BMC Psychology

RESEARCH Open Access

Validation of the English version of the TOY8 developmental screening tool: examining measurement invariance across languages, gender and income groups

Su Woan Wo[1], Ponmalar N. Alagappar[2*], Amira Najiha Yahya[3] and Pei Jun Woo[1]

Abstract

Background The National Health and Morbidity Survey in Malaysia (2022) revealed a significant increase in developmental delays among young children. Early detection using valid, accessible, and cross-culturally appropriate developmental screening tools is essential. Thus, English-language and Malay versions of the TOY EIGHT developmental screening tool (TOY8) were developed using artificial intelligence and a standardized parent-proxy questionnaire. This study aimed to examine the construct validity and reliability of the English version of TOY8, building on the previously validated Malay TOY8, and to examine measurement invariance across language versions, gender, and income groups.

Methods TOY8 was designed and developed to screen for developmental problems in children aged 3–5 years in Malay and English by an interdisciplinary research team drawing upon both national and international guidelines, and then reviewed by an expert panel (n = 5). Two samples of parents and their children were recruited: 1767 dyads to complete the English TOY8 and another 1724 dyads to complete the Malay TOY8.

Results The confirmatory factor analysis results indicated that the model structure of the English TOY8 matched that of the Malay TOY8. The split-half reliability coefficient indicated adequate to high reliability, which is also consistent with the Malay TOY8. Our results showed that all configural and metric invariance models across groups had a good fit to the data, demonstrating that multiple-group confirmatory factor analysis was appropriate. Finally, scalar invariance was only achieved in certain domains across gender and not in language versions or income groups.

Conclusion The English TOY8 demonstrates construct validity and reliable screening tool for identifying developmental milestones in children aged 3–5 years in Malaysia. In addition, configural and metric invariances across groups in all domains were established, indicating the cross-cultural equivalence of the items, and scalar invariance was established across genders in most 3- to 5-year-old domains. These findings provide preliminary evidence supporting reliability and validity that aligns with previous literature on child development, which indicates a general

*Correspondence:
Ponmalar N. Alagappar
ponmalar.a@um.edu.my

Full list of author information is available at the end of the article

similarity in the gender and cross-cultural development domains in the first years of life, but not for older children, in terms of language and socioemotional skills.

Keywords Developmental milestones, Screening tool, Validity, Reliability, Measurement invariance, Preschool children, Malaysia

Background

According to the Institute of Public Health, Malaysia [1], 7.4% of children younger than five years of age experience delays in reaching their expected developmental skills, compared to only 2.8% in 2016. One possible explanation for this increase is the significant impact of the COVID-19 pandemic which led to school closures and disruptions to regular healthcare services and highlighted the crucial need for early detection and intervention strategies [2].

Efforts to address these delays include raising awareness among parents and childcare providers and the possible collaboration of public health institutions (for example, Malaysia Ministry of Health), educational authorities, and government agencies with private sector, non-governmental organizations, startups, and other agencies to develop an effective strategy for reducing developmental delays among children [1]. Arumugam and Hock [3] supported the National Health and Morbidity Survey recommendation that early childhood education (ECE) educators acknowledge that they or the parents may overlook developmental delays in children because they are not aware, do not have knowledge, and/or may dismiss such delays as behavioral issues.

Developmental screening tools have emerged as primary resources enabling researchers to monitor children's development and identify developmental issues at an early stage [4, 5]. However, significant disparities persist worldwide in the access to and quality of services that aim to support optimal development in young children. The Sustainable Development Goals 4.2 of the United Nations [6] emphasize the importance of assessing access to and quality of early childhood care, development, education services, and early childhood development for all children [7].

Consequently, national efforts have been made to improve early childhood development and help adolescents grow healthily in Malaysia. The Ministry of Education and the National Child Development Research Centre have taken the initiative to introduce a comprehensive Developmental Monitoring Checklist to track the growth and progress of children from one month to six years of age. This checklist is a valuable tool for parents, caregivers, and educators to determine whether their children are likely to reach the expected developmental milestones during the critical early years.

