Author/s:
Matov, J;Mensah, F;Cook, F;Reilly, S

Title:
Investigation of the language tasks to include in a short-language measure for children in the early school years

Date:
2018-07-01

Citation:
Matov, J., Mensah, F., Cook, F. & 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 and Communication Disorders, 53 (4), pp.735-747. https://doi.org/10.1111/1460-6984.12378.

Persistent Link:
https://hdl.handle.net/11343/283632

JLCD12378

Research report

Investigation of the language tasks to include in a short-language measure for children in the early school years

Jessica Matov†‡, Fiona Mensah‡, Fallon Cook‡ and Sheena Reilly§

†Department of Audiology and Speech Pathology, University of Melbourne, Carlton, VIC, Australia

‡Murdoch Children's Research Institute, Royal Children's Hospital, Parkville, VIC, Australia

§Menzies Health Institute Queensland, Griffith University, Southport, QLD, Australia

(*Received May 2017; accepted January 2017*)

<FN>Address correspondence to: Jessica Matov, Department of Audiology and Speech Pathology, The University of Melbourne, Carlton, VIC 3053, Australia; e-mail: jessica.matov@mcri.edu.au

Abstract

Background: The inaccurate estimation of language difficulties by teachers suggests the benefit of a short-language measure that could be used to support their decisions about who requires referral to a speech–language therapist. While the literature indicates the potential for the development of a short-language measure, evidence is lacking about which combination of language tasks it should include.

Aims: To understand the number and nature of components/language tasks that should be included in a short-language measure for children in the early school years.

Methods & Procedures: Eight language tasks were administered to participants of the Early Language in Victoria Study (ELVS) at ages 5 (*n* = 995) and 7 (*n* = 1217). These included six language tasks measured by an omnibus language measure (which comprised a direction-following, morphological-completion, sentence-recall, sentence-formation, syntactic-understanding and word-association task) and a non-word repetition and a receptive vocabulary task, measured by two task-specific language measures. Scores were analyzed using principal component analysis (PCA), the Bland and Altman method, and receiver operating characteristic (ROC) curve analysis.

Outcomes & Results: PCA revealed one main component of language that was assessed by all language tasks. The most effective combination of two tasks that measured this component was a direction-following and a sentence-recall task. It showed the greatest agreement with an omnibus language measure and exceeded the criterion for good discriminant accuracy (sensitivity = 94%, specificity = 91%, accuracy = 91%, at 1 SD (standard deviation) below the mean).

Conclusions & Implications: Findings support the combination of a direction-following and a sentence-recall task to assess language ability effectively in the early school years. The results could justify the future production of a novel short-language measure comprising a direction-following and a sentence-recall task to use as a screening tool in schools and to assess language ability in research participants.

Keywords: assessment, language, language impairment, school-age children, screening

<A>**What this paper adds**

<B>*What is already known on the subject*

Evidence suggests that the language ability of young school-aged children can be measured without assessing all the domains and modalities of language. Studies also indicate that combining language tasks can increase the accuracy for low language-ability discrimination; however, these studies have

only investigated three types of tasks using small, and in some cases, clinically referred samples. No population-based studies have evaluated the effectiveness of combining tasks to assess language ability.

<B>*What this paper adds to existing knowledge*

This study investigated the number of components and the language tasks that should be included in a short-language measure for children in their early school years by analysing eight language tasks administered to a large-population sample. The results suggest that combining a direction-following and a sentence-recall task can effectively assess language ability in the early school years.

<B>*What are the potential or actual clinical implications of this work?*

The findings suggest there is potential to develop an accurate short-language measure that could be used to investigate the language abilities of children in schools and in research.

<A>**Introduction**

In the early school years, oral language skills are essential for learning and literacy development. Yet, approximately 10% of children start school with language difficulties (Norbury *et al.* 2016). These children are at risk of poor academic achievement (Johnson *et al.* 2010) and social, emotional and behavioural difficulties (Whitehouse *et al.* 2009), prompting the need for adequate detection and intervention services.

There are large variations in how language difficulties are identified, defined and labelled, but the consensus is that a clinical diagnosis of 'developmental language disorder' (DLD) should arise from the collection of information from multiple sources and include evidence of functional impact (Bishop *et al.* 2016, 2017). Detailed diagnosis is often not feasible in research studies, so language is frequently assessed by administering a standardized language assessment that directly samples a child's performance and compares it with that of other children (Betz *et al.* 2013). In this paper, the term 'low language (LL) ability' will be used to refer to children with significantly lower scores on a standardized language measure. These children require further evaluation for a diagnosis of DLD, which implies evidence of functional impact, sufficient language exposure and unknown aetiology (Bishop *et al.* 2017).

The identification and management of school-aged children with language difficulties is heavily reliant on teacher referral to speech–language therapy services. Yet, the literature suggests that teachers cannot always correctly differentiate children categorized with LL or normal-

range language skills when filling out a communication checklist (Antoniazzi *et al.* 2010). Consequently, over 50% of school-aged children with LL are never referred to speech–language therapy services (Bishop and McDonald 2009, Poll *et al.* 2010, Norbury *et al.* 2016) and over half of school-aged children referred to and assessed by speech–language therapists in clinics have age-appropriate communication skills (Curran *et al.* 2015).

A plausible reason for this is that teachers have limited knowledge about language concepts and lack confidence when answering questions pertaining to language (Stark *et al.* 2016). High under- and over-referral rates could also be due to lack of teacher awareness about language difficulties, confusion about the role of a speech–language therapist or poor knowledge of correct referral pathways. Language difficulties may also be confused with or masked by behavioural issues. The resulting uncertainty that teachers have about which children require referral to speech–language therapy services highlights the need for a new approach.

The inaccuracy of teacher referrals to speech–language therapy services may be reduced by teacher training to raise awareness and understanding of language difficulties and typical language development. Yet, this technique may not control for the different perceptive abilities of teachers. Teacher training may therefore be supplemented by the availability of a short-language measure that could also improve the accuracy of referrals by providing teachers with a fast way to screen the language skills of their students. Screening in this way could improve the appropriate and timely referral of children with LL (who may have DLD), who may not otherwise have been identified, if teachers can demonstrate accurate use of the screening tool. The measure could otherwise be administered by speech–language therapists during a clinician-led screening programme or to confirm the accuracy of referrals.

A limitation to this approach is that the measure can only be used to assess language ability. Thus, children with other communication impairments (e.g., speech or social) would not be identified by the measure as requiring referral to a speech–language therapy service for further assessment. Administrators would need to be mindful of this when using the measure and interpreting its results. An awareness of the measure's limitations and strategies for recognizing other communication impairments would need to be incorporated into teacher training.

A short-language measure may also improve our understanding about LL as it may be used in research to investigate language development. The administration of standardized language measures can be time consuming and costly, which can restrict the collection of language data in population studies. The availability of a short-language measure may overcome this barrier, providing a feasible method to evaluate language ability on a large-scale in population studies.

