Developing an assessment of oral language and literacy: Measuring growth in the early years

Dr Dan Cloney
Australian Council for Educational Research

Kellie Picker
Australian Council for Educational Research

https://doi.org/10.37517/978-1-74286-638-3_2

Dr Dan Cloney is a Senior Research Fellow in the Education Policy and Practice Program and a Member of the Global Education Monitoring (GEM) Centre at the Australian Council for Educational Research (ACER). Dan's expertise is in early education, cognitive development, and academic achievement.Dan's research program focuses on the potential for high-quality early childhood education and care programs to support all children to flourish and to reduce development gaps caused by inequality. Current studies include the Overcoming Disadvantage in Early Childhood study (in partnership with the Australian Literacy & Numeracy Foundation) evaluating the effectiveness of the Early Language & Literacy program in New South Wales, and the Modelling of Universal Pre-Primary Education Study (in partnership with UNICEF) evaluating the effectiveness of pre-primary programs in Bogor, Indonesia. He is also an Honorary Research Associate at the Murdoch Children's Research Institute (MCRI) and the Melbourne Graduate School of Education (MGSE).

Kellie Picker is a researcher for the Australian Council for Educational Research (ACER). She is currently working on the Overcoming Disadvantage in Early Childhood study evaluating the effectiveness of the Early Language & Literacy program in New South Wales, as well as the Learning Through Play in School research project. The Learning Through Play at School project is an intervention study aiming to help Ukrainian teachers transition from adult-centred teaching to child-centred learning through the use of play pedagogies. Kellie's expertise is in early childhood education, early literacy development and effective pedagogies. She was a researcher on the Western Australian Teaching for Growth *project and has been a classroom teacher, a learning enrichment specialist and an IT coordinator in primary education. Kellie is in the final stages of completing a PhD at the University of Melbourne with a focus on deepening our understanding of the role of teacher reading content knowledge in early primary education.*

Abstract

Children develop rapidly in their early years. A crucial component of this development is a child's ability to learn and use language. Even before they enter formal education, children have learned much about oral language and literacy through meaningful interactions with others, and from their life experiences. Children, however, do not develop at the same pace – some children arrive in early childhood education and care (ECEC) programs more advanced while others require additional support. Recent reviews of the assessment tools available to ECEC educators show a lack of good quality measurement and a reliance on checklist style inventories or narrative approaches. This paper presents a new measure of oral language and pre-literacy specifically designed to be accurate enough to reliably measure an individual child's growth. Results from a combined calibration of children's responses using a many-facets item response model show the measure to be reliable, valid and sensitive enough to measure growth within children and between groups of children over time. Implications for future assessment development and for educators' practice are discussed, including how such measures can provide insight into what children know, understand, and can do (Reynolds, 2020) and what educators can do to support future learning experiences targeted at children's specific language and literacy needs.

Introduction

Oral language and literacy

The development of oral language and literacy is a core skill, important in its own right and also predictive of later learning and development in both school curriculum areas (academic achievement) and in general capabilities, such as cognitive and social skills. Oral language research highlights that it is our innate need and ability to communicate using language that supports and develops our use of expressive and receptive language. Here, research suggests that children who start formal education with advanced oral language skills, are more likely to be successful readers (Foorman et al., 2015).

Literacy learning, including oral language, does not come naturally, and children require specific knowledge and skills so they can learn to read and write. Underpinning the construct of literacy are five key elements:

- phonemic awareness (a sub-strand of phonological awareness)
- phonics
- fluency
- vocabulary
- comprehension.

These five elements were identified by the National Institute of Child Health and Development (NICHD) (2000) and Rowe (2005) as essential to the effective teaching of reading in English speaking classrooms and have come to be known as the 'science of reading'. However, in a review of the literature, Konza (2014) argued that due to oral language being foundational to literacy learning, it should be included as an essential element of the science of reading, renaming them the 'Big Six'.

