Scandinavian Journal of Educational Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/csje20

Measuring Listening Comprehension and Predicting Language Development in At-Risk Preschoolers

Åste Mjelve Hagen, Rebecca Knoph, Hanne Næss Hjetland, Kristin Rogde, Joshua Fahey Lawrence, Arne Lervåg & Monica Melby-Lervåg

To cite this article: Åste Mjelve Hagen, Rebecca Knoph, Hanne Næss Hjetland, Kristin Rogde, Joshua Fahey Lawrence, Arne Lervåg & Monica Melby-Lervåg (2021): Measuring Listening Comprehension and Predicting Language Development in At-Risk Preschoolers, Scandinavian Journal of Educational Research, DOI: 10.1080/00313831.2021.1939136

To link to this article: https://doi.org/10.1080/00313831.2021.1939136

Published online: 17 Jun 2021.

SCANDINAVIAN JOURNAL OF EDUCATIONAL RESEARCH
https://doi.org/10.1080/00313831.2021.1939136

Routledge
Taylor & Francis Group

OPEN ACCESS

Measuring Listening Comprehension and Predicting Language Development in At-Risk Preschoolers

Åste Mjelve Hagen[a], Rebecca Knoph[b], Hanne Næss Hjetland[a], Kristin Rogde[b], Joshua Fahey Lawrence[b], Arne Lervåg[b] and Monica Melby-Lervåg[a]

[a]Department of Special needs education, University of Oslo, Oslo, Norway; [b]Department of education, University of Oslo, Oslo, Norway

ABSTRACT

Listening comprehension involves the ability to understand and extract meaning from spoken sentences, stories, and instruction. This skill is vital for young children and has long-term effects on school achievement, employability, income, and participation in society. There is a lack of measures of young children's listening comprehension skills. We report on a new measure of listening comprehension (LURI) that we tested in a group of at-risk preschoolers. Using IRT (Item Response Theory) analysis, we examined the psychometric properties of the instrument. Moreover, in a series of regressions, we found that the LURI measure predicted a range of other language skills better than standardized measures. Thus, the LURI test is a reliable and valid measure of listening comprehension. Assessing listening comprehension is a time-effective way of measuring a skill critically important to language development and represents the product of a range of different oral language processes.

ARTICLE HISTORY
Received 9 June 2020
Accepted 8 May 2021

KEYWORDS
Language development; listening comprehension; vocabulary; preschool children; validating language measure; at-risk for language problems

Introduction

Oral language skills are implicated in how children learn from adults and interact with peers, and they are vital for the development of reading skills. Research has shown that the developmental trajectories of language skills are quite stable so that the children with weaker skills at school entry tend to continue to struggle (e.g., Bornstein et al., 2016; Klem et al., 2016; Storch & Whitehurst, 2002). Children with weaker baseline language skills tend to have the same or slower growth than children with stronger language skills. These children are at high risk of developing further language problems and, also, difficulties with reading. Reading problems can, in turn, affect other academic outcomes, and many of these children will also experience social-emotional difficulties (Yew & O'Kearney, 2013).

Fortunately, several intervention studies indicate that it is possible to help these children (Fricke et al., 2013; Fricke et al., 2017; Hagen et al., 2017; Haley et al., 2017). For instance, Fricke et al. (2017) tested the effect of an oral language program for preschool children (in the UK: in Nursery and Reception class), which aimed to improve children's vocabulary, develop narrative skills, encourage active listening, and build confidence in independent speaking. Children in the intervention group participated in either a 20-week or a 30-week program where they had a combination of small-

CONTACT Åste Mjelve Hagen ammhage@isp.uio.no Department of Special needs education, University of Oslo, Oslo, Norway

groups and individual sessions each week. They found no significant differences between the 20 and 30-week intervention, but small to medium effects on measures of oral language skills for children in both intervention groups.

Similar results were found in a cluster randomized trial in 148 preschool classrooms, with an oral language intervention that lasted 1 year and 1 month, with five blocks of 6 weeks and intervention three times per week (Hagen et al., 2017). Immediately after the intervention, there were moderate effects on both measures of trained words and standardized measures. At delayed follow-up (7 months after the intervention), these positive effects remained for the distal measures.

These studies indicate that by providing systematic and intensive language support, we can boost children's language development and equip them to meet school's academic demands. However, the successful implementation of these preschool programs requires that preschools identify children in need of targeted support. This paper focus on a listening comprehension test and investigates if this short test of integrated oral skills might be a valid and reliable tool that preschool teachers could use to identify children at risk for later difficulties. We wanted to test the psychometric properties and predictive validity of such a language measure for preschool children at the lower end of the skill distribution.

Dimensionality in Oral Language and Listening Skills

Although oral language consists of vocabulary, grammar, and narrative skills, younger children's oral language skills are best understood as unidimensional (Language and Reading Research Consortium, 2015). The Language and Reading Research Consortium (2015) examined the dimensionality of language ability in a sample of 915 children from four to eight years old. They found that for the youngest children, a unidimensional model of language fitted the data best. They identified distinct factors of vocabulary, grammar, and discourse skills for older children. These results are in line with a study by Tomblin and Zhang (2006), where they found that correlations between vocabulary and grammar were high but decreasing with age from r = .94 in kindergarten to r = .78 by the time children were in eighth grade. Even though both studies found evidence of emergent dimensionality in later years, language skills in preschool children were unidimensional. Similar results are obtained in other studies (Bornstein et al., 2016; Klem et al., 2016). These results informed our hypothesis that an integrated assessment of these skills may have good psychometric properties and be useful in identifying students who need additional support.

One crucial part of children's oral language skills is listening comprehension. Listening comprehension involves the ability to process, integrate, and understand the meaning of information or text when it is heard (Hogan et al., 2014). Listening comprehension is required to extract meaning from oral communication and is also critical for social well-being and functioning. It is also clear that listening is a complex process that requires both language and cognition; vocabulary, inferencing, background knowledge, as well as working memory and attention are all working together when individuals are trying to make sense of spoken language (Alonzo et al., 2016; Cain et al., 2001; Daneman & Merikle, 1996; Florit et al., 2009; Florit et al., 2013).