While some short-language measures have been developed, existing measures lack sufficient accuracy (Dockrell and Marshall 2015, Law *et al.* 2000). Accuracy is the percentage of children correctly classified by a measure as having LL and normal-range language skills. A measure's accuracy is related to its sensitivity (the percentage of children correctly classified as having LL) and

specificity (the percentage of children correctly classified as having normal-range skills). Any short measure must meet acceptable accuracy standards as the misclassification of children can result in unnecessary distress and misuse of resources (if misclassified as having LL) or lack of adequate support (if misclassified as having normal-range skills).

One well-known, commercially available short-language measure is the Clinical Evaluation of Language Fundamentals (CELF) screening test (Semel *et al.* 2004). The measure contains 28 items that assess four different language tasks. It is purported to have sensitivity and specificity rates of 88%, but is used to assess 5–21 year olds, so contains few items that can discriminate the language abilities of young children. An alternative measure with items that are more relevant for young school-aged children is needed.

The first step in the development of a short-language measure is to determine whether it is even possible to assess language ability adequately in a short measure. Language is considered to be a complex system of knowledge, comprising several distinct domains (syntax, semantics and morphology) and modalities (expressive and receptive) (Tomblin and Zhang 2006). This is reflected in the structure of omnibus language assessments that comprehensively assess the domains and modalities of language, with the exception of assessing pragmatics or discourse skills. But to what extent do these separate domains and modalities reflect distinct language abilities? If the domains and modalities are interrelated, then it may not be necessary to administer a large range of language tasks to get an indication of a child's overall language ability.

Several studies have investigated the structure of language ability in population-representative samples. Amongst these studies, there is evidence to indicate that language in the early school years is composed of one main component: that is, the domains and modalities of language are interrelated and contribute to the assessment of one ability (general language ability) (Anthony *et al.* 2014, Colledge *et al.* 2002, Klem *et al.* 2015, Tomblin and Zhang 2006). There is also evidence indicating that language ability is composed of two distinct components of language (specifically, that grammar and semantics are distinct abilities) in the early school years (Lonigan and Milburn 2017, Pentimonti *et al.* 2015, Tomblin and Zhang 2006).

The discovery that language in the early school years is predominantly composed of one or two components indicates that it is theoretically possible to construct an effective short-language measure. However, the conflicting findings on the composition of language warrants further investigation. If grammar and semantics are indeed distinct, then the measure must include certain tasks that allow examiners to assess and mark these components separately (such as a grammatical marker and a vocabulary proficiency task). Children would be considered at risk for DLD if they obtained low scores on either of these component tasks. However, if language is predominantly composed of one component, then children would be considered at risk based on their combined results from included tasks.

The second step in the development of a short measure is to determine what language tasks it should include. The literature on the effectiveness of language tasks has focused on comparing the accuracy of single tasks (Pawlowska 2014). Measured alone, single tasks are at best moderately accurate at distinguishing LL. Combining tasks, however, has been shown to increase the accuracy of short-language measures (Conti-Ramsden *et al.* 2001, Poll *et al.* 2010). Yet, only a few studies have investigated this, and their findings are limited to the use of small, clinically referred samples and the inclusion of only a few types of language tasks (Archibald and Joanisse 2009, Conti-Ramsden *et al.* 2001, Poll *et al.* 2010).

An additional caveat of these studies is that the evaluation of language task combinations has only focused on accuracy. Accuracy calculations determine whether a measure is effective at distinguishing LL. Yet, accuracy calculations are complicated by lack of consensus on the criteria and cut-off points used to categorize LL. Furthermore, accuracy calculations do not provide any information about the capacity of a measure to investigate language ability (and distinguish the range of language skills).

A more comprehensive method to investigate the function of a short-language measure is to evaluate its agreement, or the closeness between scores, on two separate measures. Measuring agreement quantifies a language measure's capacity to differentiate children's abilities across the entire continuum of language skills. This is an important consideration if a short measure is to be used to investigate language.

<B>*Aims*

The aim of the current study was to understand the number of components and language tasks that should be included in a short-language measure for children in the early school years. Specifically, the study aimed to determine the combination of language tasks (1) with the greatest agreement with an omnibus language measure; and (2) with the highest accuracy to distinguish children with LL from those with normal-range language skills.

<A>**Materials and methods**

<B>*Sample*

The participants were drawn from the Early Language in Victoria Study (ELVS), a longitudinal study investigating language development (Reilly *et al.* 2006). The study began in 2002, recruiting a community-based sample of 1910 infants, aged between 7.5 and 10 months. Infants were recruited from six of 31 local government areas (LGAs) in metropolitan Melbourne, in the Australian state of Victoria, two from each of the three tiers in the Socio-Economic Indexes of Areas (SEIFA) (Australian

Bureau of Statistics (ABS) 2001). SEIFA is a census-based Australian Index for Relative Socio-Economic Disadvantage. All the children, aged between 7.5 and 10 months, living in the six LGAs were invited to participate through the Victorian Government Maternal and Child Health Program in Melbourne. Infants who had a known condition affecting their development were excluded. Parent questionnaire data were collected annually and the direct assessment of language was conducted at ages 4, 5, 7 and 11 years. Eligible participants for the current study were all those who received direct language assessments at ages 5 (*n* = 995) and/or 7 years (*n* = 1217). Figure 1 depicts participant retention across the first eight waves of the study and the number of participants who provided data at waves six (age 5) and eight (age 7).

<fig 1>

<B>*Procedure*

Child and family characteristics data were collected at recruitment via parent questionnaire. Parents were asked to report on factors such as whether their child was born prematurely (before 36 weeks' gestation) and whether any of their immediate family members had a speech or language-learning difficulty. Participants from non-English-speaking backgrounds were defined as families who reported mainly speaking a language other than English to their child. The socioeconomic status of each participant was measured using the SEIFA index, collected during the 2001 Census (ABS 2001).

Participating children were individually assessed by trained research assistants during a single session. Children were assessed at their local community child health centres (at age 5) and schools (at age 7). If requested, assessments were also conducted at participants' homes. Before conducting the assessments, assessors completed a one-day training course. They were also observed by an experienced assessor for a minimum of four assessment sessions to ensure consistent assessment administration.

<B>*Measures and language tasks*

The language measures administered were the Clinical Evaluation of Language Fundamentals—4th Edition (CELF-4) (Semel *et al.* 2006), the Children's Test of Non Word Repetition (CNrep) (Gathercole and Baddeley 1996) and the Peabody Picture Vocabulary Test—3rd Edition (PPVT-III) (Dunn and Dunn 1997). Children were administered all three measures at age 5. Owing to time constraints, only two measures, the CELF-4 and PPVT-III, were administered at age 7. To reduce the assessment time further, the administration of the PPVT-III was discontinued at age 7 following the assessment of 402 children. Raw total scores for the CELF-4, CNrep and PPVT-III were calculated by the trained assessors using the measures' manuals.

Six subtests from the CELF-4 were administered. The CELF-4 composite scores: core language score (CLS), expressive language score (ELS) and receptive language score (RLS), were calculated by summing age-specific scaled subtest scores, as per the directions in the manual (Semel *et al.* 2006). See table 1 for the compositions of these scores and the measure descriptions.