However, these developmental tools have limitations; they are costly and take time to administer, making them impractical for use. The majority of the tools are based on the Western cultural context and focus mainly on motor, cognitive, and language development, while neglecting other crucial components such as social and emotional development [8]. Although a large body of literature on cross-cultural child development has shown similarities in all developmental domains in the first five years of life [9], language, speech, and socioemotional skills are largely culturally specific [10]. Despite these initiatives, it is crucial to highlight that none of the developmental tools introduced by the government have incorporated artificial intelligence (AI) technology, which has the potential to enhance accuracy and efficiency.

Cultural and linguistic diversity in Malaysia

Malaysia is known for its cultural and linguistic diversity. Malay is the official language that is widely spoken across the country, but English plays a significant role, especially in the education and business sectors. Many Malaysians are multilingual, with regional languages such as Mandarin, Tamil, and various indigenous languages that are commonly spoken. The Malaysian educational system promotes bilingualism, ensuring that most children grow up in multiple languages, typically Malay and English [11].

The government supports linguistic diversity through policies that encourage the use of Malay and English in education. The Upholding the Malay Language and Strengthening the English Language policy emphasizes the importance of mastering both languages for global competitiveness while maintaining cultural identity [12]. Consequently, many students in Malaysia are proficient in both languages, with English often serving as the second language that facilitates access to global knowledge and opportunities [12].

In practice, the bilingual nature of the education system ensures that children develop proficiency in Malay while also becoming fluent in English, which is critical for academic and professional success in today's globalized world. Additionally, regional languages continue to play a significant role in maintaining the cultural heritage, contributing to Malaysia's linguistic richness [13].

In the present study, we developed a tool in both Malay and English to accommodate the linguistic diversity in Malaysia. Malay is commonly used in households, schools, and government institutions, as is English, particularly in urban areas, schools, and families with multicultural or expatriate backgrounds. Offering both

versions ensured inclusivity, allowing participants to use their preferred language and enhancing the tool's accessibility and effectiveness across Malaysia's diverse population.

Overview of the TOY8 Development Screening Tool

The current limitations of existing developmental tools highlight the need for a simple, user-friendly, and effective screening tool capable of identifying developmental delays in children aged 3–5 years. To address this issue, Toy Eight, an AI-backed Edutech start-up from Japan, together with Universiti Malaya and Sunway University, developed the TOY8 developmental screening tool for children aged 3–5 years. The TOY EIGHT team, together with AI specialists, ingeniously transformed conventional face-to-face developmental screening into a digital screening system. This digital screening tool was transformed into a fun game made available through a smartphone. This simplified screening procedure is familiar and easy to use, and can assist parents and educators in understanding and learning about children's developmental stages. This developmental screening alerts parents and educators to potential delays in development in accordance with their age, that is, 3–5 years when scores are lower than the standard norm. An additional advantage is that the TOY8 development screening tool kit is portable and enables screening to be performed anytime, anywhere, and without a specialist.

AI-based developmental screening assessments provide objective and data-driven insights into children's cognitive, physical, and socio-emotional development. These insights can be used to identify areas where additional support or interventions may be crucial for child development. Achievement gaps and disparities in educational outcomes are persistent concerns and challenges in Malaysia [1]. AI assessment is a potential tool for mitigating these concerns, as it can aid in identifying children who might lag behind their peers in specific developmental domains, enabling targeted support or interventions to be provided in a timely manner.

Importance of measurement invariance across languages, gender and income groups

A recent review by the World Health Organization [14] reported that research findings on the attainment of developmental milestones by children of different ages, genders, and cultures across countries are inconclusive. One of the major reasons for this is the variety of methodologies and the lack of psychometrically sound instruments, especially in low- to middle-income countries [14, 15]. A recent cross-sectional study investigating the early childhood development of 5,000 children aged 0–3.5 years old from low- to middle-income countries revealed that most developmental milestones were similar across genders and countries in their first year of life [9]. Similar findings were reported in another cross-cultural study conducted in Germany and India [16]. Notably, these studies revealed differences in socioemotional (e.g., play) and language milestones (e.g., receptive language) across countries later in life.

