimulus card. The sentence-recall items require children to repeat sentences verbatim. Children receive a point for each item they answer correctly. All items are administered and the final score is derived by adding all item points. Children scoring below a certain score are advised to receive additional testing by a speech-language pathologist. The SLaM was designed to assess the English language ability of children. Hence, it is likely to underestimate the true language ability of dual language learners. The SLaM was developed by the primary author, a speech-language pathologist, with guidance from an expert panel (remaining authors): a speech-language pathologist, biostatistician, audiologist, child health expert and academic. An independent linguist, Dr Coleen Holt, reviewed the final measure for grammatical correctness and appropriateness. The following outlines the SLaM's development and validation phases.

Phase one (measure creation)

Participant recruitment

Participants were recruited from Foundation year (the first year of formal education in Victoria) from three primary schools in Melbourne, Australia (age range = 5 years, 1 month to 6 years, 7 months). Foundation year teachers (*n*=12) from participating schools handed out recruitment packs to all families in their classes. These included parent questionnaires which obtained basic demographic information and ascertained whether parents were concerned about their child's language. Basic demographic questions determined a child's level of English exposure, exposure to a language other than English, school background, and whether a child had received speech-language therapy. Two hundred and forty-one recruitment packs were distributed across participating schools and 89 of these were returned. Teachers were then asked to indicate whether they had concerns about a participating child's language ability, following an initial introduction session on language. Children were excluded if they had been assessed in the last 12 months with the CELF-4 (*n*=5) or had less than two years of full-time exposure to English in early education if a child was born in a non-English speaking country (*n*=11). Consequently, 73 children were eligible to participate.

Participant selection and assessment procedure

Fifty-six of the 73 eligible children were selectively assessed. Children were selected for assessment to obtain a final sample with even proportions of children with low- and normal-range language abilities to ensure that each test item had a sufficient representation of responses across ability levels (Kline 2013). During recruitment, parents and teachers indicated whether they had any concerns about participating children's language skills. This occurred following a brief written explanation of language, which was referred to as 'the way

they use or understand spoken words or sentences'. Based on these concerns, 25 children with parent and/or teacher concern and 25 children without concern were selected for assessment. The percentage of children with low and normal-range language skills was confirmed by evaluating participant's reference measure (CELF-4) results. Low language ability was categorised if a child scored less than or equal to 1.25 SD (≤ 81) below the mean on the Receptive and/ or Expressive Language Scores of the CELF-4. A further 6 children without parent or teacher concern were assessed to increase the number of children with normal-range language skills in the final sample, to ensure a relatively even distribution of abilities.

Participants were assessed by an examining speech-language pathologist across one or two assessment sessions (within a week of each other) in a quiet room at their primary school with 160 test items and the CELF-4. The order of assessment presentation was alternated to reduce order effects.

Measures

Test items

Based on the findings of prior research, one hundred and sixty direction-following and sentence-recall test items were created for initial testing (Matov et al. 2018). The direction-following items asked children to follow instructions by pointing to pictures on a stimulus card. Before writing, two direction-following stimulus cards were created. Each contained nine pictures of common objects (card 1) or animals (card 2). Common nouns were included so that the directions measured understanding of instructional concepts and ability to process and retain verbal information, rather than the semantic understanding of nouns.

Forty-five direction-following items were written for each stimulus card. To write these items, the authors constructed a table of specifications that listed all direction class types (positional, temporal, conditional, sequential, choice, exclusion and inclusion), subdivided into directional concepts. For example, temporal directions were subdivided into directional concepts 'before', 'after', 'while' and 'same time'. At least two items were constructed for each directional concept. These usually varied in their direction level (one-, two- and three-step directions), the inclusion of additional elements, such as plurals and modifiers, or a second directional concept. An additional stimulus card and two basic direction-following items were constructed as instructional items. Participants were asked to identify the pictures in each stimulus cards before beginning to ensure they were recognised.

The sentence-recall items required children to repeat sentences verbatim. A diverse range of 70 sentences were created to assess key morphosyntactic structures and linguistic aspects that children are expected to have acquired prior to school entry. Sentences were constructed to include the full range of sentence categories (active/passive declarative, imperative and interrogative) and structures (simple, compound and complex sentences, including relative and subordinate clauses). Before writing, a table was created to list possible combinations of sentence categories and structures. Several items were then written for each of these specifications, varying by their inclusion of grammatical components (prepositions, pronouns, and modifiers), inflection (regular and irregular plural and past tense, third person singular agreement, possessive nouns and contractions) and grammatical person.

As Polisenska, Chiat and Roy (2015) recommended minimizing the impact of cultural bias and semantic familiarity by use of familiar words in sentence-recall items, sentences were constructed to exclude proper names and include common vocabulary that was specific to Australian English. The content of the items was reviewed to ensure diversity of

grammatical components via the construction of content analysis tables. Tables listed included pronouns, prepositions and conjunctions and categorised sentences by person, tense and plurality, and by sentence category, type and the inclusion of grammatical components. Two simple sentences were constructed as trial items for inclusion in the task's instructions. The instructions and content of the items was then further reviewed by the authors and an independent speech-language pathologist for grammatical and semantic correctness, gender balance and perceived administration ease.

Test item administration

The order of the direction-following items assessed by each stimulus card was randomised for each participant to reduce order effects. The order of the sentence-recall items was similarly randomised and split into two separate tests. Consequently, participants completed two direction-following and two sentence-recall tests, which were presented in random order. All test items were scored as correct or incorrect to ensure marking ease.