The CELF-4 composite and subtest scores and total CNrep and PPVT-III scores were normed to have a mean of 100 and standard deviation (SD) of 15 for the 5- and 7-year-old ELVS data by calculating and rescaling *z*-scores for the ELVS cohort at these ages. Normative scores within the ELVS cohort were used to reflect a community sample of Australian children at respective ages in preference to published norms in the CELF-4 manual as the ELVS sample was considerably larger. The distributions of the raw composite scores for the ELVS sample at ages 5 and 7 were similar to those reported in the CELF-4 manual for the sample of children used to derive normative distributions.

The CELF-4 subtest scores and CNrep and PPVT-III total scores were treated as eight independent language task scores. These scores were also combined by adding their individual scores and rescaling them to have a mean = 100 and SD = 15. Subsequently, all scores had a mean = 100 and SD = 15.

<tab 1>

<B>*Defining language ability and low language (LL)*

The three CELF-4 composite scores were used to determine language ability and define LL. Two definitions of LL were investigated to explore comprehensively the assessment of language. In the first, LL (1) was defined as performance of less than or equal to 1.25 SD (i.e., ≤ 81) below the mean on the CLS. In the second definition, LL (2) was defined as performance of ≤ 1.25 SD below the mean on the RLS and/or ELS.

<B>*Statistical analysis*

All analyses were conducted in Stata 14.1 (StataCorp 2015). Three separate analyses were conducted: principal component analysis (PCA), an agreement analysis and receiver operating characteristic (ROC) curve analysis. Participants without raw scores for a given analysis were excluded and a substantially smaller number of participants had the PPVT-III administered (n = 402) at age 7. Consequently, the number of participants included in each analysis differed (n = 899–986 at age 5, n = 399–1211 at age 7).

<B>*Principal component analysis (PCA)*

PCA was performed to explore the number of components to include in a short-language measure. It assesses a set of test scores for intercorrelations. Scores that are correlated are considered to be related, forming a component. Each component can explain a certain amount of variance in the data; this corresponds to a component's eigenvalue. Principal components with small eigenvalues may simply represent random clustering and, therefore, it is usually only recommended to retain components with eigenvalues > 1.

Before performing PCA, the sampling adequacy (suitability for PCA) of both data sets was assessed. The Kaiser–Meyer–Olkin (KMO) values were 0.88 and 0.89 for 5- and 7-year-old data respectively, exceeding the recommended value of 0.6 (Kaiser 1974). Bartlett's test of sphericity also reached statistical significance (< 0.001), indicating the suitability of the data for PCA.

<B>*Agreement analysis*

An agreement analysis was conducted to determine the combination of language tasks with the greatest agreement with the CELF-4 composite scores using the Bland and Altman (BA) method. The BA method plots the score differences of two measures against their means (Bland and Altman 1986). The plot includes reference lines at the mean difference (average difference between measures) and the limits of agreement (mean difference ± 2 × SD of the differences) in which approximately 95% of data will fall. A narrower limit of agreement indicates stronger agreement between measures. Visual examination of the plot can also reveal any evidence of systematic variation in agreement across scores.

Pearson's correlation coefficients were also calculated even though correlation is not a true measure of agreement but of the average linear association across a range of scores. It can be misleading if two measures only agree for a certain range of scores, but provides a useful, easily interpreted summary value.

<B>*Receiver operating characteristic (ROC) curve analysis*

ROC curve analysis was conducted to determine which combination of language tasks is most accurate at distinguishing children with LL. ROC curves for individual language tasks and combinations of tasks were plotted using two LL definitions. ROC curve analysis plots sensitivity over 1 – specificity for a range of diagnostic cut-off points. The area under the curve (AUC) represents the probability that a child will be categorized correctly by the task(s) under investigation. The higher the AUC, the more accurate a task is at defining LL.

<B>**Results**

Table 2 shows the characteristics of the original ELVS sample and the participants who took part in this study. The characteristics of the children who completed the PPVT-III at age 7 years are also included. As illustrated, the participants' mothers had significantly higher education levels and significantly fewer were from a non-English speaking background. The prevalence of LL in both the 5- and 7-year-old separate cohorts was 10% based on LL definition 1 (≤ 1.25 SD below the mean on the CLS) and 15% based on the LL definition 2 (≤ 1.25 SD below the mean on the RLS and/or ELS). These were calculated separately for each age group by including all eligible participants.

<tab 2>

<B>*Components of language*

The scree plots for the 5- and 7-year-old data, generated by the PCA, are displayed in figure 2. They show the eigenvalues corresponding to each extracted component. The first component, which explains the greatest amount of variance across scores, had an eigenvalue of 4.2 at 5 years and 4.0 at 7 years. This component explained 52% of the variation in all scores at 5 years and 57% of variation in scores at 7 years. The extent to which a language task contributes to the measurement of a component can be expressed by its loading value. The higher the loading value of a task, the more that task contributes to the measurement of that component. All language tasks loaded into the first component, indicating that all tasks contributed to its measurement (supporting its representation as a general language ability component). The loading values of tasks ranged from 0.41 to 0.21. The highest loading was from recalling sentences (RS = 0.41) and concepts and following directions (CFD) (0.40) at both ages.

The first component accounted for considerably greater variance in all scores than the next largest component, which had an eigenvalue of 0.9 at 5 years and 0.7 at 7 years. This second component accounted for 11% of variation in all scores at both ages. Word classes (WC) was the only language task that positively loaded onto this component (> 0.2). Conversely, the PPVT-III (another measure of receptive vocabulary) negatively loaded onto this component, indicating that the component was task specific and measured an ability that was specific to WC, such as a higher-level function that is involved in the recognition of semantic relationships, rather than the broader measurement of semantics.

The remaining components each accounted for less than 10% of the variation in scores at both ages. These components appeared to be task specific as only one language task significantly loaded onto each remaining component. There was no evidence of separate grammar, semantic, expressive or receptive language components as specific tasks that measure these aspects of language did not load into the same components. For example, tasks that measure expressive language (word structure (WS), RS, formulated sentences (FS) and the CNrep) and grammar (WS,

sentence structure (SS), FS and, to some extent, RS) did not load onto the same separate components.

<fig 2>

<B>*Agreement between language tasks and CELF-4 composite scores*

BA plots were generated for all language tasks and CELF-4 composite scores. As the BA method requires measures to be on the same scale, all language tasks were scaled to have a mean = 100 and SD = 15. A summary of the limits of agreement and correlation coefficients between language tasks and CELF-4 composite scores are displayed in table 3. The language tasks with the narrowest limits of agreement and highest correlation coefficients have the greatest agreement with a CELF-4 composite score. Therefore, the language task with the greatest agreement with the CLS and ELS was the RS at ages 5 and 7. In contrast, the language task with the greatest agreement with the RLS at ages 5 and 7 was the CFD. The limits of agreement between all single-language tasks and the CELF-4 composite scores were all > 15, indicating that the differences between single tasks and CELF-4 composite scores were > 1 SD in over 5% of cases.