In addition, the age at which milestones are attained is strongly associated with the timing of environmental exposure. The authors speculated that these domains are difficult to examine and are highly dependent on parents' expectations and perceptions of their children's comprehension levels [9]. These studies concluded that as children grow older, the influence of cultural and environmental factors on developmental milestones increases. Although previous research has focused on children's attainment of milestones between the ages of zero and three years, there is a notable gap in studies and data on children aged 3–5 years. Therefore, there is a pressing need to investigate measurement invariance across groups in Malaysia, particularly language, gender, and income groups, by employing a psychometrically robust instrument.

Research objective

The Malay version of the TOY8 developmental screening tool underwent initial testing (an early stage process where a new tool is systematically evaluated) to assess its construct validity and reliability using both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The sample size included 400 children for each age group (3–3.99, 4–4.99, and 5–5.99 years) for the EFA. Similarly, 500 children per age group (3–3.99, 4–4.99, and 5–5.99 years) were recruited for the CFA. The results of this analysis are currently being reviewed for publication in a scholarly journal.

This study aimed to examine the construct validity and reliability of the English version of the TOY8 developmental screening tool and to assess measurement invariance across language, gender, and income groups within the TOY8 developmental screening tool. Ensuring measurement invariance across language, gender and income groups is crucial for ensuring cultural appropriateness, language equivalence, validity, and generalizability of research findings when administering both the English and Malay versions of the TOY8 development screening tool to children aged three years zero months to five years 11 months 30 days in Malaysia.

Methodology

TOY EIGHT development

The TOY EIGHT developmental screening tool (TOY8) is an AI tool combined with a standardized parent-proxy questionnaire that was designed as an objective measure to assess specific developmental aspects in children aged

three years zero months to five years 11 months 30 days. The TOY8 screening tool was designed and developed in Malay and English. A validation study of both language versions was conducted in Kuala Lumpur and Selangor, Malaysia.

The screening tool was developed using a structured process. First, the research team (developmental psychologists and psychometricians) identified the key developmental milestones in the tool. Currently, there is no developmental screening tool standardized according to Malaysian norms. To ensure that the tool was comprehensive and aligned with existing research and national and international guidelines, developmental milestones were based on the Pediatrics Protocol for Malaysian Hospitals 4th Edition [17], Fernald et al.'s [18] guidelines, Singapore Health Booklet 2014 [19], and the Centers for Disease Control and Prevention (CDC)'s Developmental Milestones [20]. In addition, we referred to established developmental assessment tools that are widely used in Malaysia, such as the Mullen Scales of Early Learning [21], Griffiths Mental Development Scales [22], and Malaysia Developmental Language Assessment Kit [23], to identify key developmental milestones as a foundation for creating the original items.

Five important domains and their subdomains of development for children aged 3 years 0 months to 5 years 11 months 30 days were identified: (1) the gross motor domain assesses a child's ability to control their body, focusing on balance, movement, and coordination, with subdomains related to locomotion, balance, and manipulation of the body; (2) the fine motor domain evaluates children's use of their hands, fingers, and wrists to perform tasks with subdomains of drawing and writing, emphasizing eye-hand coordination; (3) the language domain includes both receptive and expressive communication skills, with a subdomain of children's ability to understand language, express themselves verbally, reason, name objects, understand prepositions, and solve analogies; (4) the cognitive domain tests abilities such as memory, problem-solving, and academic skills with subdomains of memory recall, spatial orientation, working with puzzles, arithmetic, shape, size, and color matching, as well as identifying letters and understanding time; and (5) the personal-social domain assesses children's ability to perform daily living skills, interact with others, and adjust to new situations, with subdomains of personal hygiene skills and social interactions.