Because of its multifaceted nature, it has been suggested that listening comprehension serves as an index of the different oral language skills that constitute a latent construct (Hogan et al., 2014; Kim & Phillips, 2014). This perspective has empirical support: in one study, listening comprehension was almost perfectly predicted by vocabulary, word definitions, grammar, inference skills, and verbal working memory (Lervåg et al., 2018). Another research team found that listening comprehension and other oral language skills (grammar, vocabulary, inference skills) were overlapping to such a degree that they were best understood as a unitary construct (LARRC 2017). Thus, listening comprehension can be understood as an index of other oral language skills. A test of the predictive validity of such a measure in a sample of students at the lower end of the language ability distribution will add to our understand of how early listening relates to the development of other discrete skills.

The rank order in children's oral language skills within a cohort is highly stable over time. Identifying children with language problems at an early age may make it possible to provide targeted support for some children and reduce the risk they will experience reading comprehension problems. In line with the simple view of reading, there are studies which suggest that oral language interventions targeting vocabulary, grammar, listening comprehension, and narrative skills, not only improve oral language but also have transfer effects to reading comprehension (Block & Mangieri, 2006; Brinchman et al., 2016; LARRC, Jiang, & Logan 2019). Unfortunately, the effects of most oral language interventions typically fade out in the months or years after the training (Rogde et al., 2019). Thus, even if the children are detected early, short duration interventions are unlikely to provide sufficient support for students in the long term. The best bet is to give children with oral language problems instructional intervention early and continue to support these children into school. However, providing early intervention and ongoing support requires effective and practical assessments to supplement teacher ratings, which are the *de facto* screening tools used in Norwegian preschools.

Language Difficulties and Detection of Language Problems in Children

Even though most children develop their native language rapidly and are advanced language users by the age of 4, the prevalence of language disorders of unknown origin (not associated with an intellectual disability or medical diagnosis) is estimated to be 7.6% (Norbury et al., 2016). This corresponds to 1–2 children in a typical Scandinavian preschool classroom, although classroom sizes vary. Children who experience poor language skill development are a heterogeneous group. While some experience language delays that resolve with time (i.e., developmental lag), other children experience more persistent language difficulties that impact their long-term outcomes (Bishop & Adams, 1990). Even among lower-skilled children, variability makes it difficult to accurately and consistently determine which students require language support and which can independently make adequate and sustained progress (see McKean et al., 2017).

Results from research into how well teachers can identify children who might need additional support is inconclusive. Cabell et al. (2009) examined how accurately teachers were able to rate the emergent literacy skills of young children (ages 3–5-year-old). The teachers used a simplified 12 item pre-literacy rating scale (Wiig et al., 2003) to make their ratings. The children also completed four language assessments. Results indicated a strong positive correlation between teachers' ratings and student performance (*r* between .49 and .77; r mean = .50), but an overall low sensitivity. That is, despite the strong correlation between ratings and assessment scores, teachers only correctly identified 165 children out of 209 students as meeting the threshold for being "at-risk." These results suggest the teacher's ratings are "somewhat valid representation of children's skills" and that in practice teachers can play a role in helping to identify a pool of children in the lower range of the skill distribution. However, teacher ratings are not accurate enough to precisely identify children who need extra support.

There are, of course, assessments that can be used to precisely and accurate identify children at risk for language and literacy difficulties. However, these tools differ in their objectives, requirements for administration, and quality. In a systematic review of assessments used to identify language difficulties, Denman et al. (2017) surveyed English language assessments for children aged 4–12 years. They only identified ten that met their inclusion criteria: six developed exclusively for children under the age of 8 years and another 4 designed for use across a wide age range, starting in the preschool years. Of these ten assessments, none met all criteria for methodological quality, and only three met criteria for reliability and validity (The Clinical Evaluation of Language Fundamentals [CELF]; The Woodcock Johnson IV Tests of Oral Language and The Test of Narrative Language–Second Edition [TNL-2; Gillam & Pearson, 2017]). However, these three tools are assessments developed for and used by experts to inform diagnostic decisions (e.g., psychologists, speech-language therapists) and required specific training to use.

Furthermore, the Woodcock Johnson and The Test of Narrative Language's listening components are integrated into larger test batteries. To our knowledge, there is no easily administered test of listening comprehension that is suited for teachers and preschool teachers in any language.

There is a similar situation among Scandinavian language assessments. A governmental expert group in Norway concluded that none of the language assessment instruments used in preschool met the basic criteria for documenting validity and reliability (Kunnskapsdepartementet, 2011). Despite this, over 90% of Norwegian preschools report that they use these instruments (Kunnskapsdepartementet, 2011; Rambøll Management, 2008). A systematic review from Sweden revealed few if any measures of early precursors of reading problems (Statens beredning för medicinsk utvärdering [Swedish Council on Health Technology Assessment], 2014). The Norwegian parliament has taken action to address this issue. They have decided that as of April 2020 that all 4-year-olds are entitled to a language evaluation. Still, at this point, there are no psychometrically sound screeners that can be administered at scale in Norway. A valid and reliable process to help identify children in need of extra language support in preschool is sorely needed. Such a process might rely on teacher ratings to identify students who are at the lower range of the language skill distributions and then invoke the administration of a short screening assessment to identify students within that pool who are at risk.

The Current Study

Given the importance of listening comprehension and that it can serve as an index for children's development on different integrated aspects of language, it is rather surprising that there are almost no existing tests to assess it. A recent meta-analysis investigated preschool predictors of reading in school (Hjetland et al., 2017). Among the 63 studies included in the quantitative synthesis, so few reported broader measures like listening comprehension it could not be included as an indicator in their models. Similarly, in a recent systematic review of oral language assessments for young children, Malec et al. (2017) found that research on oral language tests for children aged 4–8 focused on measuring vocabulary, narrative comprehension, and syntax. Although narrative comprehension is related to listening comprehension, it is important to note that the narrative comprehension tasks reported here require children to tell (or retell) stories from pictures. This task does not, therefore, elicit inferential reasoning. Therefore, having a well-validated instrument to examine listening comprehension is an important contribution to clinical work with children and a way of ensuring that young children get suitable interventions.

Thus, the aim of the current study was to examine the psychometric properties of a listening comprehension measure in a group of preschool children at the lower range of the skill distribution, and to examine how the test predicts language development in these children compared to more specific measures of different language skills. These include a standardized measure of receptive vocabulary, measures of grammatical skills (syntax and morphology) and narrative skills. As a control, we also included a measure of nonverbal ability. As guidelines for examination of psychometric properties and interpretation of results, we used the COSMIN checklist for assessing the methodological quality of studies on measurement properties (Mokkink et al., 2010). We also referred to the the EFPA review model for the description and evaluation of psychological and educational tests (European Federation of Psychologists' Associations, 2013).