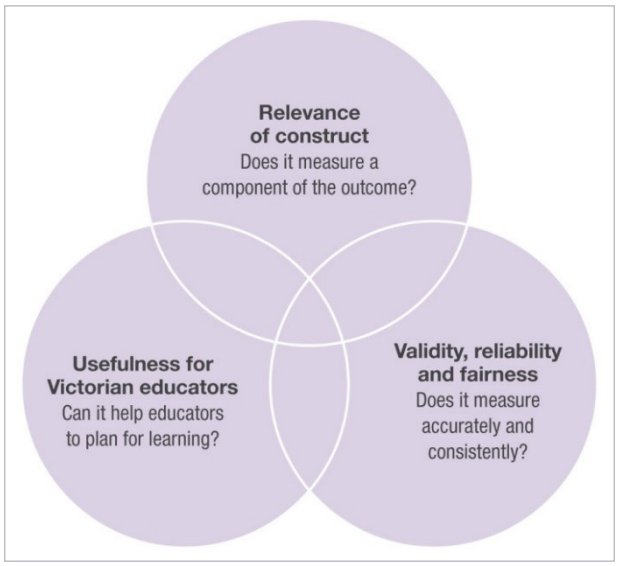
In Australia, the 'Big Six' are represented in the national learning progressions. Here, 'oral language, vocabulary and comprehension are reflected across many areas of the progressions, and phonemic awareness, phonics and fluency are addressed as sub-elements' (Australian Institute for Teaching and School Leadership [AITSL], 2020, p. 8).

Assessment

There is a lack, however, of quality assessments that measure language and literacy and describe how they develop over time. Recent literature reviews of assessment available to early educators show that assessment tends towards checklist-style inventories that are designed to screen for developmental problems and do not describe what children can do and what they might do next (and what educators might do to support this). While other forms of assessment tend toward narrative that is difficult to communicate to other educators, parents and to teachers at transition, and is applied inconsistently (Anzai et al., 2021; Cloney et al., 2020).

Assessing children's language and literacy growth is complex and multifaceted, particularly in early childhood where assessment cannot take on the form of traditional school age assessment (completing a test) but rather should be embedded within authentic interactions between adults and children. Cloney et al. (2020) developed a series of principles that help educators understand the elements of quality assessments in the early years, as shown in Figure 1. These principles are anchored in best practice for early learning, informed by frameworks such as the Early Years Learning Framework and the Victorian Early Years Learning and Development Framework.

Figure 1 Principles for assessing oral language and early literacy in children, adapted from Anzai et al. (2021, p. 23)



Principles for good quality assessment in the early years

The first two key principles apply to the selection of an assessment, to ensure it addresses the knowledge, skills and concepts (construct) being measured, in a way that is developmentally appropriate for the age and stage of the children. It is important that educators are familiar with how these constructs are discussed in the literature, to assist the educator in making decisions about the suitability of the assessment. This includes the breadth and/or depth of coverage of the construct and whether it aligns with framework and curriculum documents that govern teaching and learning. Such an assessment should also be designed to maintain or even enhance those strong, warm and responsive relationships that are foundational to children's learning (Pianta et al., 2008). Along with maintaining relationships, effective assessments engage children in meaningful interactions and experiences that are challenging but within the child's learning reach (Palermo et al., 2007).

The next principles of assessment consider the validity, reliability and fairness of the measure. Assessments are found to be 'valid' when they accurately measure the construct/s they claim to measure, for example, expressive language. 'Reliable' assessments are said to produce 'valid results consistently across contexts' and measures of 'fairness' are based on whether an assessment provides children with unbiased opportunities to express what they know, without disadvantaging specific groups of children (Cloney et al., 2020, p. 19).

Finally, it is of the utmost importance that the information collected in an assessment can be used by educators to understand and describe student's growth on a continuum of learning. Educators who have a deep understanding of how language and literacy develops and can discuss growth as a continuum of learning, such as those described in learning progressions, are well equipped to facilitate children's learning. Educators with knowledge about what children know and can do now, can use assessment information to plan for what children need to know and do next. This in-depth knowledge means educators can target learning experiences, scaffold children's learning, and contribute to the growth of all children (Cloney et al., 2020).

Research questions

This paper therefore asks whether it is possible to develop an assessment that is contextually appropriate for children (built around authentic tasks and one-on-one interactions between an adult and child) and still reliable, valid and fair. That is, can best-practice in assessment development and measurement be applied to an early years measure? Further, is such a measure sensitive enough to measure growth within children and between groups of children?