The BA plot for the agreement between the RS and CLS for the 5-year-old cohort is displayed in figure 3. It depicts slight variation in the agreement for the two measures across the range of language scores as there is greater vertical scatter in the middle of the plot. The points on the left are also more closely scattered around the midline. This suggests that the agreement between the RS and CLS is greater for children with LL abilities and weaker for children with language abilities in the middle range. There was also considerable individual variation in the differences in scores between single-language tasks and composite scores, as seen by the range on the vertical axis. There were two cases with a considerably higher CLS. In one case, the participant had a score difference of –47 points, indicating that their CLS was 47 points higher than their RS score. These individuals both had very high scores on all CELF-4 subtests except for RS.

<fig 3>

The combinations of two language tasks with the greatest agreement with the CELF-4 composite scores were CFD + RS, CFD + WS and RS + WS. Their limits of agreement with the CLS ranged from 11 to 12 score points at ages 5 and 7. This suggests that assessing a second task reduced the differences in scores to < 1 SD in 95% of cases.

The BA plot for the agreement between CFD + RS and the CLS for the 5-year-old cohort is displayed in figure 4. The scatter of points is more evenly dispersed and closer to the midline across the distribution of language scores than seen in figure 3. This indicates more uniform agreement across language ability levels when two language tasks are assessed. There was also a reduction in the individual variation in score differences when two language tasks were combined and the limits

of agreement with the CLS reduced by 6 points (age 5) and 7 points (age 7). Assessing a third task further reduced the limits of agreement with the CLS by 2 points (age 5) and 4 points (age 7).

<fig 4, tab 3>

<B>*Accuracy of language tasks*

Table 4 shows the AUCs for each language task and combination of tasks for the 5- and 7-year-old data. The tasks with the greatest AUC for distinguishing LL at both ages were the CFD (sensitivity = 90%, specificity = 87%, accuracy = 88%, at a cut-off point of 1 SD below the mean on the CFD for age 5, LL definition 1) and RS (sensitivity = 84%, specificity = 90%, accuracy = 89%, at a cut-off point of 1 SD below the mean on RS for age 5, LL definition 1).

Combining two language tasks increased the AUC by as much as 4%. The combination of two tasks with the greatest overall AUC was the CFD + RS at both ages. The AUC was just as large for two other combinations (CFD + WS and RS + WS) for LL definition 1 (age 7); however, these combinations had a lower AUC than CFD + RS when LL definition 2 was investigated.

Combining two tasks, CFD + RS, showed the highest accuracy for distinguishing LL (sensitivity = 94%, specificity = 91%, accuracy = 91%, at 1 SD below the mean for age 5, LL definition 1). Adding a third language task, WS, did not increase the AUC when considering LL definition 1 in the 5-year-old cohort (sensitivity = 98%, specificity = 90%, accuracy = 91%, at 1 SD below the mean). It did, however, slightly increase the AUC (by 1%) in the 7-year-old cohort (LL definition 1) and at ages 5 and 7 when considering LL definition 2.

<tab 4>

<A>**Discussion**

The findings from this study indicate that in relation to an omnibus language measure, the greatest agreement and highest accuracy is achieved by combining the CFD and RS. This was true for the 5- and 7-year-old cohorts and both definitions of LL investigated. The results suggest that a short-language measure can indicate a child's overall language ability and determine their need for a full diagnostic assessment.

<B>*Components of language*

The use of PCA suggested the retention of one component of language that was assessed by all language tasks. This indicated that all eight tasks contributed to the assessment of one main

language component that measured general language ability. This component explained 52–57% of the variance in language scores, signifying that much of the variance in scores was not accounted for by this component. Yet, the remaining variance did not appear to be due to distinct semantic or grammar abilities but seemed to be associated with specific tasks.

The results from this study are comparable with previous studies that have explored the composition of language ability (Anthony *et al.* 2014, Colledge *et al.* 2002, Klem *et al.* 2015, Tomblin and Zhang 2006). In accordance with these studies, no independent receptive, expressive, semantic or grammar abilities were identified. This suggests that a short-language measure does not need to assess specific domains or modalities of language independently. The results demonstrate that a short measure should assess one component (general language ability) by combining the scores from different language tasks.

<B>*Combining language tasks*

The language tasks to include in a short-language measure were investigated by measuring their agreement across the full scale of an omnibus measure and accuracy to distinguish LL. Previous studies that have investigated the value of combining tasks have only measured accuracy (Archibald and Joanisse 2009, Conti-Ramsden *et al.* 2001, Poll *et al.* 2010). While accuracy calculations enable one to determine which tasks can differentiate children with LL, they do not give any indication about a task's capacity to differentiate the range of language skills.

The single-language tasks with greatest agreement and highest accuracy in relation to an omnibus language measure were the CFD and RS. Yet, only assessing one of these tasks resulted in inconsistent agreement across the distribution of language scores and suggested that the differences between scores on these tasks and the CELF-4 composites were larger for children with middle-range language abilities. In addition, measuring the CFD or RS alone did not exceed the criterion (of 90% or above) for 'good discriminant accuracy' (Plante and Vance 1994). This indicates that only assessing one language task does not adequately differentiate the language abilities of children, suggesting that a short-language measure should include more than one task.

Combining two language tasks reduced the variation in agreement across language scores. This signifies that assessing two tasks is more effective at differentiating between the language skills of children. The two task combinations with the greatest agreement with the CELF-4 composite scores were combinations containing the CFD, RS and WS. Whereas, the combination of two tasks with the highest accuracy for LL discrimination was CFD + RS. This combination exceeded the criterion (of 90% or above) for 'good discriminant accuracy' (Plante and Vance 1994). These two tasks also had the highest loading onto the retained component in PCA, indicating that they provided the highest contribution to the measurement of general language ability.

While no known studies have compared the accuracy of a direction following task, the RS has previously been determined to have the highest accuracy for LL discrimination in comparison with non-word repetition and grammatical tasks (Archibald and Joanisse 2009, Conti-Ramsden *et al.* 2001, Poll *et al.* 2010). The effectiveness of the CFD and RS as a task combination may be explained by the collective range of knowledge and cognitive functions that these tasks assess. The RS measures linguistic knowledge such as morphosyntax, lexical phonology and, to a lesser extent, semantics (Polisenska *et al.* 2015). It can also draw on aspects of short-term and working memory, which is an underlying impairment in children with persistent language difficulties (Petruccelli *et al.* 2012). The CFD assesses semantics, specifically the understanding of instructional concepts, and draws on working memory and visual object recognition. Together these tasks are considered to assess a wider range of skills than is measured by the other two task combinations.

Assessing a third language task (WS) slightly increased the agreement with an omnibus measure (by 3–4 score points). It also increased the accuracy for LL discrimination by as much as 1%. Yet, at age 5 (considering LL definition 1), there was no increase in accuracy when assessing three tasks. It is therefore perceived that the benefits for including a third task in a short measure are not substantial enough to warrant the additional assessment time.