An expert panel ($n = 5$) consisting a play therapist, speech language therapist, preschool educator, and pediatrics convened. This expert panel critically reviewed each developmental milestone to assess its cultural relevance and applicability. The panel evaluated whether the developmental milestones should be retained, modified, or removed from the item pool. A total of 141 milestones were selected from an initial pool of 188 based on expert ratings using a scale of 1 to 10 (with 1 being the least applicable and 10 being the most applicable). For inclusion, the items needed to receive 0.8 ratings from the expert panel. This process ensured that the final selection of milestones was relevant and applicable to the target population.

Next, the item development process was carefully structured to ensure that the items were suitable for the target age group and had appropriate reading levels for parents who had completed primary education in Malaysia. These items were written in Malay and English by the research team in collaboration with a linguistic expert in both English and Malay languages. This was done to ensure linguistic equivalence between the English and Malay versions of the TOY8 screening tool. Each version was then reviewed by the expert panel ($n = 5$) who confirmed that both versions maintained a similar level of complexity and provided an equal challenge for the children. This approach allowed us to avoid direct translation, which can introduce language inconsistencies or cultural bias [24]. We also hoped to minimize the risk of bias or discrepancies between the two versions, as each version was carefully crafted to suit the developmental understanding of the children within their respective linguistic environments.

Subsequently, we identified specific items that could be assessed through an AI application, optimizing the tool's usability by leveraging technology to enhance the screening process. Owing to technological limitations at the time of app development, the AI tool was unable to accurately detect large movements by the child (which are essential for assessing gross motor skills) or to recognize emotional expressions and social interactions. These developmental areas, which could not be measured by the AI application, were supported by a parent-proxy questionnaire, which also served as a form of check and balance for a comprehensive evaluation of the child's development. Of these, 192 were child-administered and 90 were parent-proxy items (full details can be found in Appendix A). The domains measured using the AI items and parent-proxy questionnaires are listed in Table 1.

A polytomous scoring approach was employed to capture and distinguish between children's mastery of certain milestones and their emerging skills. As the child develops, this method can avoid misdiagnoses, thereby minimizing unnecessary interventions. Screening at-risk children using this approach facilitate the implementation of a more effective and targeted support system.

There were demonstration, trial, and test items in the AI application: (1) demonstration items, represented by animated blue cats, provided the opportunity to show the child how to approach new or unfamiliar tasks before each item was presented. This provided guidance and clarity in

Table 1 Number of items for the TOY-EIGHT AI application and parent-proxy questionnaire

Domain	Number of items					
	3-years-old		4-years-old		5-years-old	
	AI application	Parent-proxy questionnaire	AI application	Parent-proxy questionnaire	AI application	Parent-proxy questionnaire
Gross motor		11		9		5
Fine motor	4	4	7	3	3	7
Language	24	9	17	8	19	6
Cognitive	10	3	23	7	8	3
Personal social		12		7		7

understanding the tasks at hand. (2) Trial items (less challenging than the test items) were required to familiarize the child with the test format and structure. These items were excluded from the final scoring system. (3) For the test items, all responses were recorded and scored from zero to three.

In the Parent-Proxy Questionnaire, all items began with the statement "Your child can…" with a four-point scale ("no," "sometimes can," "can do," and "never tried") which reflected the child's current developmental milestone. Parents were required to answer all questions in the parent-proxy questionnaire to provide a comprehensive picture of the child's development. The AI component was not used to dynamically skip questions based on responses to earlier items (e.g., skipping more advanced language questions if the parent answered "no" to simpler tasks). Instead, all questions were presented sequentially to ensure that all aspects of the child's abilities were assessed even if some responses suggested a developmental delay. This approach avoided the premature exclusion of areas of development that may still be relevant or achievable in different contexts. In addition, the AI app was based on predefined developmental milestones and did not alter or tailor questionnaire items based on parents' responses.