The study examines the following research questions:

1. What are the psychometric properties of a listening comprehension assessment in terms of dimensionality, item response, and test reliability?
2. In a sample of at-risk preschoolers, how is the predictive validity of a listening comprehension assessment for language development a year later in comparison to more specific measures of language skills?

Method

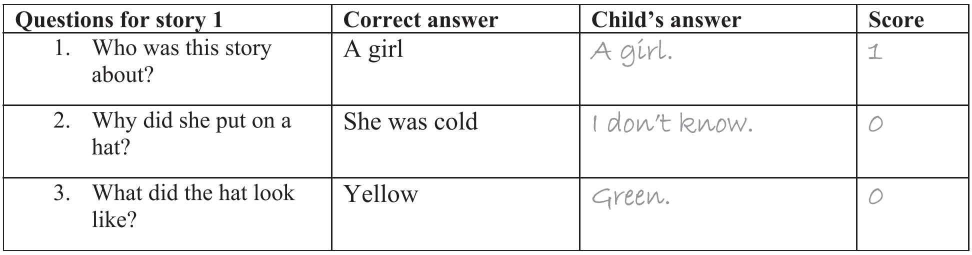
Participants

The participants ($n = 289$) were the 35% lowest scoring on a vocabulary screening measure used in a cohort of children in a longitudinal study (Hagen et al., 2017). The children were selected based on a screening with a measure consisting of 29 items from the British Picture Vocabulary Scale II (BPVS-II; Dunn et al., 1997) and 12 items from the subtest picture naming of the Wechsler Preschool and Primary Scale of Intelligence-III (Wechsler, 1989). The measure was designed to assess children's receptive and expressive vocabulary and took on average 5 min per child. The reliability of the screening measure, Cronbach's alpha, was .67. The decision to use students below the 35th percentile of skill development was pragmatic and driven by our research interests. Research on teacher ratings suggests their ratings of student language and literacy skills are valid but imprecise. Teacher ratings correlate well with assessment measures suggesting they can roughly identify students as low, medium and high skilled. The same body of research shows the teacher cannot identify individual students who are at risk reliably. Given the practical and funding limitations of this study we thought examining student in the bottom third of the distribution would approximate the range of students that could be identified by teachers. This logic also informed our design of the larger study from which we obtained these data. We discuss some implications of this decision in the limitations section below.

The children attended 148 classrooms in 75 different preschools in two municipalities in Norway. Around half of the participants (49.4%) were female, and their mean age was 57.84 months ($SD = 3.39$) when the first wave of testing was conducted. We administered the second wave of testing a year later when the children were in the final year of preschool. The sample was part of a study testing the effect of a language intervention. Here we use only data from the control group ($n = 134$) to answer the second research questions asking about the predictive value of the listening comprehension measure because we cannot be sure how participation in the intervention might influence the relationship between oral language skills at t1 and subsequent skill development.

Measures

The development of the listening comprehension test (the LURI) was inspired by a listening comprehension version of the Neale Analysis of Reading Ability-Revised (Neale et al., 2011). The LURI contains ten short stories followed by three to five questions per story, 36 questions total (see Figure 1 for an excerpt from the test). An adult reads the story and asks the child the related questions, where the child answers each item freely before advancing to the next item. Once the child has responded to all of the questions about a particular story, the adult reads the next story until all stories and questions are read and answered. The stories begin straightforward and short (12 words) but gradually increase in complexity and length (72 words). The questions are a mix of recall and inference. Thus, the test requires that children listen and understand what is being read,



Questions for story 1	Correct answer	Child's answer	Score
1. Who was this story about?	A girl	A girl.	1
2. Why did she put on a hat?	She was cold	I don't know.	0
3. What did the hat look like?	Yellow	Green.	0

Figure 1. Excerpt from the LURI test.

remember the information until questions are answered, and infer information that is not explicitly stated. The measure was named LURI as an acronym for the skills required to get items correct: Listen, Understand, Remember, and Infer. All children were asked all 36 questions. We scored all answers as either correct or incorrect. The test takes 15–20 min in total.

Receptive vocabulary was assessed by using the first 144 words of the BPVS-II (Dunn et al., 1997), Norwegian version (Lyster & Horn, 2009). In this test, the children are asked to choose a picture among four pictures one that matches the word given by the tester. Since the children in our sample were expected to have poor vocabulary knowledge, the assessment operated with custom start and stop criteria, other than what is given in the administration manual. All children started at the lowest level (item 1 in set 1) and were stopped after eight (or more) incorrect responses in two sets in a row. Cronbach's alphas were .92 for t1 and .88 for t2.

Verbal comprehension of syntax was assessed using the Test for Reception of Grammar, version 2, (TROG-2) (Bishop, 2003), Norwegian version by Lyster et al. (2010). Cronbach's alphas were .96 for t1 and .95 for t2.

Grammatical morphology was assessed with the Grammatic Closure subtest of the Illinois Test of Psycholinguistic Abilities (ITPA; Kirk et al., 1968), Norwegian version (Gjessing et al., 1975). Cronbach's alphas for the Grammatic Closure were .76 for t1 and .75 for t2.

Narrative skills were assessed using The Renfrew Bus Story test (Renfrew, 1997). The Renfrew Bus Story test has been translated to Norwegian by researchers at the Department of Special Needs Education at the University of Oslo. In this test, the administrator tells a story orally for the child alongside pictures illustrating the story. When the story has been told, the child is instructed to retell the story with the pictures available. The children's retellings were transcribed verbatim. Scores were given based on the child's use of key/informative words and story structure. Reliability was acceptably high at .76 for t1 and .79 for t2.

Raven's Progressive Matrices test was used to assess *nonverbal IQ* (Raven, 1998). In this test, the child is asked to identify which one out of six alternatives that will successfully complete a visuospatial pattern. Reliability was acceptable at .74 for t1 and .77 for t2.

Procedure

The children were tested individually with the measures; the first time (t1) at the end of their second-to-last year of preschool and the second time (t2) at the end of their last year of preschool. All testing was conducted in the children's school in a separate room without disturbances. The testing lasted a maximum of 45 min per time, and the children had breaks when necessary.

Analysis

To answer the first research question, we developed an analytic plan to asses psychometric properties of the assessment. We fit the 1PL (one parameter logistic) model (which allows the items to have different difficulty levels but restricts items to be equally perceptive in differing ability levels) and the 2PL (two parameter logistic) model (allows each item to have both different difficulty and discrimination parameters). These models were compared to determine the best fit for the data using the R package *mirt* (Chalmers, 2012). Because the scored data are binary, assessing the internal reliability of the LURI test required ordinal alpha.

