African Journal of Psychological Assessment
ISSN: (Online) 2617-2798, (Print) 2707-1618

Piloting the PuzzleBox Screener for preschool assessment in the Eastern Cape, South Africa

Check for updates

Authors:
Rivca Marais[1]
Jennifer M. Jansen[2]

Affiliations:
[1]Department of Criminology, Psychology and Social Work, Faculty of Social Sciences and Humanities, University of Fort Hare, Dikeni, South Africa

[2]Department of Psychology, Faculty of Health Sciences, Nelson Mandela University, Gqeberha, South Africa

Corresponding author:
Rivca Marais,
rmarais@ufh.ac.za

Dates:
Received: 25 Feb. 2025
Accepted: 12 Sept. 2025
Published: 12 Nov. 2025

How to cite this article:
Marais, R., & Jansen, J.M. (2025). Piloting the PuzzleBox Screener for preschool assessment in the Eastern Cape, South Africa. *African Journal of Psychological Assessment*, *7*(0), a178. https://doi.org/10.4102/ajopa.v7i0.178

Copyright:
© 2025. The Authors. Licensee: AOSIS. This work is licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/).

Read online:
Scan this QR code with your smart phone or mobile device to read online.

This article explores the shift from Western-based assessment tools to contextually relevant approaches to testing that reflect the experiences of preschool children in South Africa. This study reports on the piloting phase of the newly developed PuzzleBox Screener, with the aim of evaluating the psychometric performance of items to inform decisions on retaining, refining or removing items for the PuzzleBox Screener. The tool uses minimal, recyclable equipment to screen developmental progress in children aged 5 years to 6 years and 11 months in isiXhosa, English and Afrikaans from peri-urban, rural and informal settlements in the culturally diverse Eastern Cape. A quantitative design evaluated 65 children across five developmental domains: cognitive, language, fine motor, socio-emotional and gross motor during the piloting phase. Item selection was determined using a structured approach combining factor loadings from the exploratory factor analysis and item discrimination and difficulty from the item response theory (IRT). Outcomes from this phase included 15 cognitive, 9 fine motor, 12 emotion–social–moral, 12 language and 6 gross motor items, which can be revised or retained.

Contribution: The findings indicate that the PuzzleBox Screener can identify developmental delays across the domains investigated in the study using minimal, recyclable equipment. The results support the tool's feasibility in South Africa and emphasise the need for context-specific measures, innovation and change in early childhood assessment.

Keywords: developmental screening; early childhood; preschool assessment; the PuzzleBox Screener; screening tool.

Introduction

Developmental screening is shifting from a test-and-tell approach to a broader perspective that includes economic, societal, cultural and neuroscientific influences (Green et al., 2020). This shift highlights the adoption of more authentic and accessible assessment practices within an African context (Bagnato et al., 2024). However, this more progressive approach remains a challenge when addressing diverse preschool assessment needs in developing countries.

Children from low- and middle-income countries (LMICs) like South Africa are faced with a myriad of challenges that can impact their development when compared to first-world countries. Unfortunately, these challenges are more prevalent in rural and informal settlement areas, even after more than 30 years of an established democracy. Approximately 250 million (43%) children under the age of 5 years in LMICs, including South Africa, are failing to meet their developmental potential (Zhang et al., 2021). The repercussions of early biological and psychosocial experiences rooted in environments, which are characterised by poverty, violence, nutritional deficiencies, human immunodeficiency virus (HIV) infection and limited learning opportunities, disproportionately affect South African children and were highlighted in The Lancet as early as 2011 (Engle et al., 2011).

According to the National Association for the Education of Young Children (NAEYC, 2019), providing robust support during the formative years is essential for creating a more equitable future, breaking the cycles of disadvantage and fostering broader societal development. Despite the unique challenges faced by South African children, standardised tests imported from the Global North are still being used as a 'one-size-fits-all' approach and justified by the belief that it is preferable to having no assessments at all. It is important to develop assessment measures that consider diverse circumstances and adopt a holistic approach, integrating the child, family, school and community in the process (Laher & Cockcroft, 2017).

The interplay between these spheres is captured by the foundational work of Ruth Griffiths, a pioneering figure in developmental assessment, and the subsequent expansion and modernisation of her Avenues of Learning Model (Stroud et al., 2016) and forms the theoretical framework from which this article draws upon. Ruth Griffiths laid the groundwork for the contemporary understanding of child development with her 1954 conceptualisation of child development based on different domains of development observed during play. Stroud et al. (2016) expanded upon Griffiths's work, presenting a dynamic and systemic model that consists of three spheres: the individual, the individual's interpersonal layer and the socio-environmental context. This model also highlights the five domains of child development. Although these domains are not stand-alone areas of development, they serve as a practical and analytical framework to understand and assess the complex and interconnected nature of child development.

These domains are the Foundations of Learning (a cognitive domain), Language and Communication, Eye and Hand Coordination, Personal-Social-Emotional and Gross Motor Development. This updated model highlights the interconnected nature of development and emphasises the child's development at physiological and neurological levels, influenced by prenatal history. The recent focus on physiological processes and neuroscience aligns with the dynamic relationship between genetic underpinnings and early life events, impacting brain architecture and behaviour development. The updated model stresses that children grow at their own pace and rhythm, acknowledging the individuality of the development (Stroud et al., 2016). The second layer considers interpersonal influences such as family involvement, attachment and self-care. The third layer explores socio-environmental influences, examining the relationship between universal and unique factors like socio-economic conditions, cultural background and political influences shaping a child's development.

Compared to other frameworks like Piaget's stages of cognitive development or Vygotsky's sociocultural theory, Griffiths' framework is more domain specific and assessment oriented and allows for a comprehensive understanding of the diverse factors influencing the developmental trajectory of children and their assessment, including developmental screening (Stroud et al., 2016). Furthermore, a broad-based approach to child development recognises individual development across cognitive, language, fine motor, gross motor and socio-emotional domains. By emphasising interconnected domains and observing child-led activities, assessors can account for children's experiences within diverse sociocultural and economic contexts.

The Griffiths III and its earlier versions are based on Ruth Griffiths's work and are used internationally, including in the United Kingdom (UK), Australia, Portugal, China and South Africa (Stroud et al., 2016). As a broad-domain diagnostic measure, the Griffiths III is considered a gold standard measure in developmental assessment (Lecciso et al., 2025) and has been piloted on a South African sample, demonstrating the relevance of its theoretical framework and potential applicability to other developmental measures.