Reference measure and defining low language ability (LL)

The CELF-4 was used to determine the language ability of children and categorise them into low language (LL) and normal-range language (NL) ability subgroups (Semel, Wiig & Secord 2006). The CELF-4 is a clinical tool used to identify and evaluate language disorders in children 5-21 years old. It was selected as a reference measure because it has Australian norms and its psychometric properties exceed those of other available language measures (Betz, Eickhoff & Sullivan 2013). The CELF-4 only measures a child's English language ability and is not culturally or linguistically adapted for groups from diverse backgrounds. Hence, the measure can overidentify dual language learners (DLLs) as having a language disorder. The assessment was developed for the US population and then adapted and

standardised for the Australian population on 825 randomly recruited children, excluding those with a diagnosed condition or whose primary language was not English. Its internal consistency reliability ranged from 0.72-0.92 between subtests. Its sensitivity and specificity were 83% and 90% respectively, at a score of 1SD or more below the mean, when investigating a small sample of 49 children with clinically diagnosed language disorders. Yet, the authors noted a sensitivity of 100% at a score of 1 and 1.5SD or more below the mean in the US sample, when investigating a larger sample of 136 children with language disorders.

Six CELF-4 subtests were administered: Concepts and Following Directions, Word Structure, Recalling Sentences, Formulated Sentences, Sentence Structure and Word Classes. A raw CELF-4 composite score was created to determine the English language ability of children (irrespective of their ages) by adding the percentage of correct responses for each subtest. This was used in all analyses except when children were dichotomized into LL and NL subgroups using the CELF-4 normative data. To define LL/NL, the scores from subtests were converted to scaled scores and used to form three composites: Core Language Score (CLS), Expressive Language Score (ELS) and Receptive Language Score (RLS), as per the directions in the CELF-4 manual. Children were defined as having LL in English if they scored less than or equal to 1.25 SD (≤ 81) below the mean on their Receptive and/ or Expressive Language Scores. A cut-off point of 1.25 standard deviations below the mean was adopted as it has been widely used in several prominent epidemiological studies (Johnson, Beitchman & Brownlie 2010, Matov et al. 2018, Tomblin et al. 1997).

Analysis

The characteristics of each test item were examined through a series of calculations known as item analysis (Kline 2013). These evaluated the validity, reliability, difficulty, and

discrimination capacity of each test item. Table 1 describes the calculation and interpretation of each characteristic.

Insert Table 1.

Item selection

A stepwise procedure was used to select the final SLaM test items (Crocker & Algina 1986). Initially, each item's characteristic values were reviewed. Items that did not meet the minimally acceptable standards were excluded from item selection (Ebel & Frisbie 1991). These were items that had validity, reliability or discrimination values below 0.30 or items with total group difficulty values below 0.20 (considered too difficult) or above 0.80 (considered too easy). To evaluate the functioning of remaining items, items were ranked by highest validity, followed by highest discrimination and highest reliability.

The number of items to include in SLaM was explored by evaluating the overall performance (validity, reliability and accuracy) of different-size groups of the highest ranked items by analysing their total item scores. Validity was calculated by evaluating the Pearson's correlation coefficient between total item scores and raw CELF-4 scores. Kuder-Richardson (KR-20) calculations were performed to investigate the reliability of item groups. The KR-20 calculation investigates the internal consistency of items groups, which is the extent to which groups of items are correlated. The calculation is similar to Cronbach alpha but is used for dichotomously scored items (Crocker & Algina 1986). The accuracy of item groups was estimated, in reference to the LL/NL subgroups, using receiver operating characteristic (ROC) curve analysis. ROC curve analysis plots sensitivity against 1- specificity for a range of different cut-off points. The area under the curve (AUC) represents the probability that a child will be categorised as having LL/NL correctly by a group of items.

The content of the chosen items was reviewed. Including items that assess the same content reduces the number of items that can effectively contribute to a measure's validity, therefore an item that assessed the same concepts (e.g. directional concept and level) as a higher functioning item was substituted with the next highest priority item (Clark & Watson 1995). Items were also scanned to ensure they contained a range of difficulties with an approximate mean of 0.5, which is considered most useful at differentiating between individuals on an assessment with dichotomously scored items (Ebel & Frisbie 1991). Final items were then sorted based on descending item difficulties and collated to create SLaM.

Phase two (measure validation)

Participant recruitment and procedure

Subsequent to the completion of Phase one, a second sample of Foundation year participants (age range = 5 years to 6 years, 8 months) was recruited from three primary schools in Melbourne, Victoria. Two-hundred and twelve recruitment packs were distributed across participating schools by nine teachers and 126 of these were returned. Recruitment packs included the same parent questionnaire from Phase one, used to obtain demographic information and determine whether parents were concerned about their child's language skills. The same question was asked of teachers for participating children via administration of a teacher questionnaire. The same inclusionary criteria were used as in the first sample. All participants were assessed by the same speech-language pathologist across one or two assessment sessions (within a week of each other) in a quiet room at their primary school with the SLaM and CELF-4. The order of assessment presentation was alternated for each child.

Analysis

SLaM's properties were reviewed with respect to its validity, reliability and accuracy. Validity was calculated by evaluating the Pearson's correlation coefficient between the SLaM and raw CELF-4 scores. Internal consistency reliability was estimated using KR-20 and the accuracy of SLaM was calculated using ROC curve analysis. SLaM's properties were evaluated by examining the results from all Phase two participants (n=126) and a sub-sample who were monolingual English speakers (*n*=101). The second analysis was conducted to determine whether SLaM's properties varied with the exclusion of DLLs, given that the CELF-4 and the SLaM are designed to assess English and are not always appropriate to assess the language ability of DLLs.