To answer research question two, we fit a series of multivariate regressions predicting syntactic skills, morphological skills, narrative skills, vocabulary, and nonverbal IQ.

Results

Research Question 1. What are the psychometric properties of a listening comprehension assessment in terms of dimensionality, item response, and test reliability?

Item Response

The 2PL model significantly reduced the residual sum of squares compared to the 1PL model ($\chi^2(35) = 111.87$, $p < .001$), indicating better relative model-data fit. Most fit indices for the 2PL model also indicated adequate model fit (RMSEA = 0.03; SRMSR = 0.06). However, the *M2* statistic did indicate that the expected and observed sum score distributions were significantly different (*M2* (594) = 786.00, $p < .001$). Still, most indices show the 2PL as obtaining adequate model fit, and the M2 statistic is known for being conservative (Maydeu-Olivares & Joe, 2006). Therefore, the 2PL model was selected for further analysis.

All 36 items demonstrated adequate item fit (S-χ2 values for all items were non-significant, $p > .01$; Orlando & Thissen, 2000). Table 1 provides information for each item based on the 2PL model. See Figures 2 and 3 for the item characteristic curves and item information functions grouped by story-question block, respectively.

The total scores on the test were approximately normally distributed ($M = 16.81$, $SD = 6.18$). There is potential for a floor effect, as the observed total scores (0-31) do not represent the range of potential total scores (0-36). However, this is expected, as the test designers sought to avoid a ceiling effect during the post-test. Further analysis of the proportion of children who got each item correct shows only two items below 10% (items 11 and 30), and zero items at 0%. Regardless, the 2PL model accounts for a floor effect as the difficulty parameters are high for these two items (> 2.75). Figures 2 and 3 present the test characteristic curve (TCC) and the test information function

Table 1. Item Parameters for the 2PL Model.

Story Number	Item Number	Difficulty	Discrimination	Proportion of Correct Answers (%)	Maximum Information
1	1	−1.89	0.83	80.14	0.17
	2	−2.25	0.83	83.97	0.17
	3	−2.15	0.40	69.69	0.04
2	4	−1.05	0.93	70.03	0.21
	5	−2.00	0.58	74.91	0.09
	6	−1.88	0.52	71.78	0.07
3	7	0.68	0.69	39.72	0.12
	8	0.07	0.72	49.13	0.13
	9	1.05	1.25	26.48	0.39
4	10	0.69	0.68	39.72	0.11
	11	2.76	1.35	04.53	0.46
	12	1.42	0.66	29.97	0.11
5	13	−0.73	0.70	61.67	0.12
	14	−0.16	0.91	53.66	0.21
	15	0.52	0.81	41.11	0.16
6	16	1.01	0.99	30.31	0.24
	17	0.60	1.16	36.93	0.33
	18	−0.92	1.29	72.13	0.41
7	19	−0.95	1.59	75.26	0.63
	20	0.20	0.75	47.04	0.14
	21	−0.67	1.03	64.46	0.27
	22	0.26	0.59	46.69	0.09
8	23	1.29	1.09	24.04	0.30
	24	−0.07	1.60	53.31	0.64
	25	0.25	1.72	43.90	0.74
	26	1.46	1.62	15.33	0.66
9	27	−0.40	1.06	59.23	0.28
	28	−1.00	1.77	77.70	0.78
	29	−0.45	0.92	59.23	0.21
	30	2.36	1.12	09.76	0.31
	31	1.73	0.71	24.74	0.13
10	32	2.97	0.73	11.85	0.14
	33	−0.82	1.75	73.52	0.77
	34	0.20	1.23	45.99	0.38
	35	4.74	0.46	10.80	0.05
	36	3.84	0.50	13.59	0.06