Assessing two or three language tasks still resulted in considerable individual variation in the differences between scores on task combinations and CELF-4 composite scores. This suggests that scores on a short-language measure and omnibus measure may considerably vary and that some children who have age-appropriate language skills may be considered at risk for DLD when assessed with a short measure, or vice versa. Therefore, the results from two language tasks cannot substitute the use of an omnibus measure when significant decisions are being made about a child's access to therapy. Assessing two language tasks can, however, provide a feasible option to evaluate language ability in large-scale research studies and potentially be used in schools to identify children who need referral to a speech–language therapist for further assessment.

The sensitivity and specificity of assessing the CFD and RS exceeded values previously reported for available short-language measures (Law at al. 2000). Short-language measures often have issues around sensitivity, indicating that it is easier to determine when children have age-appropriate language skills than it is to identify LL. Available short measures can have specificity levels of over 90%, yet sensitivity levels as low as 17% (Law at al. 2000). The current study showed that it is possible to maximize sensitivity without drastically affecting the specificity levels of a short measure. Findings therefore suggest that it is possible to construct a short-language measure that would have sufficient accuracy to be useful in a practical setting.

<B>*Limitations*

The current study contains several limitations. Firstly, the initial investigation into the components of language was limited to PCA. Confirmatory factor analysis was not conducted to verify the structure of language ability and discourse and pragmatic language tasks were not included in this analysis.

Secondly, the language ability and categorization of the LL was based on participants' CELF-4 composite scores. Composite score calculations did not include measures of vocabulary or non-word repetition. This could explain the relatively low effectiveness of these tasks. Nevertheless, these two tasks had higher accuracies than one subtest of the CELF-4 (WC). In addition, it could be argued that if a short-language measure were designed to assess general language ability, a measure including the CFD and RS would be most effective given that these tasks contributed the highest loading to this component in the PCA.

Thirdly, the current analysis was largely dependent on the quality of the specific measures investigated. As the CELF-4 was used as a reference measure, the accuracy of the LL subgroup and validity of the language ability of participants were dependent on the accuracy and validity of the CELF-4 assessment. The effectiveness of each language task was also specific to the quality of the measures used to evaluate each task. For this reason, the results cannot reliably be applied to specific task categories, such as other measures that assess direction following or sentence recall. Alternative forms of these tasks will need to be explored in future.

<A>**Summary and conclusions**

The purpose of this study was to understand the number of components and types of language tasks to include in a short-language measure for children in the early school years. The study analyzed eight language tasks administered to a large-population sample. The findings suggest that it is possible to construct an accurate short-language measure to assess language ability in the early school years by including a sentence-recall and a direction-following task. Measuring these two tasks had high agreement with an omnibus measure and exceeded the criterion for good discriminant accuracy. A short measure containing these tasks could be used to assess language ability in research and in schools. The next step in making a short-language measure available would be to construct an easily administered and scored version that is independent of the CELF-4. Determining the new measure's effectiveness would then allow for the replication of findings and indicate whether the measure meets acceptable accuracy, validity and reliability standards for practical use. Its effectiveness when administered by education staff will also need to be explored to determine whether it could be delivered by teachers.

<A>**Acknowledgments**

The authors thank the Early Language in Victoria Study (ELVS) team and participating families. They especially thank Professor Richard Dowell and Professor Paul Monagle for reviewing the paper and invaluably supporting Jessica Matov during her PhD candidature; Ms Eileen Cini who assisted with data access; and Professor Edith Bavin for kindly reviewing the paper. The ELVS was funded by the Australian National Health and Medical Research Council (projects grants numbers 237106, 9436958 and 1041947) and the Centre of Research Excellence in Child Language (grant number 1023493). Fiona Mensah and Sheena Reilly received Australian National Health and Medical Research Council Early Career and Career Development Fellowships (FM: grant numbers 1037449 and 1111160) and Practitioner Fellowships (SR: grant number 491210 and 1041892). [AQ2] Jessica Matov was supported through an Australian Government Research Training Program Scholarship. ***Declaration of interest:*** The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

<<t/s Set names in caps and scaps as per usual style>>

<A>**References**

ANTHONY, J. L., DAVIS, C., WILLIAMS, J. M. and ANTHONY, T. I., 2014, Preschoolers' oral language abilities: a multilevel examination of dimensionality. *Learning and Individual Differences*, **35**, 56–61.

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.

ARCHIBALD, L. D. and JOANISSE, M. F., 2009, On the sensitivity and specificity of nonword repetition and sentence recall to language and memory impairments in children. *Journal of Speech, Language and Hearing Research*, **52**, 899–914.

AUSTRALIAN BUREAU OF STATISTICS (ABS), 2001, *Socio-Economic Indexes for Areas* (Canberra, ACT: ABS).

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 and Communication Disorders*, **44**, 600–615.

[AQ3] BISHOP, D. V. M., SNOWLING, M. J., THOMPSON, P. A., GREENHALGH, T. and the CATALISE CONSORTIUM, 2016, CATALISE: a multinational and multidisciplinary Delpi consensus study. Identifying language impairments in children. *PLoS ONE*, **11(7)**.

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 and Psychiatry*, **58**, 1068–1080.

BLAND, J. M. and ALTMAN, D. G., 1986, Statistical methods for assessing agreement between 2 methods of clinical measurement. *Lancet*, ***i***, 307–310.

COLLEDGE, E., BISHOP, D. V. M., KOEPPEN-SCHOMERUS, G., PRICE, T. S., HAPPE, F. G., ELEY, T. C., DALE. P. and PLOMIN, R., 2002, The structure of language abilities at 4 years: a twin study. *Developmental Psychology*, **5**, 749–757.

CONTI-RAMSDEN, G., BOTTING, N. and FARAGHER, B., 2001, Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology and Psychiatry*, **42**, 741–748.

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 and 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.

DUNN, L. and DUNN, L., 1997, *Peabody Picture Vocabulary Test—Third Edition (PPVT-III)* (Sydney, NSW: Psychological Corporation).

GATHERCOLE, S. and BADDELEY, A., 1996, *Children's Test of Non Word Repetition (CNrep)* (Sydney, NSW: Psychological Corporation).

[AQ4] 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*, **51**, 1058–0360.

KAISER, H., 1974, An index of factorial simplicity. *Psychometrika*, **39**, 31–36.

KLEM, M., GUSTAFSSON, J. and HAGTVET, B., 2015, The dimensionality of language ability in four-year-olds: construct validation of a language screening tool. *Scandinavian Journal of Educational Research*, **59**, 195–213.

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.

LONIGAN, C. J. and MILBURN, T. F., 2017, Identifying the dimensionality of oral language skills of children with typical development in preschool through fifth grade. *Journal of Speech, Language and Hearing Research*, **60**, 2185–2198.

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.

[AQ5] 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.

PENTIMONTI, J., O'CONNELL, A., JUSTICE, L. and CAIN, K., 2015, The dimensionality of language ability in young children. *Child Development*, **86**, 1948–1965.