Results

Participant characteristics (Phase one and two)

All participants were born in Australia and were in their first year of formal schooling. Table 2 shows the demographic data for participants. There were no significant differences in participant characteristics between the LL and NL subgroups in Phase one. In Phase two, the LL subgroup had a significantly higher rate of children who spoke a language other than English (LOTE). When comparing Phase one and two participants, Phase one participants were significantly more disadvantaged, spent significantly less time in early education, had a higher rate of speaking a LOTE and were significantly less likely to have previously seen a speech-language pathologist. This is likely to reflect the different demographics of participating schools across study phases. The prevalence of LL was 52% in Phase one (due to selective sampling) and 27% in Phase two. Table 3 depicts the CELF-4 composite score distributions for participants. Both samples had total composite score means below the population mean (100) due to a higher rate of participation of children with LL (more prevalent in Phase one).

Insert Table 2.

Insert Table 3.

Phase one (measure creation)

Item analysis

All but one item had positive item characteristic values (validity, reliability, difficulty and discrimination). This indicates that NL participants were more likely to get each item correct and that participants who were more likely to get an item correct, tended to have higher CELF-4 and overall item scores. The remaining item, a sentence-recall item which required children to repeat "the father bought a game for his daughter who had her birthday on Friday", had item characteristic values of 0 as all participants scored incorrectly. Item reliabilities ranged from 0-0.80, discrimination values ranged from 0-0.75 and validities ranged from 0-0.76. Items with the highest validities tended to have the highest reliabilities (relationship to overall item scores) and discrimination (capacity to differentiate between LL and NL), providing evidence for a strong relationship between the CELF-4 and trialed test items. Forty-six (29%) of the 160 test items were excluded from item selection: 11 items had validities below 0.30, 10 had reliabilities below 0.30, 40 had discrimination values below 0.30, 16 items had p-values (difficulties) below 0.20 and 15 items had p-values above 0.80. For example, direction-following item two (point to the small fish) had a p-value of 0.98. Only two children got this item incorrect. As a majority of children (n=54) scored correctly, it had a very low discrimination value (0.04), validity (0.15) and reliability (0.13) and was therefore excluded from item selection.

Item selection

Figure 2 depicts the validity, reliability and accuracy of different-size groups of top performing, high-functioning items. Assessing all 160 test items resulted in a validity of 0.95 and reliability and area under the ROC curve (AUC) of 0.98. The top 55 performing items had the same validity, reliability and AUC, suggesting that administering the lowest performing 105 items provided little additional value for language assessment and LL/NL discrimination.

Insert Figure 2.

Performance values were relatively high for all groups of test items investigated. Yet, these values were likely to change in the subsequent trial (especially for small groups of items) due to different sampling characteristics, test-retest reliability and measurement error (Smith & McCarthy 1995). Kruyen, Emons and Sijtsma (2012) advised retaining at least 40 test items for acceptable decision quality regarding individuals, based on empirical simulations. The AUC (related to a measure's accuracy) reached its maximum of 0.98 at 40 items. Forty items were also the minimum number of items with the ideal mean item difficulty of 0.48, where the highest discrimination between LL and NL is expected due to the sample containing 48% of children with NL. Hence, 40 test items were retained for SLaM.

The top 20 of the highest performing direction-following (10 from each stimulus card) and sentence-recall items were scanned for their content. Five pairs of items amongst the top 20 assessed the same directional concepts. For example, two items assessed the directional concept 'furthest'. The second ranked item within these pairs was substituted for the next highest functioning item that assessed a different directional concept to those already

included in the top 20. Three sentence-recall items were also substituted to include a greater variety of sentence structures (e.g. passive sentences), sentence elements (e.g. irregular noun, negative) and tenses (e.g. future).

The range of content assessed by selected items was deemed to be clinically appropriate by the authors and an independent linguist. Chosen direction-following items contained a fairly even range of possible direction levels and assessed all but two (choice and conditional) direction class types. Chosen sentence-recall items contained a range of different sentence structures. There were considerably more sentences that assessed simple sentence structures but these items all contained a range of different sentence elements (including coordination, prepositions, pronouns, noun and verb modifiers, irregular verbs and nouns, and possessive nouns) and tenses. The higher number of simple sentence structures is likely to reflect the ability level of the age group and the dichotomous item scoring method. The sentence length of sentence-recall items was normally distributed, ranging from 4-12 words with mean length of 8 words.

Characteristics of the final test items

The validity, reliability and AUC of assessing the chosen 40 test items were 0.93, 0.97 and 0.98, respectively. All items were considered 'good' quality as their validities, discrimination values and reliabilities were well above the minimum requirement (0.30) for a 'good item' (Crocker & Algina 1986, Ebel & Frisbie 1991). Item validities ranged from 0.47-0.76, discrimination values ranged from 0.44-0.75 and reliabilities ranged from 0.49-0.80. The chosen items also consisted of a range of difficulties (p–value range = 0.26-0.66) with a mean p-value of 0.48. Results from Phase one indicated high functioning of SLaM but these values needed to be replicated in a second sample to ensure consistency (Smith & McCarthy 1995).