Method

Participants

Twenty-five early childhood education and care (ECEC) services were sampled from one regional area of NSW. Twelve of these were sampled from ECEC services implementing a specific oral language and literacy intervention (the intervention group) and 10 were matched (on observed characteristics including National Quality Standard (NQS) rating, service type, size, neighbourhood socioeconomic background) controls (the control group).[1] In total, 22 agreed to participate, with one service being uncontactable (likely closed), one having fewer than 10 enrolments, and one refusing. From these services, preschool-aged children (generally children in the age range 3–5 years) were invited to participate. In total, 571 children in 27 rooms or groups within the recruited centres agreed.[2] The final achieved sample is described in Table 1 and Table 2, disambiguated by the intervention and control groups.

Table 1 Service-level characteristics achieved sample

Variable	Intervention	Control
Count	12	10
Average enrolment	42	41
Exceeding NQS	3	5
Meeting NQS	6	4
Working towards NQS	3	1
Count community preschool	8	6
Count long day care	4	4
Median SEIFA IRSAD*	863	923

* Socio-Economic Indexes for Areas, Index of Relative Socio-Economic Advantage and Disadvantage

1 Although not a focus of this paper, it is important to note that the centres are not a random sample of all ECEC services in the region nor is the sample representative of the services operating in the region.

2 All classrooms or groupings of children were recruited to the Study where they included at least five children of preschool age (e.g. who would be eligible to attend school in 2019, whether their parents intended them to or not). Within each sampled room or group, all children and their main caregiver (a caregiver who lives with the child, and usually the person who drops them off to the centre, or the person who knows the most about the child and their daily routine) were invited to participate.

Table 2 Counts of participants in achieved sample

Variable	Intervention	Control
Centres	12	10
Rooms/groups	14	13
Educators	14	13
Children	263	308
Main caregivers	263	307

Instrumentation

The purpose of this study was to generate a measure of oral language and literacy that was suitable to measure within-child growth and that represented a developmental continuum that covered the range of abilities of children from age (approximately) 2 to 8 years. The assessment was designed in line with the Big 6 framework and assessment principles described in the introduction.

Items were developed around a range of interactions, including a picture book reading comprehension and extended response activity (children are read a picture book and asked a range of questions including extended response), a scene picture (a familiar scene with actions happening that the child is asked to observe and or discuss; for example see Figure 2) and more traditional question-and-answer items (a child is given a prompt and asked to respond). Some responses were audio recorded for the purpose of later coding against a rubric to allow deeper understanding of children's oral language – particularly concepts around sophistication of ideas, fluency, and clarity of expression.

Figure 2 Beach scene oral language stimulus, interview script and scoring guide

1	**Instructions and Questions** *Pass the child the beach scene.* Now please hold this picture and look at it carefully. Then I will ask you some questions about it. *Give the child 30 seconds to look at the picture.* There are lots of people doing things. Tell me something that is happening in the picture. *If no response, prompt with:* Can you tell me what someone is doing? *If the child shakes head then stop; if nods head then say:* Could you please tell me, even in a whisper is ok.		
	Responds with an observation about the picture **1**	Responds but unrelated to the picture **0**	Child says I do not know or no response **99**
Expressiveness	○	○	○

In addition to the main assessment instrument, contextual questionaries and publicly available data were used to elicit demographic and background information about children, families, home environments, ECEC programs, and local communities.

Analytic approach

All analysis was conducted in ACER ConQuest Version 5 (Adams et al., 2020) and the conquestr package (Cloney & Adams, 2021) in R (R Core Team, 2020).

Calibration

With up to two observations of each child in the study, all information is used in a concurrent calibration to yield item parameters. To do this, the data are represented as a single row per child and cycle combination. That is, children who were observed twice have two rows in the data set.

We estimate a one parameter item response model (1PL) – the many facets model (facet model) (Linacre, 1993) – an extension of the Partial Credit Model (PCM) (Masters, 1982). This facet model allows the responses (at all time points) to all the items, to be decomposed into an item difficulty component (the location of the items on the oral language continuum) and some average deviation from that difficulty at each time point. In this model, the assumption is that as time increases, the items get easier.