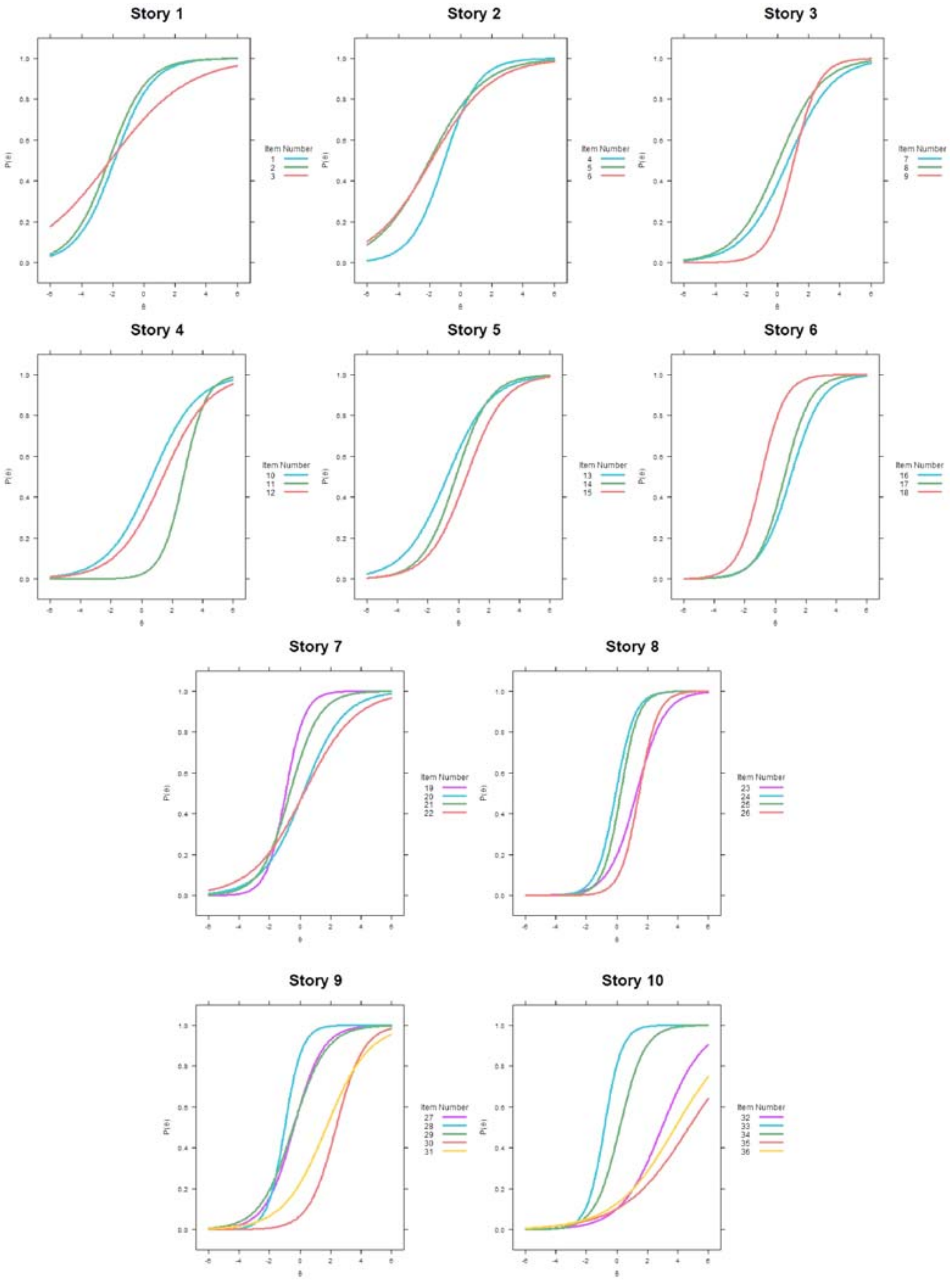


Figure 2. Item Characteristic Curves.

Note: Items are split into their story-question blocks for reading ease.

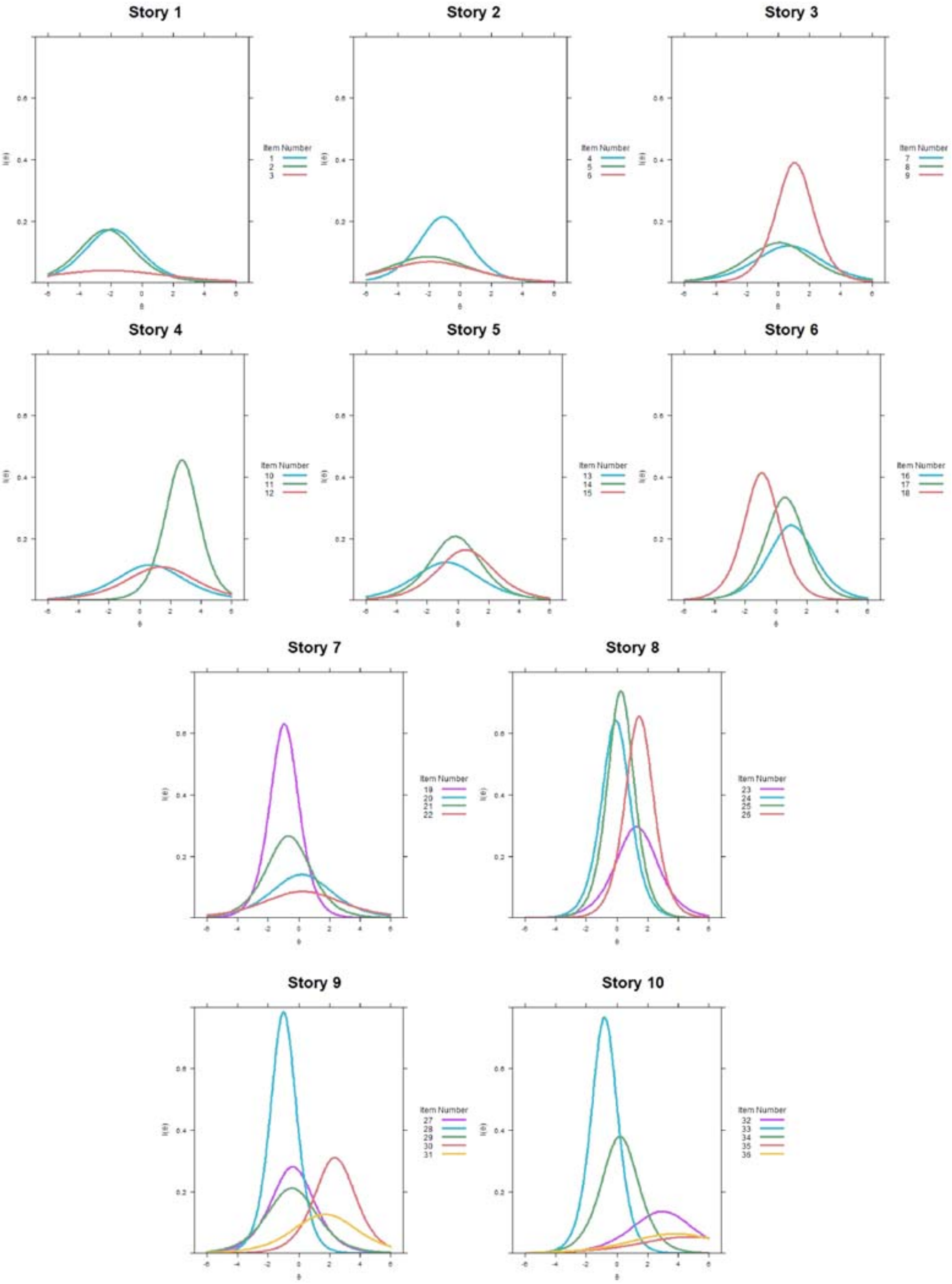


Figure 3. Item Information Functions.

Note: Items are split into their story-question blocks for reading ease.

(TIF) for the test as a whole. The TCC indicates that for an at-risk child of average listening comprehension ability ($\theta = 0$), we expect they will achieve a total score on the test of 16.85. The TIF indicates that the most precise estimates occur slightly below this same point (at approx. $\theta = -0.31$).[1]

Test Reliability

Using the R package *userfriendlyscience* (Peters, 2014), ordinal alpha indicated high internal consistency for the listening comprehension measure ($\alpha = .91$). Conditional reliability (Raju et al., 2007) using item response theory also indicated moderate consistency (ρ^θ^θ = .86). A further correlation between estimated ability scores and sum scores on the test also indicated high reliability in the usage of sum scores (r (285) = .99, $p < .001$). Ordinal alpha indicated high internal consistency for the listening comprehension measure ($\alpha = .91$).

Research Question 2. In a sample of at-risk preschoolers, how is the predictive validity of a listening comprehension assessment for language development a year later in comparison to more specific measures of language skills?

Predictive Validity

Predicting syntactic skills, the regression analysis indicated that the predictors explained 44% of the variance in growth in syntactic skills (R^2= .44, F(19, 105) = 4.26, p<.01). Listening comprehension was the only significant predictor of syntactic skills at age 5.5 ($\beta = .30$, p<.001). Neither syntactic skills at age 4.5 ($\beta = .17$, p=.10) or vocabulary at 4.5 predicted syntactic skills at age 5.5 ($\beta = .16$, p=.14). We controlled for mean differences across our test administrators in all our models.