Phase two (measure validation)

Measure characteristics

The validity and reliability of the SLaM in the Phase two sample (*n*=126) were 0.89 and 0.93, respectively. ROC curve analysis, depicted in Figure 3, indicated an AUC of 0.98, with a sensitivity of 94%, specificity of 93% and accuracy of 94% at a cut-off point of scoring less than or equal to 22 (specified as it resulted in the highest accuracy, followed by sensitivity). At this specific cut-off point, 32 out of 34 participants with LL and 86 out of 92 participants with NL were correctly classified by the measure. Table 4 lists the range of cut-off points and their respective sensitivity, specificity and accuracy values. Administration and scoring of the SLaM required approximately 10 minutes. The validity and reliability of SLaM in the Phase two sub-sample who were monolingual (*n*=101) were 0.87 and 0.91, respectively. ROC curve analysis indicated an AUC of 0.97, with a sensitivity of 91%, specificity of 94% and accuracy of 93% at a cut-off point of 22 or less. The same cut-off point was identified in both samples as having the highest accuracy, followed by sensitivity. SLaM's properties only differed slightly when DLLs were excluded (validity and reliability reduction of 0.02, accuracy reduction of 0.01%). The slight changes in exact values were likely to reflect greater uncertainly (an increase in measurement error) as the number of participants reduced. Including DLLs also increased the classification accuracy of the SLaM as DLLs were more likely to be classified into the LL subgroup based on their CELF-4 results and score below the cut-off point on the SLaM. Yet, without an appropriate battery of assessments which considers the language ability of children in all their spoken languages, it is unknown which of these children truly have low language ability.

Insert Figure 3.

Insert Table 4.

Parent and teacher low language ability detection

The sensitivity, specificity and accuracy of LL/NL detection by parents and teachers was investigated to explore the capacity of referrers to detect children with low English language ability. These values are presented in Table 5 and were measured by comparing parent/teacher concern about children's language skills (from parent questionnaire data) to children's LL/NL (CELF-4) categories. The results from both study phases indicate that parents and teachers were moderately accurate at detecting low and normal-range language ability. Parents in Phase one and two had a relatively high specificity for NL detection (96% and 89%, respectively), as parents were in most cases, not concerned about a child with NL. Conversely, parent sensitivity for LL detection was very low (31% and 23%), indicating that parents participating in both study phases were unlikely to be concerned about the language skills of a child with LL. Teachers had higher sensitivity (69% and 50%), signifying they were more likely to be concerned about a child with LL than parents. However, teachers had a lower specificity for NL detection (74% and 87%), so they were more likely to be concerned about a child with NL than parents. The classification accuracy for true low language ability in this analysis was confounded by the inclusion of DLLs due to the use of the CELF-4 as the sole indicator of low language ability.

Insert Table 5.

Discussion

The present study created and evaluated the reliability and validity of a short language measure (SLaM) for children in their first year at school. Results indicate that SLaM has excellent psychometric properties and is suitable for clinical use to identify children needing further language evaluation.

Implications for practice

Current approaches for detecting children with language difficulties rely on parent and/or teacher report of concern, and subsequent referral to speech-language services. However, results from the present study confirm previous findings, which indicate that parents and teachers are not sufficiently accurate at identifying children who require referral (Antoniazzi, Snow & Dickson-Swift 2010, Bishop & McDonald 2009, Broomfield & Dodd 2004, Jessup et al. 2008). Both parents and teachers were only moderately accurate at identifying children with normal or low language ability based on concern. This indicates that many children scoring low on a language assessment would not be considered by their parents or teachers as having language difficulties, and that teachers are sometimes concerned about the language skills of children with normal-range language ability. As SLaM's detection rate of children identified as having language difficulties was substantially higher than that of parents and teachers, its administration could improve the future accuracy of language referrals and ensure that children with language difficulties are appropriately identified and supported. The use of SLaM may also afford speech-language pathologists with more time to manage those in need by reducing the rate of inappropriate language referrals. Hence, speech-language pathologists could spend more time organising language intervention programs for children.

Study and measure development strengths

SLaM was developed by including a task combination found to be sufficiently accurate at identifying low language skills (Matov et al. 2018). It was also developed using item analysis and selection procedures that were omitted in previously published short language measure development studies (Dollaghan & Campbell 1998, Gardner et al. 2006, Gathercole et al. 1994, Seeff-Gabriel, Chiat & Dodd 2010, Sturner et al. 1993). Application of these additional measure development steps may explain SLaM's high psychometric properties.

Language measures are often designed to identify children who score at the lower end of the normal distribution based on arbitrary cut-off points (Plante & Vance 1994, Spaulding et al. 2006). The current study used ROC curve analysis to determine a specific cut-off point that ensured maximal effectiveness in differentiating children, prioritising accuracy, followed by sensitivity. Use of ROC curve analysis also ensured that the distinct characteristics of participants (for example, the proportion of LL) had minimal impact on cut-off point calculations. Other cut-off points, along with their sensitivity and specificity rates, are also reported (Table 4). Administrators may wish to choose a different cut-off point depending on SLaM's intended purpose. For example, clinicians using the measure as confirmatory screening tool (to assess referrals) may choose a cut-off point with higher specificity (rate of children correctly identified as having age-appropriate language skills) to combat over-referrals; whereas administrators using the measure as a general screener (to assess all children) may decide to use a cut-off point with higher sensitivity (detection of children with language difficulties).

The number of items to include in SLaM was investigated by exploring the overall properties of different-size groups of high functioning items. Forty items were selected because this was the minimum number of items with the highest area under the ROC curve (which corresponds to a measure's accuracy). Kruyen, Emons and Sijtsma (2012) also

determined that measures should include at least 40 items for acceptable decision accuracy. However, this study was not specific to language ability assessments.

Limitations and suggestions for future research

More research is needed to investigate the ideal length of a short language measure that considers the compromise between administration time and its effects on the psychometric properties of an assessment. SLaM requires significantly less time to administer than an omnibus language measure but could still be considered too time consuming for certain language screening programs. Hence, it would be beneficial to trial a shorter version of the assessment in the future, given that the assessment of fewer items in Phase one still resulted in acceptable accuracy levels.