In the South African context, developmental assessment and, in particular, screening tools must acknowledge the country's unique political, economic and social history (Laher & Cockcroft, 2019). The complexities of South Africa's context pose challenges in psychological assessment and test development, leading to limited efforts to address this process (Laher, 2024). A South African survey conducted 14 years ago highlighted the urgent need for school readiness assessment tools (Engle et al., 2011). It revealed significant disparities between communities with access to professional services and those in poverty. These disparities persist, with further concerns raised by Mcaleni (2025). Mcaleni's survey found that South African psychologists stressed the need for culturally relevant and contextually appropriate assessment tools, as many existing ones are outdated or unsuitable for the diverse cultural and linguistic backgrounds of South African children. Psychologists are increasingly relying on observations from teachers and parents because of the limited availability of standardised assessment materials. While moving beyond a test score is encouraged, a standardised assessment tool remains a crucial starting point for guiding investigations and forms the backbone of psychometric inquiry (Foxcroft & De Kock, 2023). Recognising the importance of assessing children in this diverse cultural setting, there is a need for developmental assessments that are culturally relevant, up to date, affordable and accessible, particularly for the multicultural learning environment where South African children develop (Laher & Cockcroft, 2019).

Despite advancements in the field, significant challenges persist, underscoring the need for innovative approaches. A recent review by Laher (2024), focusing on the development of psychological assessment in South Africa, emphasises the need for developing reliable, valid, affordable and accessible tests through inclusion and innovation. Currently, there are limited tests developed within South Africa. One of the few available developmental measures for assessing preschool children is the Early Learning Measurement (ELOM) (Giese et al., 2023). The aim of ELOM is to assess the school-based outcomes of young children in South Africa, focusing on their progress and the quality of early childhood education. The ELOM addresses one of the key gaps identified by Laher (2024), offering a more contextually appropriate measure for the South African preschool landscape. While the ELOM represents progress, there is still a need to develop more measures to incorporate innovative approaches and ensure that tools remain relevant and applicable in addressing children's current and future needs (Giese et al., 2023). Laher (2024) stresses that the development of local measures should be prioritised to enhance relevance and accessibility. This involves combining indigenous methods together with adaptations of international tests to better suit the local context and improve performance and engagement.

The paucity of South African developed measures, coupled with outdated and culturally biased tests, creates a significant obstacle in identifying and addressing developmental concerns, particularly among South African children in rural and informal settlement areas such as the Eastern Cape. The scarcity of free or cost-effective psychological services further compounds the issue, hindering timely intervention. The need for African innovations in the African context has never been more opportune for development than it is at present. One such innovation leveraging local knowledge to create tools that serve African children is the newly developed screening tool: the PuzzleBox Screener (Jansen & Marais, 2024). The PuzzleBox Screener is a reimagined puzzle made from recyclable material that is multifunctional in its application. Minimal equipment is required, with a single puzzle and a bean bag and recyclable add-on components such as sticks, stones and bottle tops used to test 83 experimental items across five domains. The measure is developed as a normed-referenced assessment; therefore, typically developing children are included to establish essential benchmarks for identifying potential developmental delays. These are children who show age-appropriate physical, cognitive, emotional and social development and who do not have any clinical diagnoses or identified neurodiversity (Foxcroft & De Kock, 2023). The aim of this article is to report on the piloting phase of this tool, focusing on the psychometric performance of items across five domains to inform decisions on retaining, refining or removing items for the PuzzleBox Screener. These include the cognitive, fine motor and emotion–social–moral domains, chosen to ensure broad developmental coverage while enabling an in-depth analysis of the tool's item performance and factor structure.

Method

Research design

The piloting phase of this study followed a cross-sectional design to collect data from the children at a single point in time, capturing a snapshot of their performance on the PuzzleBox Screener (Creswell & Creswell, 2022). A quantitative approach was applied by collecting categorical data through the pass or fail scoring of each item across the five domains of the tool.

Participants

A non-probability, purposive sampling technique was followed to select a sample of $N = 65$ children. As can be seen from Figure 1, the sample was drawn from early childhood development settings, including playschools and preschools, in rural and informal areas within the Raymond Mhlaba Municipality (Bedford, iKhobonqaba 3.1%, KwaMaqoma 52.3%) and Nelson Mandela Bay Municipality (Gqeberha, 44.6%) in the Eastern Cape.

Furthermore, the sample consisted of children from five different home language groups: isiXhosa (46.2%), English (20%), isiZulu (1.5%), Afrikaans (46%) and Shona (24.6%). Despite the linguistic diversity, all participants were proficient in English, Afrikaans and isiXhosa, as 70.8% of the children received educational instruction in English, 24.6% in isiXhosa and 4.6% in Afrikaans. All participants had a typical developmental history (non-clinical population), ensuring a representative and well-rounded sample for the study. The sample consisted of 56.9% boys and 43.1% girls.

The age of the children in the sample ranged from 61 months to 83 months (mean [M] = 74.37, standard deviation [s.d.] = 5.746) or approximately 5.08 years to 6.92 years (M = 6.20, s.d. = 0.48) with a mean age of 74.37 months (s.d. = 5.75), equivalent to 6.20 years (s.d. = 0.48).

Data collection

Data collection tools

The PuzzleBox Screener

The data collection tool used for this study is the PuzzleBox Screener, a newly developed South African multicultural developmental screening tool that informs intervention in