In predicting morphological skills, we found that the predictors explained 51% of the variance (R^2=.51, F(19,105) = 5.73, p<.01). Controlling for test leader, listening comprehension significantly predicted morphological skills at age 5.5 ($\beta = .40$, p<.001), as did morphological skills at age 4.5 (β = .28, p<.01), but not vocabulary ($\beta = .18$, p=.05).

In predicting growth in narrative skills, we found that the predictors explained 53% of the variance (R^2=.53, F(19,105) =6.30, p<.01). Listening comprehension significantly predicted skills at age 5.5 ($\beta = .19$, p<.05), as did narrative skills at age 4.5 ($\beta = .52$, p<.01), but not vocabulary (β =-.02, p = .84).

Predicting vocabulary at age 5.5, the predictors explained 42% of the variance (R^2=.42, F(18,103) = 4.05, p<.01). Both vocabulary at age 4.5 ($\beta = .32$, p = .001) and listening comprehension ($\beta = .24$, p = .007) significantly predicted vocabulary at age 5.5. Thus, the listening comprehension test was a significant predictor of student's vocabulary skills a year later, even when controlling for vocabulary skills at t1.

We did not find that the listening comprehension measure predicted growth in nonverbal IQ (R^2=.29, F(19,103) = 2.21, p < .01). Nonverbal IQ at age 4.5 ($\beta = .29$, p <.01), but not listening comprehension (β = -.01, p = .89) or vocabulary (β = -.05, p = .66) significantly predicted nonverbal IQ at age 5.5.

[1] There was an assumption that children's responses would be impacted by response time and working memory efficacy. That is, a child who spent a lot of time answering the first question would then have to keep the story in their working memory for a longer amount of time than a child who answered the first question quickly. This notion can be problematic in that IRT models assume local independence: responses to earlier items do not impact responses to later items. A test of local dependence using the Q3 test statistic (Yen, 1984) was conducted to ensure that this was not the case. Four pairs of questions from the same story-question blocks did contain a Q3 test statistic above the suggested cutoff of 0.20 (Christensen et al., 2017), and thus indicated potential item dependence: items 1 and 2 from story 2; items 2 and 3 from story 3; items 1 and 2 from story 6; and items 4 and 5 from story 10. However, this is few item pairs (8%) and there is no systematic pattern (i.e., there is not a single problematic story or placement of items). Hence, local independence is established and IRT is appropriate for assessing the test items.

Discussion

In this study, we wanted to investigate the psychometric features of a short test of integrated oral language, listening comprehension (the LURI), with preschool students at the lower range of language skill development and examine its utility in predicting later discrete language skill development. Our factor analyses show that the LURI is capturing one language factor. This is consistent with studies finding that language skills are initially uni-dimensional (Language and Reading Research Consortium, 2015; Tomblin & Zhang, 2006). Moreover, the listening comprehension assessment gives the most precise estimates for children who answered approximately 50% of the questions correctly. The individual item curves show that all items perform well. Test-reliability was high, and there was no ceiling effect. Thus, the psychometric properties of the listening comprehension measure are good in this sample of at-risk preschool students.

To understand the predictive validity of the test we fit a series of multivariate models predicting children's scores on measures of syntax, morpheme generation, narrative skills, and vocabulary one year later, controlling for measured baseline scores. The LURI was a robust predictor of all later language skills, but not of nonverbal IQ. We found that the LURI predicted growth in language skills, and we were able to predict around 50% of the variance in t2 in the different language skills measured. Interestingly, our listening comprehension assessment predicted children's language development better than the BPVS. This study contributes to our understanding of language testing in young children by examining these issues in a sample of at-risk preschoolers. Our findings suggest that when measuring language in young children at risk for language problems, a measure that captures language skills' integrated use is better than assessments of discrete skills. These findings are consistent with research on the dimensionality of oral language skills in young children and demonstrate the practical utility of this research for assessing young children.

We also found that vocabulary skills at t1 did predict vocabulary skills a year later, but t1 vocabulary was not predictive of growth in any of the other language skills when LURI was also a predictor. These results are consistent with the idea that an assessment that taps specific skills might not be best suited to measure language in young children at risk. It has been relatively common to use vocabulary to measure language skills (e.g., Lonigan et al., 1998; Storch & Whitehurst, 2002). Our results suggest that we should be more precise in how we operationalize and conceptualize language skills. Our findings also emphasize that we need integrated measures in order to predict young children's language development better. A practical implication of this is that specific measures of separate skills (e.g., vocabulary tests) are not sufficient when testing young children about whose language we are concerned.

An important part of preschool teachers' job is to monitor children's language development and to identify those in need of support. Research suggests that teachers' ratings of child's language skills may not be sensitive enough (Cabell et al., 2009), and some children at-risk may go through their preschool years unidentified. Our findings suggest that a relatively quick and straightforward assessment of children's integrated oral language skills may be an excellent supplement to teacher's ratings.

Our study has limitations. One limitation is our selection criteria for identifying children at-risk for language problems. We used a screening measure testing children's receptive and expressive vocabulary. This is partly what we have criticized in this paper. However, several studies, including our own, show that children with language difficulties showed weaker vocabulary skills at an early age (e.g., Bornstein et al., 2016; Klem et al., 2016). The sample in our study was originally selected for a language intervention and the children were therefore identified as the ones that we believed would be most in need of a systematic language intervention in preschool.

A second limitation is that the vocabulary measure we compared the listening comprehension measure with is a receptive vocabulary measure. Although some would define listening comprehension measures as first and foremost receptive, our listening comprehension measure requires verbal answers. It would be interesting to include an expressive measure of vocabulary also to see if the listening comprehension measure is still the better predictor of language development.

A third limitation is that we did not include children with normal language development as a comparison. However, considering that we included 35% of the lowest scoring children on the screening measure, the majority of children in our sample are likely within the normal range, although in the lower end. In line with the COSMIN checklists (Mokkink et al., 2010) and EFPA (2013) for assessment validity, several important main steps have been taken here to establish validity and reliability of the measure. For the usability of the test it will be important to establish a critical limit for performance and analyze sensitivity and specificity of different critical limits. Since the current sample is based on children selected for language difficulties, the variation in the test scores are accordingly limited, and the sample will not be representative for norming. It should also be noted that this restriction of range in the current data set can attenuate correlations and predictive relationships. An important next step will be to select a representative sample and do norming of the test. It will be important then to examine the dimensionality of the test, whether items behave the same way in an unselected sample and if the measure can predict language development for children with typical language development. In line with the EFPA (2013) it will then also be possible to use confirmatory factor analysis to confirm the factor structure of the test.