A single speech-language pathologist conducted all of the assessments in Phase one and two of this study to ensure consistency of assessment delivery. However, further study is needed to evaluate SLaM's effectiveness when administrated by different assessors, such as teachers and teaching aides. This would establish the degree to which different administrators might influence test results. Permitting teachers and paraprofessionals to administer SLaM would provide speech-language pathologists with more time to deliver therapy. Sturner et al (1993) previously found that paraprofessionals were able to achieve high degrees of agreement with speech-language pathologists when investigating the effectiveness of a sentence-recall measure. Hence, it is assumed that sufficient training of professionals would not affect children's SLaM results, but this will need to be determined in future trials.

The sample size used to validate SLaM in Phase two (n=126) was larger than the samples in each age group that are used to validate omnibus language measures (e.g. the CELF-4). However, future SLaM studies will benefit from including a larger sample of

children from a range of different schools. The ages of children in their first year at school can vary substantially. Investigation in a larger sample would enable examination of the impact of age on SLaM's accuracy rates and cut-off scores. In addition, while including 30-50 participants is sufficient for an item trial, including a larger sample (of over 100) is often recommended and should be considered in future development studies (Johanson & Brooks 2010).

SLaM's psychometric properties were investigated with respect to children's performance on a single omnibus language measure. Yet, the identification of language difficulties requires the collection and integration of information from multiple sources, including evidence of functional impact (Bishop et al. 2017). Future investigation of SLaM's effectiveness warrants comparison to a range of language measures, including clinician diagnosis of language difficulties.

SLaM was developed and validated using samples of children that included DLLs. This was conducted to respond to the current climate in Australia, in which 21% of Australians speak a language other than English at home (Australian Bureau of Statistics 2016). The inclusion of DLLs had little impact on SLaM's final properties and no impact on the interpretation of its final scores because of the specific analyses applied. There are, however, limitations to solely relying on the SLaM's results to make referrals and get indications about the language skills of DLLs. DLLs may experience language delays and warrant referral to a speech-language pathologist if they fall outside of the normal range. However, the true language ability of DLLs (determined by the comprehensive evaluation of language skills in English and any other spoken language(s)), is likely underestimated by the SLaM – just as the language ability of DLLs is underestimated by other screeners and assessments designed primarily for monolingual, English speakers. This is because the SLaM

and the CELF-4 only measure a child's English language ability. SLaM administrators must be mindful of a child's level of English exposure and proficiency in their other spoken language(s) before considering their suitability for in-depth language evaluation. The impact of administering the SLaM to children who speak a language other than English warrants future investigation.

Summary and Conclusions

Identifying and supporting young school-aged children with language difficulties may improve their learning opportunities.The current study developed and evaluated the effectiveness of a short language measure (SLaM) for children in their first year at school which could improve the accurate identification of children needing language support. The SLaM has high validity, reliability and discrimination accuracy when compared to a readily-used omnibus language measure. It is an effective measure that is suitable for clinical use to assess language ability and determine whether children need additional consultation with a speech-language pathologists.

References

ANTONIAZZI, D., SNOW, P. and DICKSON-SWIFT, V., 2010, Teacher identification of children at risk for language impairment in the first year of school. *International Journal of Speech-Language Pathology*, **12**, 244-252.

AUSTRALIAN BUREAU OF STATISTICS, 2016, *Multiculturalism* (Canberra, ACT: Australian Bureau of Statistics).

BETZ, S. K., EICKHOFF, J. R. and SULLIVAN, S. F., 2013, Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools*, **44**, 133-146.

BISHOP, D. V. M. and MCDONALD, D., 2009, Identifying language impairment in children: combining language test scores with parental report. *International Journal of Language & Communication Disorders*, **44**, 600-615.

BISHOP, D. V. M., SNOWLING, M. J., THOMPSON, P.A. and GREENHALGH, T., 2017, Phase 2 of CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development: terminology. *Journal of Child Psychology & Psychiatry*, **58**, 1068-1080.

BROOMFIELD, J. and DODD, B., 2004, Children with speech and language disability: caseload characteristics. *International Journal of Language & Communication Disorders*, **39**, 303-324.

CLARK, L. A. and WATSON, D., 1995, Constructing validity: basic issues in objective scale development. *Psychological Assessment*, **7**, 309–319.

COLE, K. N. and FEWELL, R. R., 1983, A quick language screening test for young children: the Token Test. *Journal of Psychoeducational Assessment*, **1**, 149–153.

CROCKER, L. and ALGINA, J., 1986, *Introduction to Classical and Modern Test Theory* (New York, NY: Holt, Rinehart and Winston).

CURRAN, A., FLYNN, C., ANTONIJEVIC-ELLIOTT, S. and LYONS, R., 2015, Non-attendance and utilization of a speech and language therapy service: a retrospective pilot study of school-aged referrals. *International Journal of Language & Communication Disorders*, **5**, 665-675.

DOCKRELL, J. E. and MARSHALL, C. R., 2015, Measurement issues: assessing language skills in young children. *Child and Adolescent Mental Health,* **20**, 116-125.

DOLLAGHAN, C. and CAMPBELL, T. F., 1998, Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, **41**, 1136–1146.

EBEL, R. L. and FRISBIE, D. A., 1991, *Essentials of Educational Measurement* (Englewood Cliffs, NJ: Prentice Hall).

FLUHARTY, N. B., 1973, The design and standardization of a speech and language screening test for use with preschool children. *Journal of Speech and Hearing Disorders*, **39**, 75–88.

GARDNER, H., FROUD, K., MCCLELLAND, A. and Van Der Lely, H. K. J., 2006, Development of the Grammar and Phonology Screening (GAPS) test to assess key markers of specific language and literacy difficulties in young children. *International Journal of Language and Communication Disorders*, **41**, 513–540.

GATHERCOLE, S. E., WILLIS, C. S., BADDELEY, A. D. and EMSLIE, H., 1994, The children's test of nonword repetition: a test of phonological working memory. *Memory*, **2**, 103–127.