Child responses to items are integer scored from 0 (most incorrect) to m (most correct) at each time point t. If we denote the latent ability of child n as θ_n, and the 'difficulty' of each item, i, is made up by the item category boundaries (e.g. the boundary between scoring 1 rather than 0, and between scoring 2 rather than 1), which have three components, δ_i (the 'average' difficulty of the item), plus $\tau_{i\kappa}$ (the deviation from the average difficulty for this category boundary), plus α_t (the average change in the difficulty of items over time), then the probability of child n scoring x on item i at time t is given by:

Equation 1 Probability model of the many facets item response model

$$p(X_{nit} = x) = \frac{exp\ \Sigma_{j=0}^{x}(\theta_n - (\delta_i + \tau_{i\kappa} + \alpha_t))}{\Sigma_{\kappa=0}^{m} exp\ (\Sigma_{j=0}^{\kappa}(\theta_n - (\delta_i + \tau_{i\kappa} + \alpha_t)))}$$

The continuing product of the probabilities for child n's responses to many items represents a likelihood, given their response vector. The unknown parameters θ[3], δ_i, $\tau_{i\kappa}$ and α_t are then estimated by maximum likelihood.

Scaling and modelling

Taking the item parameters (δ_i, $\tau_{i\kappa}$) from the calibration stage as fixed, a two-dimensional 1PL item response model (one dimension for each time point) is estimated. That is, the time facet is removed from the specification and each time point is modelled as a dimension. This yields a measurement model where the increasing ability is reflected in the change in the response vectors (i.e. older children tend to get more items correct) holding the item difficulties constant (note that only some items are used between time points). This is equivalent to a latent growth model, as there is some fixed (average) growth between time point 1 and 2, and a random effect within children at each time point (some random deviation from the average at each time point). This model is used to calculate

3 The distribution of θ involves a population distribution assumption (estimation is by marginal maximum likelihood) that means the parameters of the model are influenced by both the responses to the assessment items as well as the regression variables in the model.

factor scores (plausible values [Von Davier et al., 2009]) at each time point that are then used in secondary analysis. A limited set of conditioning variables are included in the model (population model) to support the secondary analysis in this paper – including a variable indicating whether children are in the intervention and control group – a discussion of the generation of plausible values with regression variables is outside the scope of this work (e.g. Wu, 2005).

Results

Calibration

In general, the model shows good fit to the data (Adams & Wu, 2009). Most items fit between rule-of-thumb guides for weighted mean square (WMNSQ). An example item characteristic curve is given in Figure 3 – in this figure the WMNSQ (sometimes called infit) is 1.06 (95%CI 0.87–1.13). A plot of the item (deltas) fits against the quantiles of the normal distribution (a Q-Q plot) reinforce this – the expectation is that WMNSQ is distributed with expectation 1 and variance 1. Note that in Figure 4, 10 items showing poor fit (all underfit) to the model are excluded.

Figure 3 Example item characteristic curve for a polytomous item with good fit to the model