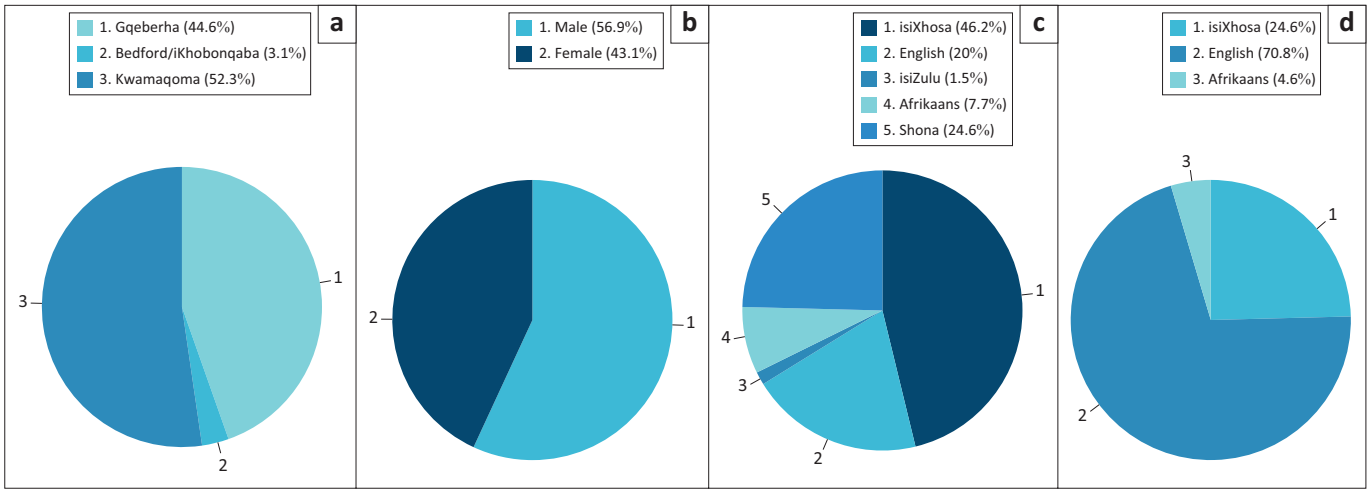

FIGURE 1: Sample characteristics by (a) place of assessment; (b) gender; (c) home language; (d) test language.

multi-domain development before formal schooling. The experimental version of the tool includes 25 cognitive items, 18 language items, 18 fine motor items, 19 emotional-social-moral items and 25 gross motor items. Figure 2 presents a construct map of the PuzzleBox Screener, illustrating the constructs covered within each domain.

These include cognitive (reasoning, memory and learning), language (receptive and expressive communication) and fine motor (manual dexterity, coordination and visual-motor integration). Each key construct comprises several sub-constructs, which are mapped in more detail in Figure 2.

The development process was driven by key economic considerations, including the need for portability, affordability and ease of component replacement, with the aim of reducing costs by 80% compared to imported screening tools. At the same time, societal, cultural and neuroscientific factors were considered to enhance the broader applicability and effectiveness of the measure. The tool takes a more holistic approach by keeping the child's context in mind and including relevant factors extending beyond school-based outcomes. For example, it was designed to assess five broad domains of development, incorporating well-established constructs from existing developmental measures, direct testing of emotional, social and moral functioning and considering environmental and contextual factors in scoring. In addition, the tool was designed for simplicity and ease of administration, allowing for faster, less subjective screening; therefore, binary scoring without partial credit was implemented.

The data collected from the piloting phase informs the next phase, which focuses on the validation and standardisation of the tool. The tool was first translated from English into isiXhosa and Afrikaans, ensuring that the language was both linguistically and culturally appropriate and accurately conveyed. Afterwards, a back-translation process was carried out, where the isiXhosa and Afrikaans versions were translated back into English to verify the consistency and correctness of the original message, ensuring that the meaning was preserved across languages.

Biographical questionnaire

The study also used a biographical questionnaire to gather baseline information on participants' age, gender, location, developmental history, language of instruction, home language and school grade. These data ensured that the sample met the study's parameters and confirmed that the participants were typically developing children with no developmental delays.

Procedure

The procedural steps involved obtaining ethical consent and permission to conduct the study from applicable gatekeepers, including the relevant ethics committee, the Department of Education and the school principals, regarding the study's intent. Formal invitations to participate were extended to educational authorities, school leaders and parents. The children were given a visual story booklet with study information and the opportunity to provide assent before testing. Once informed consent and assent were obtained from the child's guardian and the child, the pilot testing phase began. Testing was administered by a team of psychologists and postgraduate psychology students specially trained for this purpose. The testing sessions were conducted individually in a controlled environment to minimise distractions. These sessions took place on the school premises during the morning hours.

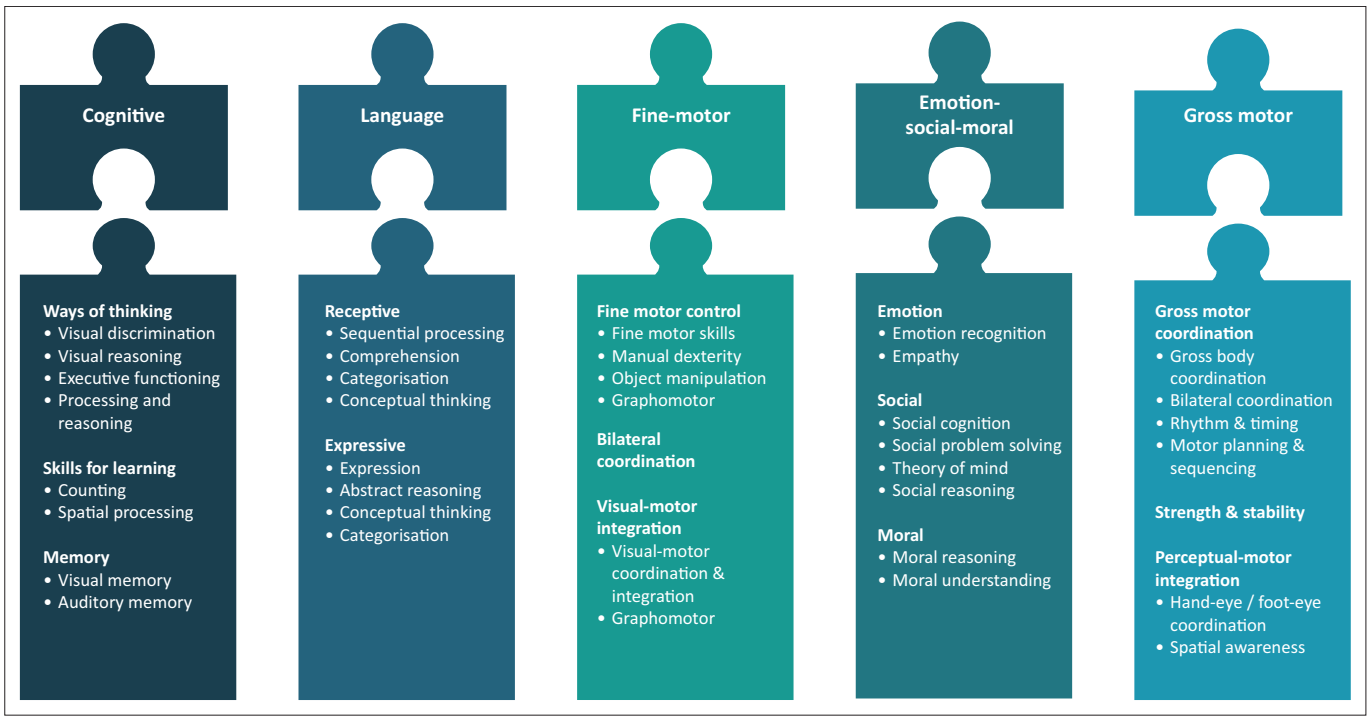

FIGURE 2: Construct map of the PuzzleBox Screener.