PETRUCCELLI, N., BAVIN, E. L. and BRETHERTON, L., 2012, Children with specific language impairment and resolved late talkers: working memory profiles at 5 years. *Journal of Speech, Language and Hearing Research*, **55**, 1690–1703.

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 and 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 and Hearing Research*, **53**, 414–429.

REILLY, S., EADIE, P., BAVIN, E., WAKE, M., PRIOR, M., WILLIAMS, J., BRETHERTON, L., BARRETT, Y. and UKOUMUNNE, O., 2006, Growth of infant communication between 8 and 12 months: a population study. *Journal of Paediatrics and Child Health*, **42**, 764–770.

SEMEL, E., WIIG, E. H. and SECORD, W., 2004, *Clinical Evaluation of Language Fundamentals Screening Test* (San Antonio, TX: PsychCorp).

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

STARK, H. L., SNOW, P. C., EADIE, P. A. and GOLDFELD, S. R., 2016, Language and reading instruction in early years' classrooms: the knowledge and self-rated ability of Australian teachers. *Annals of Dyslexia*, **66**, 28–54.

StataCorp, 2015, *Stata Statistical Software: Release 10* (College Station, TX: StataCorp LP).

TOMBLIN, J. B. and ZHANG, X., 2006, The dimensionality of language ability in school-age children. *Journal of Speech, Language and Hearing Research*, **49**, 1193–1208.

WHITEHOUSE, A. J. O., WATT, H. J., LINE, E. A. and BISHOP, D. V. M, 2009, Adult psychosocial outcomes of children with specific language impairment, pragmatic language impairment and autism. *International Journal of Language and Communication Disorders*, **44**, 511–528.

Table 1. Measures, language tasks administered at 5 and 7 years

Assessment	Subtest/language task	Subtest or assessment description	Derived scores		
			CLS	RLS	ELS
CELF-4[a,b]	Concepts and following directions (CFD)	Evaluates the ability to interpret and follow spoken directions	*	*	
	Word structure (WS)	Evaluates the ability to use pronouns and apply word structure rules	*		*
	Recalling sentences (RS)	Evaluates the ability to repeat spoken sentences	*		*
	Formulated sentences (FS)	Evaluates the ability to formulate complete and meaningful sentences	*		*
	Sentence structure (SS)	Evaluates the ability to interpret sentences		*	
	Word classes (WC)	Evaluates the ability to understand relationships between words		*	
CNrep[a]	Non-word repetition	Evaluates the ability to repeat non-words	Total score		
PPVT-III[a,b]	Receptive vocabulary	Evaluates the ability to match words with pictures	Total score		

Notes: [a]Data collected for the 5-year cohort.

[b]Data collected for the 7-year cohort.

*Inclusion of the subset to form the CELF-4 composite score; CLS, core language score; RLS, receptive language score; ELS, expressive language score.

Table 2. Summary of the participant characteristics

Characteristic	Original sample (baseline)	Age 5 cohort	*p*-value	Age 7 cohort	*p*-value	PPVT at 7 years	*p*-value
Included participants (*n* =)	1910	995	–	1217	–	402	–
Age, mean (SD)	8.5 (0.6) months	5.2 (0.1)	–	7.4 (0.2)	–	7.2 (0.1)	–
Male, *n* (%)	965 (50.5)	488 (49)	0.17	595 (48.9)	0.05	184 (45.8)	0.03
Premature birth (< 36 weeks), *n* (%)	59 (3.1)	31 (3.1)	0.94	40 (3.3)	0.50	20 (4.9)	0.01
Maternal education, *n* (%)[b]			< 0.01		< 0.01		0.13
Completed Year 10 or less	179 (9.4)	69 (6.9)[a]		86 (7.1)[a]		26 (6.5)	
Completed Year 12	765 (40.2)	379 (38.2)[a]		478 (39.4)[a]		160 (40.0)	
Completed a university degree	693 (36.4)	422 (42.5)[a]		488 (40.2)[a]		153 (38.3)	
Family history of speech/language-learning difficulties, *n* (%)	475 (24.9)	248 (24.9)	0.95	296 (24.3)	0.46	90 (22.4)	0.19
Non-English speaking background (NESB) (%)	126 (6.6)	38 (3.8)[a]	< 0.01	47 (3.9)[a]	< 0.01	14 (3.5)[a]	< 0.01
SEIFA Index of Disadvantage, mean (SD)[c]	1036 (60)	1042 (53)	0.95	1040 (56)	0.46	1036 (58)	0.19

Notes: [a]Significant difference between the included group and the excluded group at $p < 0.01$.

[b]Years 10–12 are the last three years of school in Australia.

[c]Socio-Economic Index for Areas (SEIFA) Disadvantage. The lower the index score, the more disadvantaged the sample (Australian population mean = 1000, standard deviation (SD) = 100).

Table 3. Limits of agreement and Pearson correlation coefficients for agreement analysis between language tasks and CELF-4 composite scores for 5- and 7-year-old data

	Limit of agreement						Pearson correlation coefficient					
	5 years			7 years			5 years			7 years		
	CLS	RLS	ELS	CLS	RLS	ELS	CLS	RLS	ELS	CLS	RLS	ELS
CFD	(–17.3, 17.5)	(–19.1, 19.2)	(–23.0, 23.3)	(–18.0, 17.9)	(–19.9, 19.8)	(–23.9, 23.9)	0.83*	0.79*	0.69*	0.82*	0.77*	0.68*
WS	(–17.6, 17.9)	(–27.8, 27.7)	(–17.0, 17.2)	(–18.5, 18.5)	(–28.6, 28.5)	(–17.2, 17.3)	0.82*	0.57*	0.83*	0.80*	0.54*	0.83*
RS	(–17.2, 17.4)	(–26.9, 27.0)	(–17.2, 17.3)	(–15.6, 15.6)	(–27.3, 27.3)	(15.2, 15.2)	0.83*	0.59*	0.83*	0.86*	0.58*	0.87*
FS	(–19.4, 19.4)	(–29.1, 29.1)	(–18.3,18.4)	(–19.0, 19.0)	(–28.2, 28.2)	(–17.6, 17.6)	0.79*	0.52*	0.83*	0.79*	0.55*	0.82*
WC	(–35.2, 35.2)	(–22.7, 22.8)	(–35.8, 35.9)	(–33.1, 33.1)	(–22.0, 21.9)	(–33.6, 33.6)	0.31*	0.71*	0.28*	0.38*	0.72*	0.36*
SS	(–26.7, 26.8)	(–20.2, 20.2)	(–28.1, 28.2)	(–28.4, 28.4)	(–19.6, 19.5)	(–29.5, 29.4)	0.60*	0.77*	0.55*	0.54*	0.78*	0.51*
PPVT-III	(–27.9, 27.7)	(–27.9, 27.	(–29.2, 28.9)	(–27.3, 28.7)	(–30.7, 29.	(–27.1, 29.4)	0.56*	0.56*	0.52*	0.57*	0.50*	0.56*