JESSUP, B., WARD, E., CAHILL, L. and KEATING, D., 2008, Teacher identification of speech and language impairment in kindergarten students using the Kindergarten Development Check. *International Journal of Speech-Language Pathology*, **10**, 449-459.

JOHANSON, G. A. and BROOKS, G. P., 2010, Initial scale development: sample size for pilot studies. *Educational and Psychological Measurement*, **70**, 394–400.

JOHNSON, C. J., BEITCHMAN, J. H. and BROWNLIE, E., 2010, Twenty-year follow-up of children with and without speech-language impairments: family, educational, occupational, and quality of life outcomes. *American Journal of Speech-Language Pathology*, **19**, 51-65.

KLINE, P., 2013, *The New Psychometrics: Science, Psychology, and Measurement* (New York, NY: Routledge).

KRUYEN, P. M., EMONS, W. H. M. and SIJTSMA, K., 2012, Test length and decision quality in personnel selection: when is short too short? *International Journal of Testing*, **12**, 321–344.

LAW, J., BOYLE, J., HARRIS, F., HARKNESS, A. and NYE, C., 2000, The feasibility of universal screening for primary speech and language delay: findings from a systematic review of the literature. *Developmental Medicine and Child Neurology*, **42**, 190-200.

MATOV, J., MENSAH, F., COOK, F. and REILLY, S., 2018, Investigation of the language tasks to include in a short-language measure for children in the early school years. *International Journal of Language & Communication Disorders*. https://doi.org/10.1111/1460-6984.12378.

MROZ, M., 2006, Providing training in speech and language for education professionals: challenges, support and the view from the ground. *Child Language Teaching and Therapy*, **22**, 155-176.

NASH, H., LEAVETT, R. and CHILDS, H., 2011, Evaluating the GAPS test as a screener for language impairment in young children. *International Journal of Language & Communication Disorders*, **46**, 675–685.

NORBURY, C., GOOCH, D., WRAY, C., BAIRD, G., CHARMAN, T., SIMONOFF, E., VAMVAKAS, G. and PICKLES, A., 2016, The impact of nonverbal ability on prevalence and clinical presentation of language disorder: evidence from a population study. *Journal of Child Psychology and Psychiatry*, **57**, 1247-1257.

PAWLOWSKA, M., 2014, Evaluation of three proposed markers for language impairment in English: a meta-analysis of diagnostic accuracy studies. *Journal of Speech, Language, and Hearing Research*, **6**, 2261-2273.

PLANTE, E. and VANCE, R., 1994, Selection of preschool language tests: a data-based approach. *Language, Speech, and Hearing Services in Schools*, **25**, 15–24.

POLISENSKA, K., CHIAT, S. and ROY, P., 2015, Sentence repetition: what does the task measure? *International Journal of Language & Communication Disorders*, **50**, 106-118.

POLL, G. H., BETZ, S. K. and MILLER, C. A., 2010, Identification of clinical markers of specific language impairment in adults. *Journal of Speech, Language & Hearing Research*, **53**, 414-429.

SADLER, J., 2005, Knowledge, attitudes and beliefs of the mainstream teachers of children with a preschool diagnosis of speech/language impairment. *Child Language Teaching and Therapy*, **21**, 147-163.

SEEFF-GABRIEL, B., CHIAT, S. and DODD, B., 2010, Sentence imitation as a tool in identifying expressive morphosyntactic difficulties in children with severe speech difficulties. *International Journal of Language and Communication Disorders*, **45**, 691–702.

SEMEL, E., WIIG, E. H. and SECORD, W., 2006, *Clinical Evaluation of Language Fundamentals—Fourth Edition, Australian Standardised Edition* (Sydney, NSW: PsychCorp).

SMITH, G. T. and MCCARTHY, D. M., 1995, Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment*, **7**, 300–308.

SPAULDING, T. J., PLANTE, E. and FARINELLA, K. A., 2006, Eligibility criteria for language impairment: is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools*, **37**, 61–72.

STURNER, R. A., KUNZE, L., FUNK, S. G. and GREEN, J. A., 1993, Elicited imitation: its effectiveness for speech and language screening. *Developmental Medicine and Child Neurology*, **35**, 715–726.

TOMBLIN, J. B., RECORDS, N. L., BUCKWALTER, P., ZHANG, X., SMITH, E. and O'BRIEN, M., 1997, Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research*, **40**, 1245–1260.