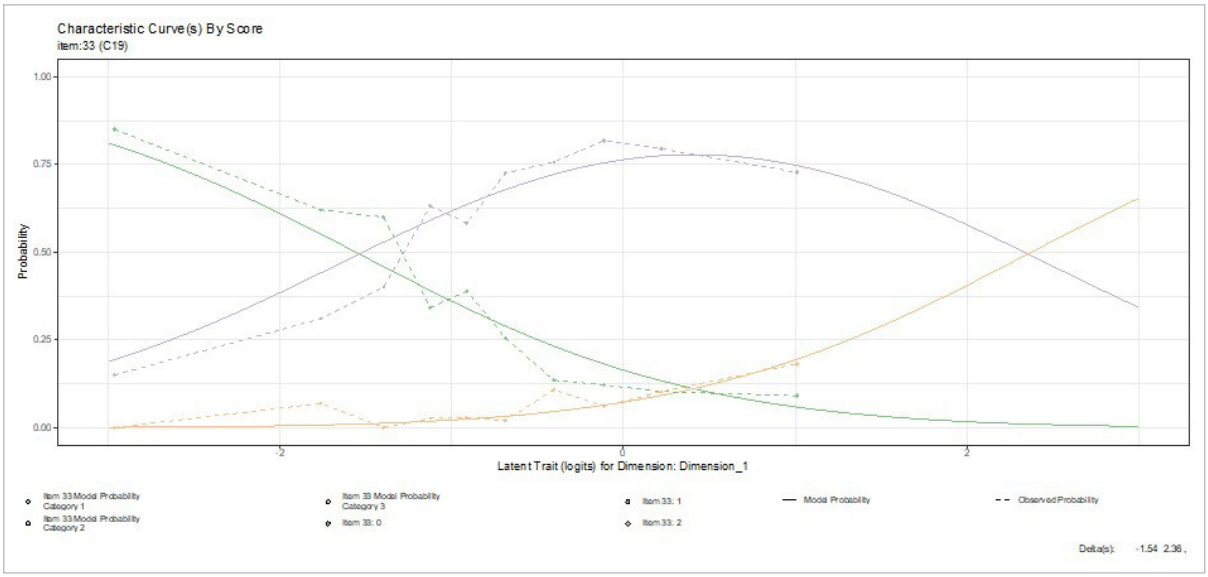
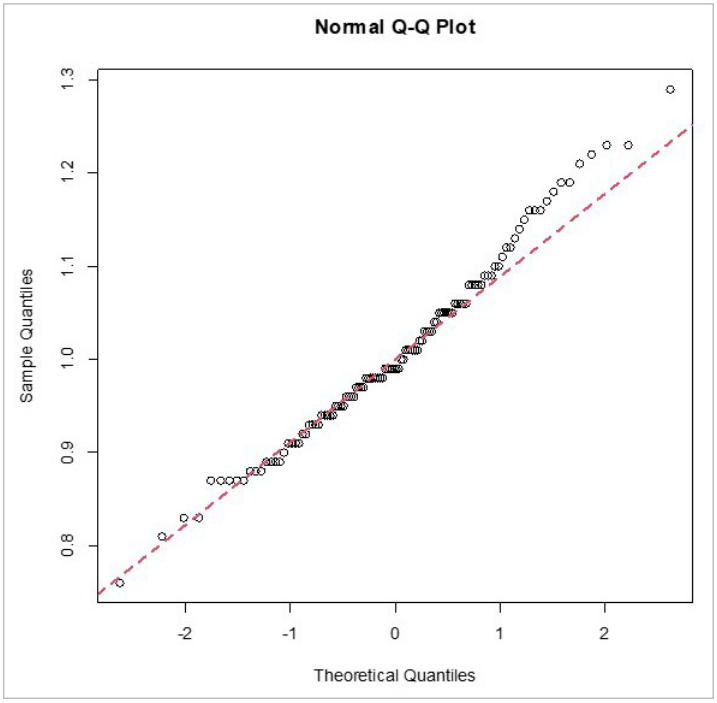


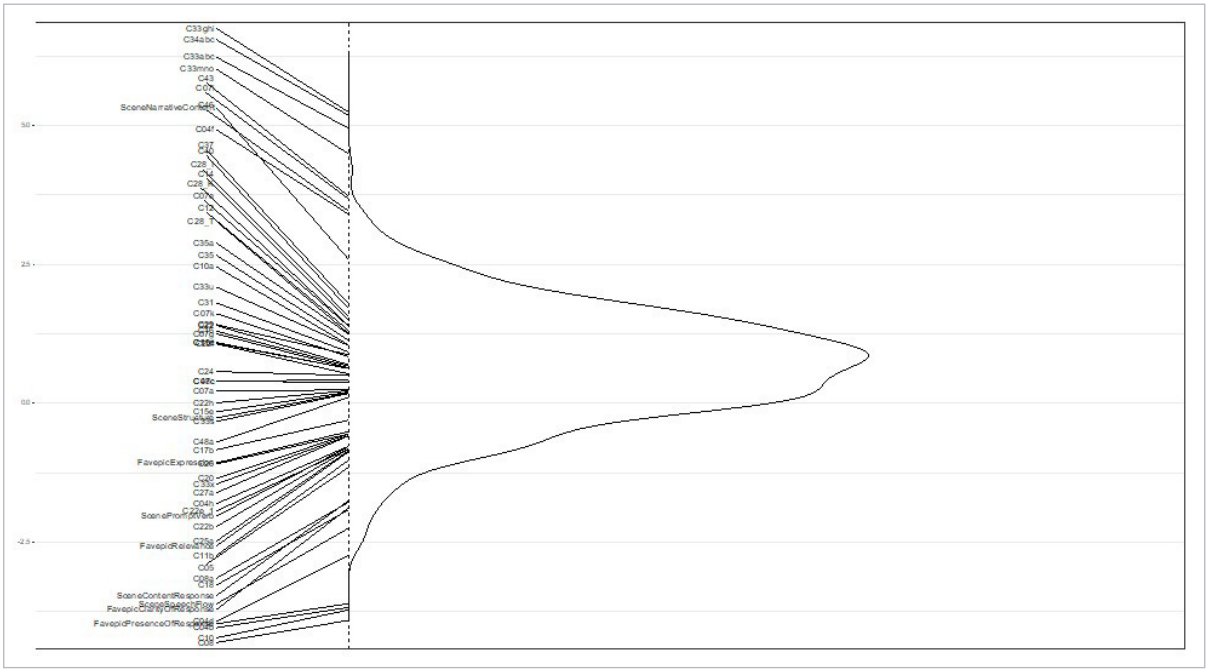
Figure 4 Plot of item delta weighted mean square fit statistics against quantiles of the standard normal distribution