		6)			7)							
CNrep	(–28.8, 27.6)	(–32.9, 31.9)	(–29.0, 27.9)	–	–	–	0.54*	0.40*	0.53*	–	–	–
CFD + RS	(–12.3,12.5)	(–20.6, 20.7)	(–16.7, 16.9)	(–11.2,11.1)	(–21.0, 21.0)	(–15.9,15.9)	0.91*	0.76*	0.84*	0.93*	0.75*	0.85*
CFD + WS	(–11.7, 11.9)	(–20.7, 20.7)	(–15.9, 16.1)	(–11.9, 11.9)	(–21.1, 21.1)	(–16.0, 16.0)	0.92*	0.76*	0.85*	0.92*	0.74*	0.85*
RS + WS	(–12.8, 12.9)	(–25.4, 25.4)	(–12.3, 12.4)	(–11.1, 11.1)	(–25.7, 25.7)	(–9.5, 9.5)	0.90*	0.64*	0.94*	0.93*	0.63*	0.94*
CFD + WS + RS	(–9.4, 9.5)	(–21.2, 21.2)	(–13.0, 13.1)	(–7.6, 7.6)	(–21.4, 21.4)	(–11.4, 11.4)	0.94*	0.74*	0.90*	0.96*	0.74*	0.92*

Note:*Significant at $p < 0.01$. Age 5 years: *N* for CELF-4 subtests = 980–986; *N* for CNrep = 934–935; *N* for PPVT-III = 925–926. Age 7: *N* for CELF-4 subtests = 1208–1211; *N* for PPVT-III = 398; CLS, core language score; RLS, receptive language score; ELS, expressive language score; CFD, concepts and following directions; WS, word structure; RS, recalling sentences; FS, formulated sentences; WC, word classes; SS, sentence structure.

Table 4. Areas under receiver operating characteristic (ROC) curves for language tasks and task combinations for 5- and 7-year-old data

Language task/subtest	LL definition 1				LL definition 2			
	Age 5 cohort		Age 7 cohort		Age 5 cohort		Age 7 cohort	
	AUC	*N*	AUC	*N*	AUC	*N*	AUC	*N*
Concepts and following directions (CFD)	0.95	981	0.94	1206	0.91	980	0.88	1202
Word structure (WS)	0.94	981	0.93	1206	0.89	980	0.85	1202
Recalling sentences (RS)	0.95	981	0.96	1206	0.89	980	0.88	1202
Formulated sentences (FS)	0.88	981	0.90	1206	0.83	980	0.85	1202
Word classes (WC)	0.66	980	0.76	1206	0.76	980	0.86	1202
Sentence structure (SS)	0.81	980	0.85	1202	0.85	980	0.85	1202
CNrep	0.83	934	–	–	0.77	933	–	–
PPVT-III	0.83	925	0.80	399	0.83	924	0.83	399
CFD + RS	0.98	981	0.98	1206	0.94	980	0.92	1202
CFD + WS	0.97	981	0.98	1206	0.94	980	0.91	1202
RS + WS	0.97	981	0.98	1206	0.92	980	0.90	1202
CFD + RS + WS	0.98	891	0.99	1206	0.95	980	0.93	1202

Note: AUC, area under the curve; LL definition 1 = ≤ 1.25 SD below the mean on the core language score (CLS) of the CELF-4; LL definition 2 = ≤ 1.25 SD below the mean on receptive and/or expressive language scores (ELS) of the CELF-4.

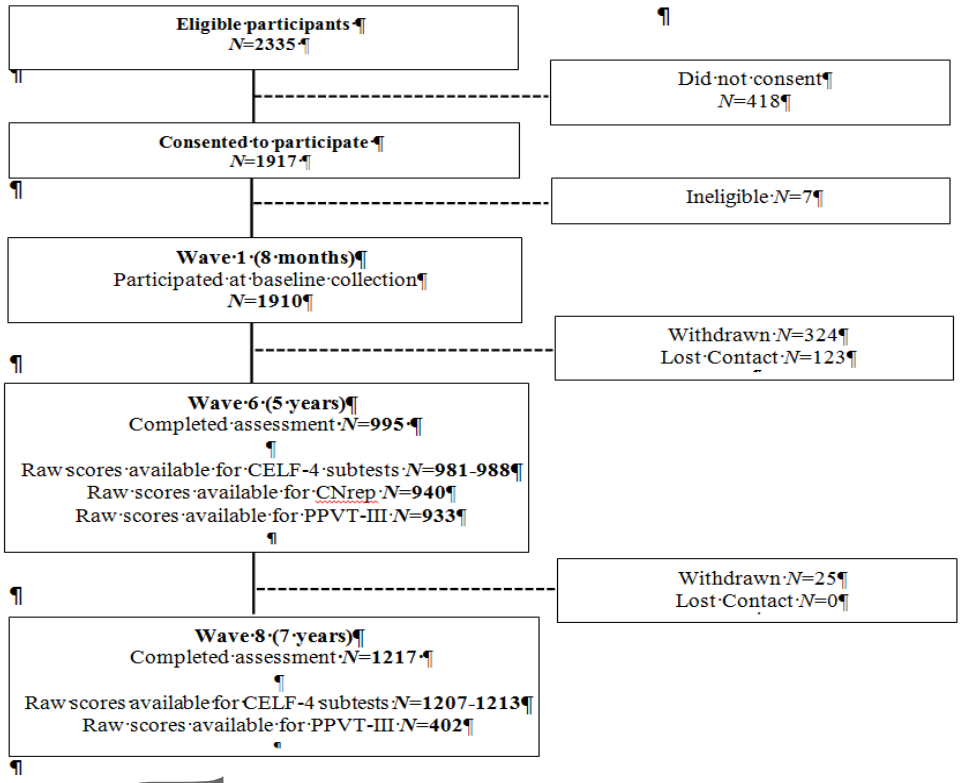

Figure 1. Flowchart showing Early Language in Victoria Study (ELVS) participation from baseline (8 months) to 7 years. CELF-4 = Clinical Evaluation of Language Fundamentals—4th edition; CNrep = Children's Test of Non Word Repetition; PPVT-III = Peabody Picture Vocabulary Test—3rd edition

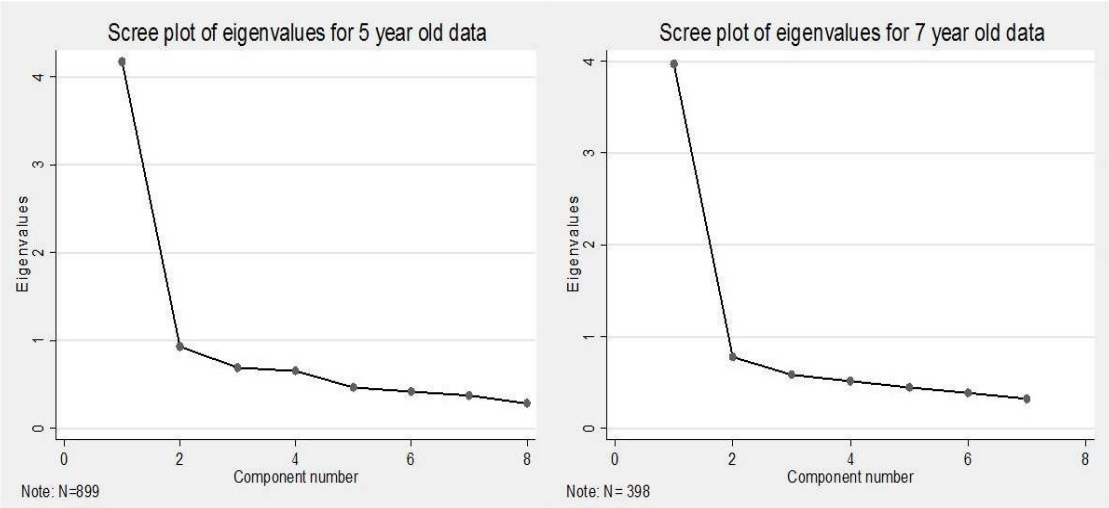

Figure 2. Scree plots for principal component analysis (PCA), depicting eigenvalues corresponding to extracted components.