Data analysis

A structured approach combining classical test theory (CTT) and item response theory (IRT) was applied to evaluate whether individual test items should be retained, revised or removed. This combination allowed for a comprehensive evaluation of item performance, leveraging the structural insights of CTT and the precision of IRT to inform item retention decisions (El-Hamamsy, 2023). Unlike CTT, IRT provides a more precise and robust basis for decisions on item retention, revision or removal by detailing key characteristics such as discrimination and difficulty (Çetinkaya-Rundel & Hardin, 2021). Additionally, IRT helped identify items that function differently across the two age bands of the screening tool, ensuring fairness and validity across developmental stages. According to Hu et al. (2021, p. 1397), combining CTT and IRT is the 'direction of psychological and educational measurement in the future'. Combining IRT and exploratory factor analysis (EFA) has been applied in other early learning environments. For example, IRT has been used alongside CTT to validate the competent computational thinking test for learners aged 7–11 years, ensuring reliable, developmentally appropriate and gender-fair assessment across multiple grades (El-Hamamsy et al., 2023).

Before running the IRT analysis, an EFA was first conducted to assess the factor loadings of each item and to ensure that the items aligned with the intended constructs. Item response theory, on the other hand, provided detailed item-level information on discrimination and difficulty to guide precise item selection. Item response theory and EFA were preferred to multidimensional models during the piloting because they enabled the evaluation of individual items and accommodated a smaller sample size.

Items with loadings above 0.40 were retained for further analysis, while those with cross-loadings or loadings below 0.32 were considered for removal. This step helped provide an initial indication of item performance. In cases where the two-parameter logistic (2PL) IRT models did not converge, only one-parameter logistic (1PL) models were applied. As a result, for those items, only the difficulty parameters are reported. Overall, items were retained if they demonstrated strong factor loadings (≥ 0.40), good discrimination values ($a \geq 1.50$) and balanced difficulty levels ($-1.0 \leq b \leq 1.0$). Items outside this range were deemed too easy or too difficult (Jabrayilov et al., 2016).

Reliability and validity

Several factors were considered to enhance the reliability and validity of the preliminary findings of the piloting of the tool. As multiple test takers were involved, training sessions were conducted to ensure consistency in administration and scoring. Testing of the tool was supervised by two registered psychologists with expertise in developmental assessment and the tool's development to ensure consistency in administration and scoring. Additionally, a developmental assessment expert consulted on the tool's design, reviewing and validating item content to enhance face validity. The input from a statistician further contributed to the selection of statistical methods to validate, ensure quality and enhance the reliability and validity of the data. To this end, EFA examined the item structure, retaining those with strong factor loadings to ensure conceptual alignment with developmental theory. Additionally, the IRT assessed item discrimination and difficulty, with well-performing items retained. Item response theory also provided item and test information functions, supporting reliable measurement across the tool's intended developmental range.

Ethical considerations

Ethical clearance to conduct this study was obtained from the Pharma-Ethics Independent Research Ethics Committee (No. 240226207).

Informed consent and assent were obtained from legal guardians and children, ensuring ethical participation in the study. The physical records, including the scoring of the children and biographical questionnaires, were collected from testers and securely stored to maintain confidentiality. Access to these records was restricted to the principal investigators and the co-investigator of the study, ensuring that sensitive information remained protected. Additionally, all personal identifiers, such as names, were removed from the dataset to further safeguard participants' identities. Each child was assigned a unique code, with no names allocated to the protocols where scoring took place. Only these codes were recorded for data capturing and analysis, ensuring that the data remained anonymous and confidential.

It would be important to acknowledge the country's unique historical context when considering ethical test development practices to avoid replication of existing discriminatory practices or contributing to further marginalisation. To this end, ethical considerations in the development of the PuzzleBox Screener included designing graphics and content with the lived experiences of South African children in mind, including familiar objects, scenarios and cultural elements. In addition, a context form was designed to be used alongside the tool to flag a lack of prior exposure to puzzles, allowing for a familiarisation phase with practice tasks that helped results reflect developmental ability rather than familiarity.

Results

The results are interpreted using three core metrics: factor loadings, discrimination and difficulty. These metrics provide the basis for determining whether items are retained, revised or removed. Results are presented separately for each domain, specifying the number of items retained, revised and removed, together with statistical evidence informing these decisions. To assist with interpreting these statistics, the following criteria have been applied for item selection and provide descriptive explanations for each metric:

1. Retain: Items were retained when they showed high or ideal properties: factor loadings greater than 0.40, discrimination values above 0.80 and difficulty levels between −1.0 and 1.0. Such items are strong indicators of the construct, effectively differentiate between individuals and are appropriately challenging for most respondents.
2. Revise: Items were flagged for revision when the factor loadings fell between 0.32 and 0.40 and the discrimination values ranged from 0.50 to 0.79. Items were considered slightly easy if their difficulty parameter ranged between −2.0 and −1.0 and slightly difficult if it ranged between 1.0 and 2.0. These items are acceptable but may require adjustment because of moderate discrimination or difficulty that is potentially too easy or too hard.
3. Remove: Items were removed if they demonstrated low or weak psychometric properties: factor loadings below 0.32, discrimination values under 0.50 or extreme difficulty levels ($b < -2.0$ or $b > 2.0$), rendering them unsuitable for reliable scoring.
4. In addition to the above core metrics, communalities served as a supporting measure, indicating how much of an item's variance was explained by the factors (≥ 0.50 = good, 0.30–0.49 = moderate, < 0.30 = weak) and providing further evidence to guide item retention, revision or removal.

Cognitive domain

The psychometric evaluation of the cognitive items, as illustrated in Table 1, indicated that 11 of the 20 items showed acceptable to strong factor loadings (0.40 to 0.77), good to high discrimination (1.650 to 3.650) and difficulty levels ranging from slightly easy to ideal (–1.911 to 0.101), indicating strong measurement properties.