Figures and tables

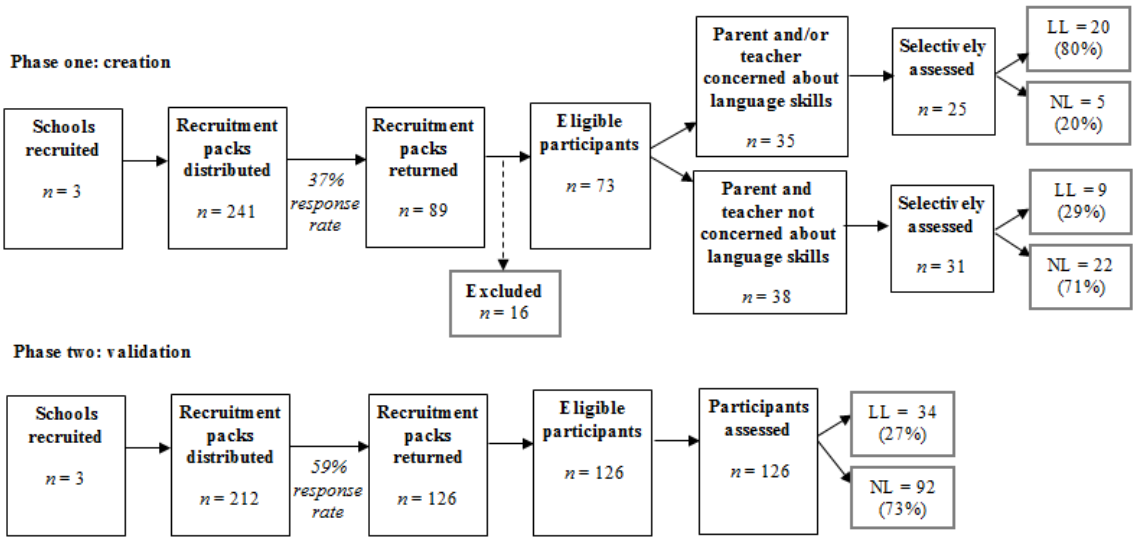

Figure 1. Participant recruitment and selection for assessment flowchart, Phase one and two

Note: LL = low language ability = less than or equal to 1.25 SD below the mean (≤ 81) on the Receptive and/ or Expressive Language Scores of the CELF-4; NL = normal-range language ability.

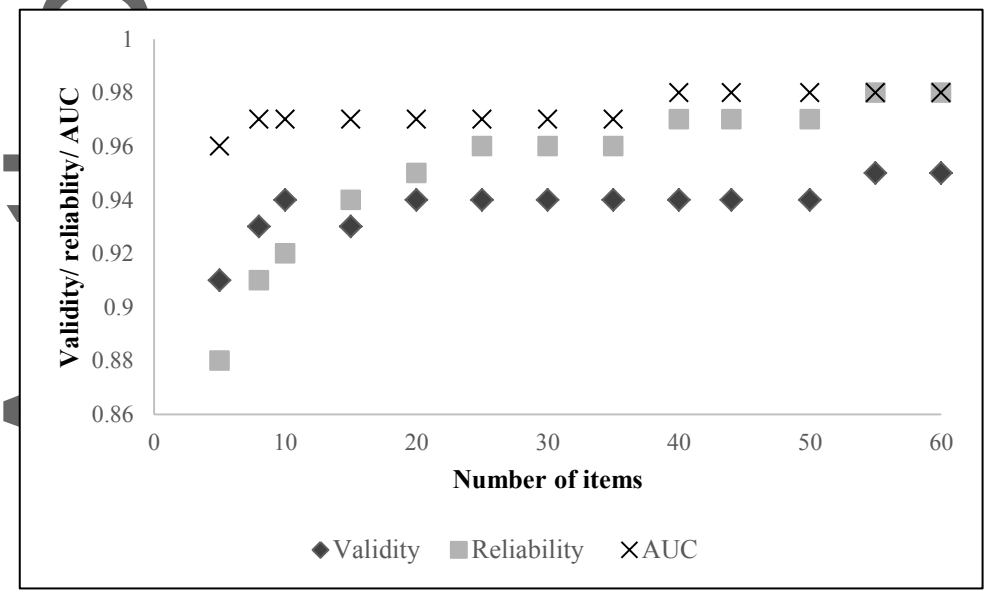

Figure 2. Validity, reliability and area under the Receiver Operating Characteristics curve (AUC) for top performing test items

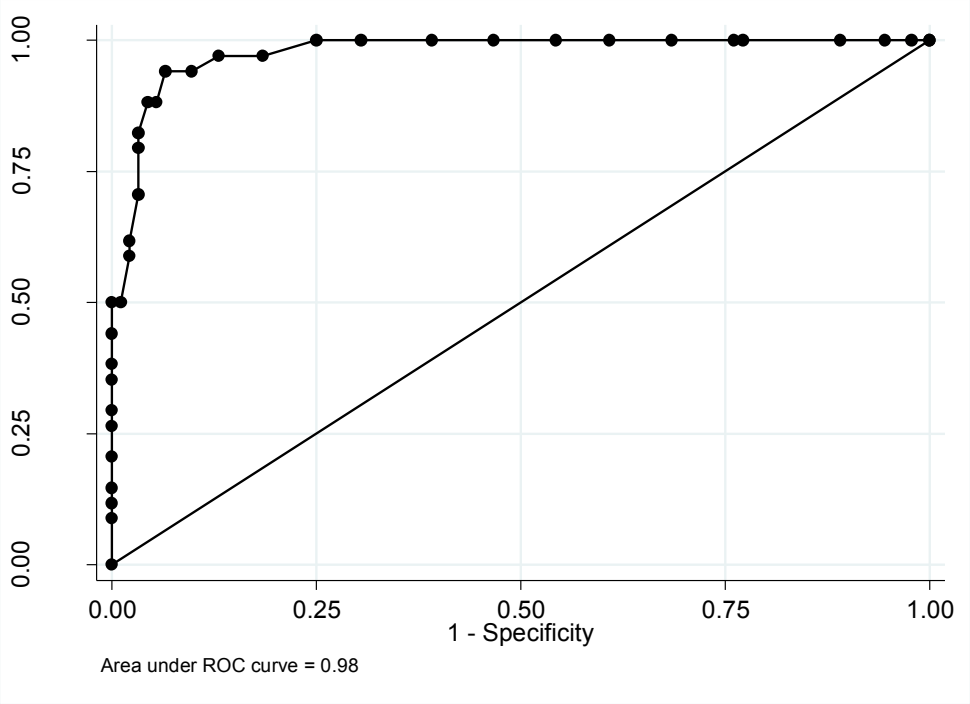

Figure 3. Receiver Operating Characteristics (ROC) curve for the Short Language Measure (SLaM)