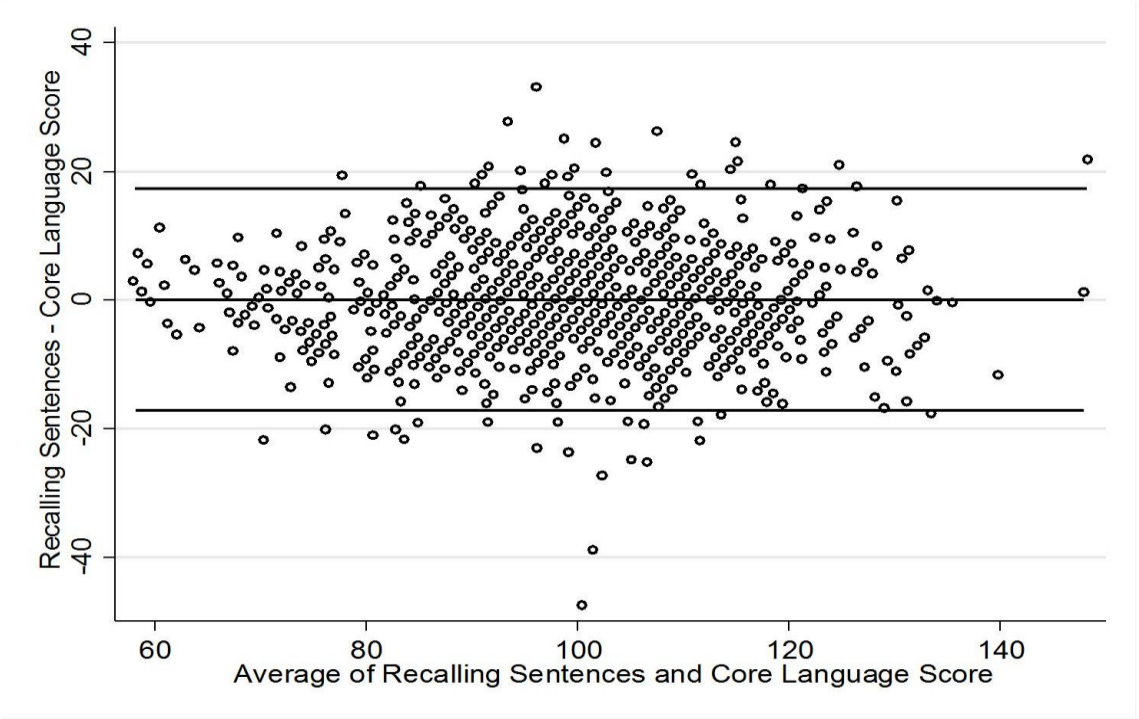

Figure 3. Bland and Altman plot for agreement between recalling sentences (RS) and the core language score (CLS) at age 5 years.

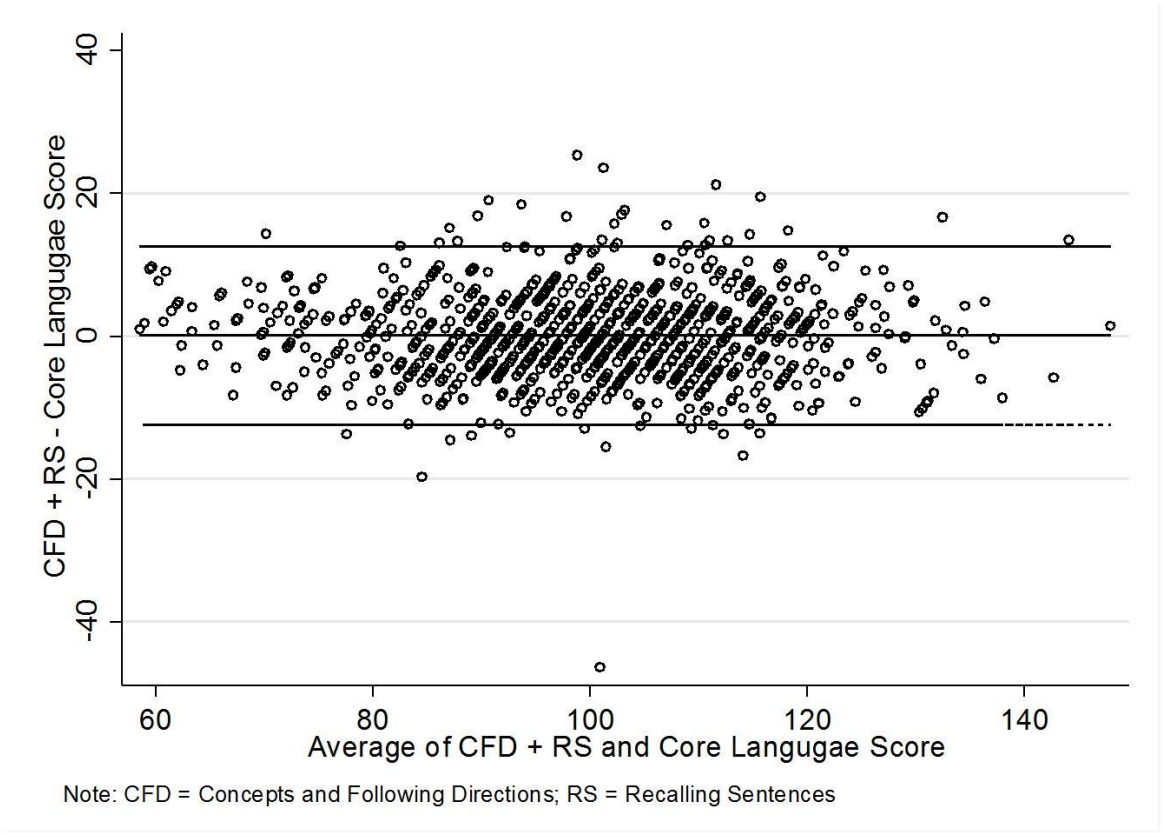

Figure 4. Bland and Altman plot for agreement between concepts and following directions (CFD) and recalling sentences (RS) and the core language score (CLS) at age 5 years.