Four items (Cog3, Cog11, Cog12 and Cog13) demonstrated moderate factor loadings (0.33 to 0.36) and lower communalities (0.12 to 0.14), with Cog3 and Cog11 being too easy, Cog12 showing good difficulty (0.513) and Cog13 displaying poor difficulty balance (–0.089). These items were therefore earmarked for revision.

Five items are recommended for removal because of their weak psychometric properties. Item 20 (Cog20) exhibited poor discrimination (0.527), indicating limited differentiation ability, while item 2 (Cog2), item 17 (Cog17), item 18 (Cog18) and item 19 (Cog19) had psychometric properties that were unacceptable across all parameters.

Language domain

Results for the language items, as illustrated in Table 2, indicated that 8 of the 17 items demonstrated moderate to strong factor loadings (0.42 to 0.86), with item 3 (Lang3) (2.479) and item 10 (Lang10) (2.040) displaying good discrimination and difficulty levels ranging from slightly easy to ideal, supporting their retention (–1.868 to 0.880).

Three items (Lang1, Lang6 and Lang13) showed moderate factor loadings (0.36 to 0.38) and higher difficulty levels (–1.747 to 1.821). Six items (Lang2, Lang4, Lang8, Lang11, Lang16 and Lang17) are recommended for removal because of weak psychometric properties, including low factor loadings and low communalities.

Fine motor domain

Assessment of the fine motor items, illustrated in Table 3, indicated that 7 of the 14 items demonstrated moderate to

TABLE 1: Cognitive domain: Factor loadings, item discrimination and item difficulty.

Number	Item description	Item number	Factor loading	Communality	Discrimination		Item difficulty	
					Coeff	*p*	Coeff	*p*
65	Build Puzzle < 10 min	Cog1	0.77	0.60	-	-	-1.120	< 0.001
65	Count 10 oranges	Cog4	0.52	0.28	3.650	0.125	-1.702	< 0.001
65	Puzzle piece to left	Cog5	0.56	0.52	-	-	-0.199	0.376
65	Puzzle piece to right	Cog6	0.57	0.43	1.980	0.005	-0.639	0.006
65	Direction Up/down	Cog7	0.61	0.37	1.726	0.006	-0.737	0.005
65	Visual match	Cog8	0.58	0.37	2.824	0.030	-1.475	< 0.001
65	Counting	Cog9	0.52	0.55	3.222	0.030	-1.192	< 0.001
65	More and Less	Cog10	0.43	0.20	1.706	0.046	-1.911	0.001
65	Copy pattern	Cog14	0.66	0.44	2.586	0.022	-1.257	< 0.001
65	Count 15 bottle tops	Cog15	0.50	0.26	1.653	0.009	-1.161	0.001
65	Count bottle tops backwards	Cog16	0.40	0.21	1.726	0.008	0.101	0.644
65	Which wheel is the biggest	Cog3	0.34	0.12	-	-	-3.400	< 0.001
65	Find all 5 red apples	Cog11	0.36	0.14	-	-	-1.718	< 0.001
65	Growth Pumpkin	Cog12	0.33	0.12	-	-	0.513	0.028
65	Differences between grannies	Cog13	0.35	0.14	-	-	-0.089	0.688
65	Fit puzzle pieces back	Cog2	-	0.02	-	-	-	-
65	Which stick is longer	Cog17	-	-	-	-	-	-
65	Which stick is shorter	Cog18	-	-	-	-	-	-
65	Which bottle top is heavier	Cog19	-	-	-	-	-	-
65	Which cost more	Cog20	-	0.07	0.527	0.12	0.998	0.193

Coeff, coefficient.

strong factor loadings (0.39 to 0.70), good to very strong discrimination (1.121 to 2.575) and difficulty levels ranging from slightly easy to ideal (–1.411 to –0.343), supporting their retention (Fine1, Fine7, Fine8, Fine10, Fine11, Fine12 and Fine14).

Two items (Fine9 and Fine13) showed moderate factor loadings (0.36 and 0.39) and borderline or sufficient discrimination, with item 9 (Fine9) having ideal difficulty (0.075) and item 13 (Fine13) slightly easy (–1.796) but reasonable and were therefore earmarked for revision. Five items (Fine2, Fine3, Fine4, Fine5 and Fine6) are recommended for removal because of their weak psychometric properties, exhibiting low factor loadings, poor discrimination and unacceptable difficulty levels.

Emotion–social–moral domain

The emotion–social–moral items, as illustrated in Table 4, indicate that 10 of the 17 demonstrated moderate to strong factor loadings (0.39 to 0.84), good to high discrimination (1.189 to 3.728) and difficulty levels ranging from slightly easy to ideal (–1.693 to 0.259), supporting their retention. These items highlight the tool's ability to assess children's capacity to understand and interpret social cues, recognise emotions in others and navigate social interactions effectively.

Two items were identified for revision because of factor loadings slightly below the retention threshold and lower communality. Item 1 (Emo1) had a factor loading of 0.37, lower communality (0.14) and weaker discrimination (0.907) than ideal, although still acceptable, with ideal difficulty (–0.645). Item 4 (Emo4) demonstrated a moderate factor loading (0.37) and lower communality (0.17) but showed fine discrimination (1.086) and ideal difficulty (–0.795).

A total of five items (Emo8, Emo9, Emo13, Emo14 and Emo15) are recommended for removal because of their inadequate measurement properties. Notably, Emo14 demonstrated extreme difficulty (–2.31), making it unsuitable for the intended age range.

TABLE 2: Language domain: Factor loadings, item discrimination and item difficulty.

Number	Item description	Item number	Factor loadings	Communality	Discrimination		Item difficulty	
					Coeff	*p*	Coeff	*p*
65	Memory: items on list	Lang3	0.86	0.79	2.479	0.138	0.806	0.002
65	What do you know about?	Lang5	0.43	0.30	-	-	-0.735	0.006
65	Find fruit	Lang7	0.53	0.19	-	-	0.677	0.010
65	One step instruction	Lang9	0.49	0.05	-	-	-1.868	< 0.001
65	Two step instruction.	Lang10	0.50	0.24	2.040	0.069	0.414	0.058
65	Repetition of sentence	Lang12	0.42	0.14	-	-	0.880	0.002
65	Repetition longest sentence	Lang14	0.47	0.14	-	-	0.183	0.450
65	What is a ... (two features)	Lang15	0.49	0.22	-	-	-1 436	< 0.001
65	Story about puzzle	Lang1	0.36	0.18	-	-	-1.747	< 0.001
65	Similarities	Lang6	0.38	0.27	-	-	0.065	0.788
65	Repetition of longer sentence	Lang13	0.38	0.18	1.451	0.076	1.821	0.007
65	Verbs	Lang2	-	0.16	-	-		-
65	Opposites	Lang4	-	0.15	-	-		-
65	Day / Night	Lang8	-	0.39	-	-		-
65	Three step instruction	Lang11	-	0.28	-	-		-
65	What is a fire? (two features)	Lang16	-	0.25	-	-	-	-
65	What is a spoon?	Lang17	-	0.17				

Coeff, coefficient.