Table 1. Item analysis calculations with item exclusion and selection criteria

Item value	Interpretation	Calculation	Item exclusion	Item selection
Validity	Capacity to measure language ability/ the relationship between an item and overall CELF-4 results	Point-biserial correlation coefficient[1] between a test item and raw CELF-4 scores	< 0.3	Highest value (first priority)
Reliability	Internal consistency of an item/ the relationship between an item and total item scores[2]	Point-biserial correlation coefficient[1] between a test item and overall test item scores[2]	< 0.3	Highest value (third priority)
Difficulty (p-value)	Difficulty level of an item	Proportion of correct item responses. Calculated for the total sample and LL and NL subgroups	< 0.2 & > 0.8	Range of values, with a mean ~0.5
Discrimination	Item's capacity to differentiate between	Subtracted proportion of correct responses of LL	< 0.3	Highest value

NL and LL participants	from NL subgroup	(second priority)	

Note: [1]Point biserial correlation coefficient is used to calculate correlations between dichotomous and continuous variables. [2]Total test scores were calculated by adding scores from 160 test items. Item exclusion = items were excluded if their characteristic score fell within this value. Item selection = values of a characteristic prioritised for item selection.

Table 2. Summary of participant characteristics (Phase one and two)

Characteristic	Phase one sample			Phase two sample		
	Total	LL	NL	Total	LL	NL
Participants (*n*=)	56	29	27	126	34	92
Percentage rate of subgroup [c]	100	52	48	100	27	73
Age (mean, SD) [c]	5.7 (0.4)	5.6 (0.3)	5.7 (0.4)	5.3 (0.5)	5.5 (0.5)	5.3 (0.5)
Male, n (%)	55	66	44	56	71	51
Mean SEIFA index [c]	977	971	983	1060	1069	1056
English main language (%) [b c]	77	76	78	87	76	91
LOTE (%) [b c]	45	52	37	20	32	15
Kindergarten attendance (%)	95	90	100	97	97	97
Mean years in early education [c]	2	1.8	2.2	2.9	2.6	3
Previously seen SLT (%) [c]	7	11	4	29	35	26
Parent concern (%) [a b]	18	31	0.04	13	26	11
Teacher concern (%) [a c]	48	69	26	23	50	13

Note: LL = low language ability; NL = normal-range language ability; Mean SEIFA index = Socio-Economic Index for Areas (SEIFA) Disadvantage (Australian population mean=1000, SD=100) (Australian Bureau of Statistics 2011); English main language = English main language spoken at home; LOTE = language other than English spoken at home; parent concern = parent concerned about their child's language skills, teacher concern = classroom teacher concerned about a child's language skills; [a] indicates significant difference between LL and NL subgroups (Phase one) at $p<0.05$; [b] indicates significant difference between LL and NL subgroups (Phase two) at $p<0.05$; [c] indicates significant difference between Phase one and Phase two total samples at $p<0.05$.

Table 3. Reference measure (CELF-4) composite score distributions

	Phase one sample			Phase two sample		
	Total	LL	NL	Total	LL	NL
N	56	29	27	126	34	92
CLS mean (SD)	83.5 (17.3)	70.6 (10.7)	97.4 (11.0)	92.7 (17.2)	71.5 (10.0)	100.5 (11.9)
min/max	45/126	45/84	84/126	45/135	45/87	82/135
RLS mean (SD)	89.5 (17.0)	76.4 (8.9)	103.6 (11.3)	93.5 (15.1)	77.3 (11.7)	99.5 (11.4)
min/max	60/127	60/94	82/127	51/130	51/100	82/130
ELS mean (SD)	82.5 (17.0)	69.9 (11.2)	96.0 (10.8)	91.7 (16.8)	70.7 (9.1)	99.5 (11.6)
min/max	47/122	47/86	82/122	51/132	51/89	82/132

Note: LL = low language ability = less than or equal to 1.25 SD below the mean (≤ 81) on the Receptive and/ or Expressive Language Scores of the CELF-4, NL = normal-range language ability, CLS = Core Language Score, RLS = Receptive Language Score, ELS = Expressive Language Score. Population mean=100, SD=15.

Table 4. Sensitivity, specificity and accuracy for various Short Language Measure (SLaM) cut-off points (*n*=126)

Cut-off point	Sensitivity	Specificity	Accuracy	% scored
≤ 26	100%	75%	82%	45
≤ 25	97%	82%	86%	36
≤ 24	97%	87%	90%	35
≤ 23	94%	90%	91%	32
≤ 22	94%	93%	94%	30
≤ 21	88%	95%	93%	28
≤ 20	88%	96%	94%	27
≤ 19	82%	97%	93%	24
≤ 17	79%	97%	92%	23
≤ 16	71%	97%	90%	21
≤ 15	62%	98%	88%	18
≤ 14	59%	98%	87%	17
≤ 13	50%	99%	86%	14
≤ 12	50%	100%	87%	13

Note: % scored = percentage of children scoring this value. Mean=25, SD=10.

Table 5. Sensitivity, specificity and accuracy of low language ability detection by parents and teachers

		Sensitivity		**Specificity**		**Accuracy**
			95% CI		**95% CI**	
Phase one	Parents	31.0%	(15.3, 50.8)	96.3%	(81.0, 99.9)	62.5%
	Teachers	69.0%	(49.2, 84.7)	74.1%	(53.7, 88.9)	71.4%
Phase two	Parents	23.5%	(10.7, 41.2)	89.1%	(80.9, 94.7)	71.4%
	Teachers	50.0%	(32.4, 67.6)	87%	(78.3, 93.1)	77.0%

Note: CI = 95% confidence interval. Sensitivity, specificity and accuracy were calculated in relation to low language ability detection based on children's CELF-4 (full language measure) results.