In this study, we have demonstrated that a listening comprehension assessment can serve as an index of young children's language skills that is both valid and reliable in a sample of children with weaker language skills. This is an important step towards creating a tool that can be used by preschool teachers to identify children in need of extra language support. Listening comprehension arguably is closer to real-life language skills and less tedious for the child and the administrator than traditional measures of grammar and vocabulary, and it could, therefore, serve as a useful screening for teachers and preschool teachers.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Research council of Norway, Finnut [Grant Number 203335].

References

Alonzo, C. N., Yeomans-Maldonado, G., Murphy, K. A., Bevens, B., & Language, & Reading Research Consortium (LARRC), (2016). Predicting second grade listening comprehension using prekindergarten measures. *Topics in Language Disorders*, *36*(4), 312. https://doi.org/10.1097/TLD.0000000000000102

Bishop, D. V. M. (2003). *Test for reception of grammar. Version 2. Trog-2 manual*. London, UK: Harcourt Assessment.

Bishop, D. V., & Adams, C. (1990). A prospective study of the relationship between specific language impairment, phonological disorders and reading retardation. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *31*(7), 1027–1050. https://doi.org/10.1111/j.1469-7610.1990.tb00844.x

Block, C. C., & Mangieri, J. N. (2006). *The effects of powerful vocabulary for reading success on students' reading vocabulary and comprehension achievement*. Research Report 2963-005, Institute for Literacy Enhancement. Retrieved from http://teacher.scholastic.com/products/fluencyformula/pdfs/powerfulvocab_Efficacy.pdf

Bornstein, M. H., Hahn, C. S., & Putnick, D. L. (2016). Stability of core language skill across the first decade of life in children at biological and social risk. *Journal of Child Psychology and Psychiatry*, *57*(12), 1434–1443. https://doi.org/10.1111/jcpp.12632

Brinhmann, E., Hjetland, H. N., & Lyster, S.-A. H. (2016). Lexical quality matters: Effects of word knowledge instruction on the language and literacy skills of third- and fourth-grade poor readers. *Reading Research Quarterly*, *51*(2), 165–180. https://doi.org/10.1002/rrq.128

Cabell, S. Q., Justice, L. M., Zucker, T. A., & Kilday, C. R. (2009). Validity of teacher report for assessing the emergent literacy skills of at-risk preschoolers. *Language, Speech, and Hearing Services in Schools*, *40*(2), 161–173. https://doi.org/10.1044/0161-1461(2009/07-0099)

Cain, K., Oakhill, J. V., Barnes, M. A., & Bryant, P. E. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Memory & Cognition*, *29*(6), 850–859. https://www.ncbi.nlm.nih.gov/pubmed/11716058 https://doi.org/10.3758/BF03196414

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

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for yen's Q3: Identification of Local Dependence in the rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178–194. https://doi.org/10.1177/0146621616677520

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*(4), 422–433. https://doi.org/10.3758/BF03214546

Denman, D., Speyer, R., Munro, N., Pearce, W. M., Chen, Y.-W., & Cordier, R. (2017). Psychometric properties of language assessments for children aged 4–12 years: A systematic review. *Frontiers in Psychology*, *8*, 207. https://doi.org/10.3389/fpsyg.2017.01515

Dunn, L. M., Dunn, L. M., Whetton, C., & Burley, J. (1997). *British Picture Vocabulary Scale 2nd edition (BPVS-II)*. NFER-Nelson.

European Federation of Psychologists' Associations. (2013). *EFPA review model for the description and evaluation of psychological tests*. EFPA Board of Assessment. Downloaded June 11th 2021 from http://assessment.efpa.eu/documents-/

Florit, E., Roch, M., Altoè, G., & Levorato, M. C. (2009). Listening comprehension in preschoolers: The role of memory. *The British Journal of Developmental Psychology*, *27*(Pt 4), 935–951. https://doi.org/10.1348/026151008X397189

Florit, E., Roch, M., & Chiara Levorato, M. (2013). The relationship between listening comprehension of text and sentences in preschoolers: Specific or mediated by lower and higher level components? *Applied Psycholinguistics*, *34*(2), 395–415. https://doi.org/10.1017/S0142716411000749

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

Fricke, S., Burgoyne, K., Bowyer-Crane, C., Kyriacou, M., Zosimidou, A., Maxwell, L., Lervåg, A., Snowling, M. J., & Hulme, C. (2017). The efficacy of early language intervention in mainstream school settings: A randomized controlled trial. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *58*(10), 1141–1151. https://doi.org/10.1111/jcpp.12737

Gillam, R. B., & Pearson, N. A. (2017). *TNL-2: Test of Narrative language*. Pro-ed.

Gjessing, H. J., Nygaard, H. D., & Solheim, R. (1975). *ITPA: Illinois Test of Psycholinguistic Abilities: En orientering om den norske utgaven av ITPA forfatterne: Bergen*. Oslo, Norway: Universitetsforlaget.

Hagen, ÅM, Melby-Lervåg, M., & Lervåg, A. (2017). Improving language comprehension in preschool children with language difficulties: A cluster randomized trial. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *58*(10), 1132–1140. https://doi.org/10.1111/jcpp.12762

Haley, A., Hulme, C., Bowyer-Crane, C., Snowling, M. J., & Fricke, S. (2017). Oral language skills intervention in preschool—a cautionary tale. *International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists*, *52*(1), 71–79. https://doi.org/10.1111/1460-6984.12257

Hjetland, H. N., Brinchmann, E. I., Scherer, R., & Melby-Lervåg, M. (2017). Pre-school predictors of later reading comprehension ability: A systematic review. *Campbell Systematic Reviews*, *13*(1), 1–155. https://doi.org/10.4073/csr.2017.14

Hogan, T. P., Adlof, S. M., & Alonzo, C. N. (2014). On the importance of listening comprehension. *International Journal of Speech-Language Pathology*, *16*(3), 199–207. https://doi.org/10.3109/17549507.2014.904441

Kim, Y. S., & Phillips, B. (2014). Cognitive correlates of listening comprehension. *Reading Research Quarterly*, *49*(3), 269–281. https://doi.org/10.1002/rrq.74

Kirk, S. A., McCarthy, J. J., & Kirk, W. D. (1968). *Illinois test of psycholinguistic abilities*. University of illinois press Urbana.