TABLE 3: Fine motor domain: Factor loadings, item discrimination and item difficulty.

Number	Item description	Item number	Factor loadings	Communality	Discrimination		Item difficulty	
					Coeff	*p*	Coeff	*p*
65	Weave baskets to carry fruit	Fine1	0.44	0.19	1.232	0.012	-0.658	0.036
65	Make fence with sticks in holes	Fine7	0.47	0.22	1.289	0.020	-1.411	0.003
65	Copy gate pattern	Fine8	0.39	0.15	1.121	0.015	-0.343	0.241
65	Draw a bicycle	Fine10	0.79	0.49	1.981	0.013	-0.519	0.022
65	Draw line within rectangle	Fine11	0.69	0.48	2.575	0.016	-0.524	0.009
65	Draw on puzzle	Fine12	0.46	0.21	2.519	0.052	-1.334	< 0.001
65	Write your name	Fine 14	0.55	0.30	1.641	0.018	-1.234	0.001
65	Copy light	Fine9	0.36	0.13	0.943	0.026	0.075	0.810
65	Draw straight line	Fine13	0.39	0.15	1.087	0.043	-1.796	0.012
65	Remove bottle top & screw back	Fine2	-	0.02	-	-	-	-
65	Thread 6 bottle tops	Fine3	-	-	-	-	-	-
65	Thread 10 bottle tops	Fine4	-	-	-	-	-	-
65	Stone in bottle tops	Fine5	-	0.08	-	-	-	-
65	Balance 10 stones on bottle tons	Fine6	-	0.06	-	-	-	-

Coeff, coefficient.

Gross motor domain

As can be seen in Table 5, the gross motor domain showed the weakest measurement properties, with 4 of the 14 items being retained. Item 5 (Gross5) had a strong factor loading (0.61) and very high discrimination (1.866) although its difficulty was slightly below the ideal range (–1.319). Item 7 (Gross7) demonstrated a moderate factor loading (0.36) with acceptable discrimination (1.125) and difficulty (0.070), while item 9 (Gross9) had an acceptable factor loading (0.43) and acceptable discrimination (1.242), and its difficulty fell within the ideal range of −1.0 to 1.0. Item 12 (Gross12) showed a strong factor loading (0.51), very high discrimination (1.752) and difficulty within the ideal range.

Items flagged for revision included item 8 (Gross8), which had a factor loading below the threshold, poor communality, discrimination close to the acceptable threshold (0.716) and difficulty within the ideal range. Similarly, item 3 (Gross3) had a factor loading below the threshold, poor communality, discrimination slightly below the acceptable threshold (0.409) and difficulty within the ideal range (–0.170).

Eight items (Gross1, Gross2, Gross4, Gross6, Gross10, Gross11, Gross13 and Gross14) performed poorly across all evaluated parameters, including factor loading, discrimination and difficulty.

Discussion

Overall, the psychometric evaluation across the cognitive, language, fine motor, emotion–social–moral and gross motor domains indicated that a substantial number of items demonstrated acceptable to strong factor loadings, good to high discrimination and difficulty levels ranging from slightly easy to ideal, supporting their retention. Several items in each domain showed moderate factor loadings, lower communalities or borderline difficulty and discrimination and were therefore earmarked for revision. A subset of items across all domains exhibited weak or unacceptable psychometric properties, including low factor loadings, poor discrimination and extreme difficulty, and were recommended for removal. Subsequent refinement efforts will prioritise increasing the proportion of easier items to optimise the tool's sensitivity for identifying children at risk.

TABLE 4: Emotion–social–moral domain: Factor loadings, item discrimination and item difficulty.

Number	Item description	Item number	Factor loadings	Communality	Discrimination		Item difficulty	
					Coeff	*P*	Coeff	*P*
65	Feelings	Emo2	0.45	0.21	1.189	0.010	0.259	0.348
65	Problem solving	Emo3	0.55	0.31	1.429	0.007	-0.125	0.602
65	Moral: Drop oranges	Emo5	0.39	0.25	1.205	0.016	-1.208	0.005
65	Social: Ask permission	Emo6	0.63	0.40	2.007	0.006	-0.162	0.428
65	Social: good or bad?	Emo7	0.42	0.27	3.728	0.290	-1.693	<.001
65	Problem solving: One cold drink left	Emo10	0.69	0.49	-	-	-0.300	0.123
65	Kindness	Emo11	0.84	0.70	-	-	0.033	0.863
65	Perspective Taking	Emo12	0.48	0.30	-	-	-0.156	0.415
65	Problem solving: bumped into lady	Emo16	0.65	0.43	2.932	0.034	-0.676	0.001
65	Self description	Emo17	0.46	0.22	1.198	0.016	-1.129	0.007
65	Social:sharing	Emo4	0.37	0.17	1.086	0.015	-0.795	0.029
65	Identification feelings	Emo1	0.37	0.14	0.907	0.023	-0.645	0.095
65	Moral: Good or bad	Emo8	-	0.09	-	-	-	-
65	Robbing the shop. Is this good or bad.	Emo9	-	0.11	-	-	-	-
65	Social: helpful behaviour	Emo13	-0.34	0.18	-	-	-	-
65	Empathy	Emo14	0.35	0.29	0.566	0.150	-2.310	0.127
65	Best friend	Emo15	-	0.09	-	-	-	-

Coeff, coefficient.

TABLE 5: Gross motor domain: Factor loadings, item discrimination and item difficulty.