Scale reliability is very high. Plausible value (PV) reliability is 0.97 and weighted likelihood estimation (WLE) separation reliability is 0.94. Note that with more than 100 polytomous items in the test (summing to more than 200 category boundaries) administered to more than 500 children on two time points, a high reliability is likely for most models. For example, PV reliability at any one time point, is 0.79 in 2018 and 0.88 in 2019.

Targeting is good, although there are a number of items outside the range of the ability distribution (the density shown on the right of Figure 5). There are also relatively fewer items in the upper range of the ability distribution and this is something to be better targeted in future item development.

Figure 5 Wright Map of item location (difficulties) relative to the distribution of children's abilities

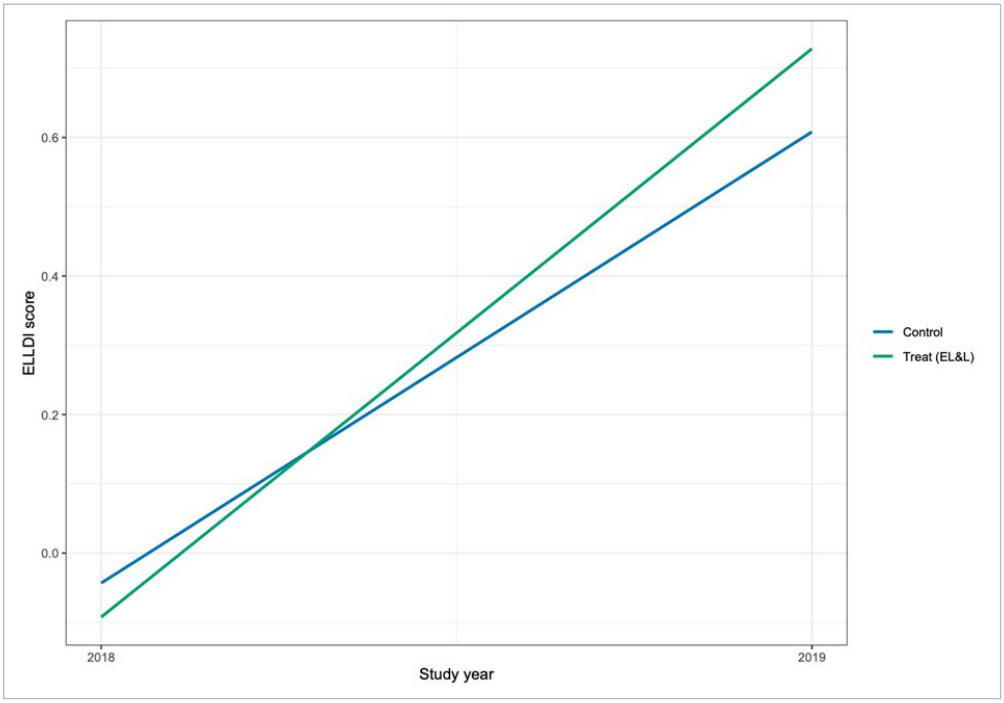


Note only a selection of item labels are printed to minimise overlap.

Scaling and modelling

Holding item parameters constant at both time points, the average growth between 2018 and 2019 is 0.70 logits (approximately 0.63 standard deviations [SD] of the 2019 ability distribution). The correlation between children's scores at time 1 and 2 is 0.78. In an unconditional regression model, there is a small positive effect for the interventions. Children in the intervention group start marginally behind the control group (ES = −0.05, *ns*) and have grown slightly faster to be, on average, above the control group at the end of time 2 (0.14 SD, $p < 0.05$), as shown in Figure 6. Caution should be made with interpreting these interim results as the longitudinal design is not complete, with a third wave of data collection completed in 2021.