Klem, M., Hagtvet, B., Hulme, C., & Gustafsson, J.-E. (2016). Screening for language delay: Growth trajectories of language ability in Low- and high-performing children. *Journal of Speech, Language, and Hearing Research: JSLHR*, *59*(5), 1035–1045. https://doi.org/10.1044/2016_JSLHR-L-15-0289

Kunnskapsdepartementet [The Norwegian Ministry of Education]. (2011). Vurdering av verktøy som brukes til å kartlegge barns språk i norske barnehager: Rapport fra Ekspertutvalget nedsatt av Kunnskapsdepartementet 2010/2011. [Assessment of tools used to map children's language in Norwegian kindergartens: Report from the Expert committee reduced by Ministry of Education 2010/2011.].

Language and Reading Research Consortium (LARRC). (2017). Oral language and listening comprehension: Same or different constructs?. *Journal of Speech, Language, and Hearing Research*, *60*(5), 1273–1284. https://doi.org/10.1044/2017_JSLHR-L-16-0039

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

Language and Reading Research Consortium, Jiang, H., & Logan, J. (2019). Improving reading comprehension in the primary grades: Mediated effects of a language-focused classroom intervention. *Journal of Speech, Language, and Hearing Research*, *62*(8), 2812–2828. https://doi.org/10.1044/2019_JSLHR-L-19-0015

Lervåg, A., Hulme, C., & Melby-Lervåg, M. (2018). Unpicking the Developmental relationship between oral language skills and Reading comprehension: It's simple, But complex. *Child Development*, *89*(5), 1821–1838. https://doi.org/10.1111/cdev.12861

Lonigan, C. J., Burgess, S. R., Anthony, J. L., & Barker, T. A. (1998). Development of phonological sensitivity in 2- to 5-year-old children. *Journal of Educational Psychology*, *90*(2), 294–311. https://doi.org/10.1037/0022-0663.90.2.294

Lyster, S. A. H., & Horn, E. (2009). *Test for reception of grammar (TROG – 2): norsk versjon*. Pearson Assessment.

Lyster, S. A. H., Horn, E., & Rygvold, A.-L. (2010). Ordforråd og ordforrådsutvikling hos norske barn og unge. Resultaterfraen utprøving av[British Picture Vocabulary Scale, Second Edition (BPVS II)]. *Spesialpedagogikk*, *9*, 35–43. https://scholar.google.com/scholar_lookup?title=Ordforr%C3%A5d+og+ordforr%C3%A5dsutvikling+hos+norske+barn+og+unge%3A+resultater+fra+en+utpr%C3%B8ving+av+British+Picture+Vocabulary+Scale+II%2C+second+edition+%5BVocabulary+and+vocabulary+development+in+Norwegian+children+and+youth%3A+results+from+testing+with+British+Picture+Vocabulary+Scale+II%5D&author=Lyster+S.-A.+H.&author=Horn+E.&author=Rygvold+A.-L.&publication+year=2010&journal=Spesialpedagogikk&volume=9&pages=35-43

Malec, A., Peterson, S. S., & Elshereif, H. (2017). Assessing young children's oral language: Recommendations for classroom practice and policy. *Canadian Journal of Education/Revue Canadienne de L'éducation*, *40*(3), 362–392. https://journals.sfu.ca/cje/index.php/cje-rce/article/view/3119

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713. https://doi.org/10.1007/s11336-005-1295-9

McKean, C., Wraith, D., Eadie, P., Cook, F., Mensah, F., & Reilly, S. (2017). Subgroups in language trajectories from 4 to 11 years: The nature and predictors of stable, improving and decreasing language trajectory groups. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *58*(10), 1081–1091. https://doi.org/10.1111/jcpp.12790

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international delphi study. *Quality of Life Research*, *19*(4), 539–549. https://doi.org/10.1007/s11136-010-9606-8

Neale, M. D., McKay, M. F., & Childs, G. H. (2011). The Neale Analysis of Reading ability-revised. *British Journal of Educational Psychology*, *56*(3), 246–256. https://doi.org/10.1111/j.2044-8279.1996.tb01194.x

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

Orlando, M., & Thissen, D. (2000). Likelihood-Based item-Fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Peters, G. Y. (2014). The alpha and the omega of Scale reliability and validity: Why and how to abandon Cronbach's alpha. *European Health Psychologist*, *16*(S), 576.

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized ConditionalSEM: A case for Conditional reliability. *Applied Psychological Measurement*, *31*(3), 169–180. https://doi.org/10.1177/0146621606291569

Rambøll Management. (2008). *Kartlegging av språkstimulering og språkkartlegging i kommunene. [assessment of language stimulation and language assessments in the municipalities]*. The Norwegian Ministry of Education.

Raven, J. C. (1998). *Raven's progressive matrices and vocabulary scales*. http://www.v-psyche.com/doc/IQ/Raven-Vocabulary.doc

Renfrew, C. (1997). *Bus story test: A test of narrative speech (4th ed.)*. Winslow Press.

Rogde, K., Hagen, ÅM, Melby-Lervåg, M., & Lervåg, A. (2019). The effect of linguistic comprehension instruction on generalized language and reading comprehension skills: A systematic review. *Campbell Systematic Reviews*, *15*(4), e1059.

Statens beredning för medicinsk utvärdering [Swedish Council on Health Technology Assessment]. (2014). *Dyslexi hos barn och ungdomar - tester och innsatser. En systematisk litterauröversik [Dyslexia in children and adolescence – tests and efforts: A systematic review]*. Statens beredning för medicinsk utvärdering [Swedish Council on Health Technology Assessment].

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology*, *38*(6), 934–947. https://doi.org/10.1037/0012-1649.38.6.934

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

Wechsler, D. (1989). *Wechsler preschool and primary scale of intelligence-revised*. Psychological Corporation.

Wiig, E. H., Semel, E. M., & Secord, W. (2003). *CELF 5: Clinical evaluation of language fundamentals*. Pearson/PsychCorp.

Yen, W. M. (1984). Effects of Local item Dependence on the Fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145. https://doi.org/10.1177/014662168400800201

Yew, S. G. K., & O'Kearney, R. (2013). Emotional and behavioural outcomes later in childhood and adolescence for children with specific language impairments: Meta-analyses of controlled prospective studies. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *54*(5), 516–524. https://doi.org/10.1111/jcpp.12009