Number	Item description	Item number	Factor loadings	Communality	Discrimination		Item difficulty	
					Coeff	*p*	Coeff	*p*
65	Throw bean bag underarm	Gross5	0.61	0.37	1.866	0.070	-1.319	0.002
65	Throw bean bag into air and catch it again	Gross7	0.36	0.13	1.125	0.029	0.070	0.808
65	Balance bean bag on wrist > 115sec	Gross9	0.43	0.19	1.242	0.037	-0.885	0.024
65	Walk like granny	Gross12	0.51	0.26	1.752	0.035	-0.737	0.110
65	Bear crawl, straight line 3m	Gross3	-	0.07	0.409	0.272	-0.170	0.800
65	Balance bean bag on head > 1	Gross8	-	0.06	0.716	0.086	-0.633	0.211
65	Jump like a frog	Gross1	-	0.02	0.162	0.658	-3.841	0.660
65	Jum over river	Gross2	-	0.02	0.393	0.597	-7.662	0.582
65	Throw bean bag	Gross4	-	0.08	0.563	0.206	-2.254	0.177
65	Catch the bean bag with both hands	Gross6	-	0.00	-0.120	0.730	4.552	0.731
65	Balance bean bag on foot 15sce	Gross10	-	0.07	0.567	0.305	-3.780	0.262
65	Bowling 3 bottles.	Gross11	-	-	-	-	-	-
65	March like boy	Gross13	-	-	-	-	-	-
65	Fly like bird	Gross14	-	-	-	-	-	-

Coeff, coefficient.

The literature supports this approach, as exemplified by the Denver Developmental Screening Test II (Denver II), which includes items completed by 75% – 90% of children (Frankenburg et al., 1990). Using an IRT approach affords the PuzzleBox Screener the ability to strategically include items that are both easy and highly discriminative, increasing the likelihood of detecting at-risk individuals while preserving specificity. This is facilitated by IRT, which identifies items that offer the most information at the lower end of the ability spectrum, the range most relevant for detecting delays or risk.

Initial findings also highlighted the need for flexibility in language items to accommodate regional and dialectal differences and the inclusion of classroom-based socio-emotional competencies via the teacher-report form. Additionally, a context checklist was also developed to provide clearer descriptions of each participant's background in preparation for norming.

While the tool has shown potential, the variability in developmental skill levels within the 5 years to 6 years and 11 months age range remains an important consideration. This period is marked by significant cognitive and developmental changes resulting from individual factors, such as neural plasticity and ongoing brain development, as well as environmental factors linked to exposure and exploration (Gualtieri & Finn, 2022). This variability in developmental skill levels further underscores that a structured statistical approach alone for the retention, revision and removal of items is insufficient. While such an approach is critical in contributing to the psychometric properties of the items, it highlights the necessity of achieving a fine balance between theoretical and psychometric soundness. In the cognitive domain, for example, item 11 is likely to be retained as a precursor to an incidental memory task, as it will contribute to the functionality of subsequent memory items and allow the items to be more clearly assimilated into the child's cognitive schemas. Although item 11 does not meet the statistical criteria for retention, it may ultimately be included without revision. In this case, theoretical consideration takes precedence, underscoring that test development is both science and art. Furthermore, the strong loadings for key items across different factors in each domain emphasise the tool's ability to assess diverse aspects of children's development. Utilising a broad-domain approach aligns with Ruth Griffiths's Avenues of Learning Model, which underscores the complex interaction of various factors and avenues of learning in child development (Stroud et al., 2016). Exploring these areas of development provides valuable insights into the developmental trajectories of preschool children, where socio-economic challenges often influence outcomes and lay the foundation for later academic and social success (Zhang et al., 2024). The ability to assess a wide range of developmental domains using a single reimagined, recyclable puzzle with limited equipment further offers a valuable resource for early childhood educators and healthcare professionals working with young children in diverse South African communities. This makes the PuzzleBox Screener particularly relevant and potentially feasible as a broad-domain-based screening tool within the South African context. Furthermore, this article acknowledges the significant gap in the availability of locally developed, comprehensive developmental assessments in South Africa because of cost and resource limitations. The preliminary results from the PuzzleBox Screener thus align with the global call for developmental tools that are not only scientifically valid but also suitable for diverse socio-economic environments (Faruk et al., 2020; Laher, 2024). The development of such tools requires re-imagination to suit the needs of children, challenging the reliance on Western-designed tools that may not fully capture their developmental profiles in these contexts. This is in line with Faruk et al. (2020), who noted the scarcity of research from LMICs on cost-effective, low-resource methods to identify children with developmental delays.

Limitations and future directions

The relatively small sample size remains a limitation in terms of the statistical power. Nevertheless, the EFA as well as IRT analysis identified problematic items for all domains, such as those with poor factor loadings or those that were too easy, which will inform the further development, piloting and final selection of items for the tool. In addition, although the sample size was small it provided preliminary evidence of the tool's potential to assess multiple developmental domains and distinguish between different performance levels. Another limitation was that, at the time of piloting, the tool had only been translated into three languages. Consequently, the study did not assess children from all five home language groups represented in the sample, focusing only on those who received educational instruction in isiXhosa, Afrikaans or English. Furthermore, the results are preliminary, with further research needed to validate the factor structure, confirm the tool's reliability and validity and strengthen its psychometric evidence. The planned steps include using psychometric data from the study to evaluate each item's theoretical fit, developmental relevance and direct connection to the puzzle equipment. This will ensure the statistical and theoretical alignment for the final test selection. Following the final test selection, the subsequent phase will involve implementing the PuzzleBox Screener on a larger scale, adapting it for digital administration and simplifying the tool by removing the items pertaining to the gross motor domain and consolidating the remaining domain items into a single item set. This approach will focus on generating a global score with cut-off points to identify children at risk rather than interpreting scores at the domain level. Nevertheless, the tool will retain a broad-based domain structure in its conceptual design and will provide qualitative information across the four developmental domains, offering meaningful insight into children's strengths and areas of concern.

Conclusion

The piloting phase of the PuzzleBox Screener provides support for its further development and standardisation within the South African preschool context. The tool's ability to assess a broad range of developmental domains demonstrates its capacity to identify meaningful performance differences. Its simplicity and relevance to the local context make it a promising option as part of locally appropriate solutions for preschool assessment in South Africa. Using modern measurement theory to inform item selection and enhance the PuzzleBox Screener design may represent an effective approach for other test developers wanting to develop a precise and contextually relevant assessment tool.

This article supports the need for innovative, locally developed assessment tools that reflect the realities of South African children and their communities. Results from the piloting phase align with broader calls for such tools, stressing the feasibility and potential of the PuzzleBox Screener to contribute meaningfully to the South African early childhood assessment landscape, particularly in the Eastern Cape (Dawes et al., 2018; Laher, 2024). However, future validation and standardisation efforts, including the establishment of precise cut-off scores, will be needed to refine the PuzzleBox Screener and ensure its applicability within the South African context.