Figure 6 Average linear growth trajectories for study children between time 1 and 2



Discussion

This paper shows that it is possible to develop an assessment that aligns both with guiding principles of good quality assessment in the yearly years and cutting-edge psychometric techniques. The development of assessments that access a developmental continuum (and describe them – so called *learning progressions*) is growing quickly in school education circles. We show that this kind of progress can be made in ECEC and can be done in a way that is familiar to educators: one-on-one interactions with children around activities like reading a picture book.

The utility of strong measurement and description of oral language and pre-literacy development is clear. If educators can see, on a continuum, 1) where a child is currently performing in terms of their oral language and pre-literacy skills and 2) what are the skills and abilities that typically develop next, then they can focus their efforts on making conceptually sound incremental improvements that are targeted as individual child needs. This approach will also build educator capacity by providing a conceptual understanding of the construct they are supporting children's growth in.

Ongoing work is required to develop an item bank that covers the range of abilities that cover the typical oral language and literacy development of children aged 2 to 8 years. Further, it is important to carefully assess the calibration of this measure as more data is collected. Future work should include moving towards adaptive assessment or at least providing a way for educators to ensure the tasks and items given to children are optimally targeted at their level of ability. The development of automatic scoring and reporting will also ensure educators get access to high-quality learning data on the same day they complete an assessment of a child.

References

Adams, R. J., & Wu, M. (2009). The construction and implementation of user-defined fit tests for use with marginal maximum likelihood estimation and generalized item response models. *Journal of Applied Measurement*, *10*(4).

Adams, R. J., Wu, M. L., Cloney, D., & Wilson, M. R. (2020). *ACER ConQuest: Generalised Item Response Modelling Software* (Version 5) [Computer software]. Australian Council for Educational Research.

Anzai, D., Knowles, S., Cloney, D., Munro-Smith, P. & Mitchell, P. (2021). *Assessment of oral language and early literacy in early childhood education and care: Literature review*. [Manuscript submitted for publication]. Victorian Curriculum and Assessment Authority.

Australian Institute for Teaching and School Leadership. (2020). *Reading instruction evidence guide*. AITSL. https://www.aitsl.edu.au/docs/default-source/initial-teacher-education-resources/reading-instruction/reading-instruction-evidence-guide.pdf?sfvrsn=5465d73c_2

Cloney, D., & Adams, R. J. (2021). *conquestr* (Version 0.9.9) [Computer software]. Australian Council for Educational Research. https://cran.r-project.org/package=conquestr

Cloney, D., Jackson, J., & Mitchell, P. (2020). *Assessment of children as confident and involved learners in early childhood education and care: Literature review*. Victorian Curriculum and Assessment Authority. https://www.vcaa.vic.edu.au/Documents/earlyyears/EYLitReviewLearning.pdf

Foorman, B. R., Herrera, S., Petscher, Y., Mitchell, A., & Truckenmiller, A. (2015). The structure of oral language and reading and their relation to comprehend in Kindergarten through Grade 2. *Reading and Writing*, *28*(5), 655–681. https://doi.org/10.1007/s11145-015-9544-5

Konza, D. (2014). Teaching reading: Why the 'Fab Five' should be the 'Big Six'. *Australian Journal of Teacher Education*, *39*(12), 153–169.

Linacre, J. M. (1993). *Generalizability theory and many-facet Rasch measurement.* American Educational Research Association.

National Institute of Child Health and Development. (2000). *Report of the National Reading Pane. [Teaching children to read] An evidence-based assessment of the scientific research literature on reading and its implications for reading instructions*. (No. 00-4769). Washington DC: Government Printing Office. http://www.nichd.nih.gov/publications/nrp/smallbook.cfm

Palermo, F., Hanish, L., Martin, C., Fabes, R., & Reiser, M. (2007). Preschoolers' academic readiness: What role does the teacher–child relationship play?, *Early Childhood Research Quarterly*, *22*(4), 407–22. https://doi.org/10.1016/j.ecresq.2007.04.002.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System: Manual K–3*. Brookes Publishing Co.

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. www.R-project.org/

Reynolds, B. (2020). Assessment and documentation for children's learning and development. In Kilderry, A. & Raban, B. (Eds.), *Strong foundations: Evidence informing practice in early childhood education and care* (pp. 250–265). ACER Press.

Rowe, K. (2005). *Teaching reading: National inquiry into the teaching of literacy*. Australian Government Department of Education, Science and Training. https://research.acer.edu.au/tll_misc/5

Von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? *IERI Monograph Series*, *2*, 9–36.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2–3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005