Acknowledgements

The authors would like to acknowledge Prof. Leon van Niekerk for the preliminary statistical analysis of the first round of the piloting data and Dr Kirsty Donald for the second round of statistical analysis. We would also like to acknowledge Prof. Cheryl Foxcroft, who served as a consultant on the development of the PuzzleBox Screener.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Authors' contributions

R.M. and J.M.J. conceived and wrote the article. Both authors made substantial contributions to the acquisition, analysis and interpretation of data. They have both reviewed and approved the final version of the article and agree to be accountable for all aspects of the work, ensuring the accuracy and integrity of the content.

Funding information

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Data availability

The datasets analysed in this study are available from the corresponding author, R.M., upon reasonable request.

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. They do not necessarily reflect the official policy or position of any affiliated institution, funder, agency or that of the publisher. The authors are responsible for this article's results, findings and content.

References

Bagnato, S.J., Macy, M., Dionne, C., Smith, N., Robinson Brock, J., Larson, T., Londono, M., Fevola, A., Bruder, M.B., & Cranmer, J. (2024). Authentic assessment for early childhood intervention: In-vivo & virtual practices for interdisciplinary professionals. *Perspectives on Early Childhood Psychology and Education, 8*(1), 43–74. https://doi.org/10.58948/2834-8257.1066

Çetinkaya-Rundel, M., & Hardin, J. (2021). *Introduction to modern statistics*. OpenIntro.

Creswell, J.W., & Creswell, J.D. (2022). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage.

Dawes, A., Biersteker, L., Girdwood, E., Snelling, M., & Tredoux, C.G. (2018). Early learning assessment innovation in South Africa: A locally appropriate monitoring tool. *Childhood Education, 94*(1), 12–16. https://doi.org/10.1080/00094056.2018.1420358

El-Hamamsy, L., Zapata-Cáceres, M., Martín-Barroso, E., Mondada, F., Zufferey, J.D., Bruno, B., & Román-González, M. (2023). The competent computational thinking test: A valid, reliable, and gender-fair test for longitudinal CT studies in grades 3–6. Tech Know Learn, 30, 1607–1661. https://doi.org/10.1007/s10758-024-09777-8

Engle, P.L., Fernald, L., Walker, S., Wachs, T., Black, M., McGregor, S., & the Global Child Development Group. (2011). Strategies for reducing inequalities and improving developmental outcomes for young children in low-income and middle-income countries. *Lancet, 378*(9799), 1339–1353. https://doi.org/10.1016/S0140-6736(11)60889-1

Faruk, T., King, C., Muhit, M., Islam, M.K., Jahan, I., ul Baset, K., Badawi, N., & Khandaker, G. (2020). Screening tools for early identification of children with developmental delay in low- and middle-income countries: A systematic review. *BMJ Open, 10*(11), e038182. https://doi.org/10.1136/bmjopen-2020-038182

Frankenburg, W.K., Dodds, J., & Archer, P. (1990). *Denver II technical manual*. Denver Developmental Materials, Inc.

Foxcroft, C., & De Kock, F. (2023). *Psychological assessment in South Africa: An introduction* (6th ed.). Oxford University Press.

Green, E.M., Stroud, L., & Marx, C., Cronje, J. (2020). Child development assessment: Practitioner input in the revision for Griffiths III. *Child: Care, Health and Development, 46*(6), 682–691. https://doi.org/10.1111/cch.12796

Giese, S., Dawes, A., Biersteker, L., Girdwood, E., & Henry, J. (2023). Using data tools and systems to drive change in early childhood education for disadvantaged children in South Africa. *Children (Basel), 10*(9), 1470. https://doi.org/10.3390/children10091470

Gualtieri, S., & Finn, A.S. (2022). The sweet spot: When children's developing abilities, brains, and knowledge make them better learners than adults. *Perspectives on Psychological Science, 17*(5), 1322–1338. https://doi.org/10.1177/17456916211045971

Hu, Z.F., Lin, L., Wang, Y.H., & Li, J.W. (2021). The integration of classical testing theory and item response theory. *Psychology, 12*, 1397–1409. https://doi.org/10.4236/psych.2021.129088

Jabrayilov, R., Emons, W.H.M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement, 40*(8), 559–572. https://doi.org/10.1177/0146621616664046

Jansen, J., & Marais, R. (2024). The Puzzle Box. *Oral Presentation at the Faculty of Health Sciences Research Conference*. Port Elizabeth: Nelson Mandela University.

Laher, S. (2024). Advancing an agenda for psychological assessment in South Africa. *South African Journal of Psychology, 54*(4), 515–528. https://doi.org/10.1177/00812463241268528

Laher, S., & Cockcroft, K. (2017). Moving from culturally biased to culturally responsive assessment practices in low-resource, multicultural settings. *Professional Psychology: Research and Practice, 48*(2), 115–121. https://doi.org/10.1037/pro0000102

Laher, S., & Cockcroft, K. (2019). *Psychological assessment in South Africa: Research and applications*. Wits University Press.

Lecciso, F., Martis, C., & Levante, A. (2025). The use of Griffiths III in the appraisal of the developmental profile in autism: A systematic search and review. *Brain Sciences, 15*(5), 506. https://doi.org/10.3390/brainsci15050506

Mcaleni, A. (2025). South African psychologists' view of screening assessment needs for early childhood development. Unpublished master's thesis. University of Fort Hare, South Africa.

National Association for the Education of Young Children. (2019). *Advancing equity in early childhood education: A position statement of the NAEYC*. Retrieved from https://www.naeyc.org/resources/position-statements/equity

Stroud, L., Foxcroft, C., Green, E., Bloomfield, S., Cronje, J., Hurter, K., Lane, H., Marais, R., Marx, C., McAlinden, P., O'Connell, R., Paradice, R., & Venter, D. (2016). *Griffiths scales of child development 3rd edition; Part I: Overview, development and psychometric properties*, Hogrefe.

Zhang, L., Sewanyana, D., Martin, M.C., Lye, S., Moran, G., Abubakar, A., Marfo, K., Marangu, J., Proulx, K., & Malti, T. (2021). Supporting child development through parenting interventions in low- to middle-income countries: An updated systematic review. *Frontiers in Public Health, 9*, 671988. https://doi.org/10.3389/fpubh.2021.671988

Zhang, Y., Zhang, J., & Zhang, L. (2024). Understanding the role of socio-emotional development in early childhood education. *Journal of Early Childhood Development, 35*(2), 98–105. https://doi.org/10.1111/jecd.12456