Once all items were finalized, they were embedded in an AI application. The application presented instructions, tasks, and interactive games to the child and recorded their responses in real time. To achieve a play-based approach, a fictional blue-colored cat character was created as an interactive agent in the TOY8 app. The cat was carefully crafted and driven by AI algorithms to provide instructions and demonstrate various tasks and activities in a dialog tailored to the understanding of children aged 3–5 years. This helped keep the child engaged throughout the session and ensured that the instructions were delivered appropriately based on the child's responses and progress. In terms of scoring and evaluation, AI tracked the accuracy and speed of a child's responses during the screening process. This automated processing provided objective and consistent results that were free from human bias.

Before the launch of the TOY8 app, both the Malay and English versions underwent feasibility testing via convenience sampling in a pilot study with 30 parents and their children. After obtaining written consent from the parents, the AI application was set up (Fig. 1) and the research assistants were thoroughly trained to ensure consistency in how they interacted with the children. Throughout the pilot study, detailed observations were made and any challenges or difficulties encountered while using the AI tool were carefully documented. Each parent was invited to complete the parent-proxy questionnaire and provide feedback on its feasibility, ease of use, and overall comprehensibility.

Fig. 1 TOY EIGHT AI Application Setup: A captivating experience where a child interacts with a smartphone screen guided by a fictional character. The set includes engaging materials such as stacking blocks, a drawing pen, and drawing sheets

Based on feedback and observations from 30 parents and their children, several areas for improvement were identified in both the screening tool and the parent-proxy questionnaire. Overall, most children were able to follow instructions and complete the tasks within 15–20 min. However, adjustments were made to the font size, color contrast, and voice clarity during AI administration to enhance usability based on the children's performance and reactions. For the parent-proxy questionnaire, parents reported that the items were easy to understand and relevant to their child's development. Nevertheless, further modifications were made to improve the clarity and refine the language based on parental feedback. These

adjustments enhanced the usability and effectiveness of both the screening tool and questionnaire, ensuring that they better addressed the needs of the target population and fulfilled their intended purpose.

The Malay version of the TOY8 underwent a vigorous validation process involving EFA, CFA, gross motor, fine motor, cognitive, language, and personal social subscale intercorrelations, and split-half reliability. The EFA results were consistent for all domains and subdomains that the construct intended to measure (three items were removed), which was reconfirmed by running the CFA. Table 1 shows the final items ($n = 138$) of the TOY8 developmental screening tool. The inter-correlations of the gross motor, fine motor, language, cognitive, and personal-social domains ($r = 0.225–0.577$, $p < 0.01$) showed evidence of convergent validity. Finally, the split-half reliability coefficients ranged from 0.600 to 0.804.

Participants

This study was conducted between 2021 and 2023. During this period, approximately 2,400 parents and their children were approached and recruited through convenience sampling using a dyadic approach. Recruitment took place at playschools, kindergartens, daycare centers, and shopping malls across Klang Valley. Recruitment posters were circulated on social media platforms. Consistent with a previous study (manuscript currently under review), the inclusion criteria were Malaysian parents and their children aged between 3 years 0 months and 5 years 11 months 30 days. TOY8 was designed to include all children within this age range, with no exclusions based on physical or developmental disabilities or chronic illnesses. The exclusion criteria included children who were unable to use the tool or whose parents failed to submit the questionnaire within two weeks of the child's assessment to ensure the integrity of the assessment. Parents reported their child's medical history, and 3.7–4.6% of the children included in the study had medical conditions or neurodevelopmental diagnoses but successfully completed the AI-based assessment. Demographic information is shown in Table 2.

A total of 2,178 parents consented to participate in the English version of the TOY8. However, some participants were excluded because of their children's inability to complete the assessment, as observed by the research assistants. Specifically, 34, 65, and 56 children from the three age groups were excluded for reasons such as being unwell or unable to follow instructions. Additionally, 11.75% of parents ($n = 256$) did not complete the parent-proxy questionnaire within the seven-day period, despite several reminders. Consequently, 1,767 participants were included in the final CFA.

During the same period, approximately 2,400 parents whose primary language of communication was Malay and their children were approached. Informed consent was obtained from 2,143 parents, and 1,724 dyads successfully completed the Malay version of the TOY8. This sample was independent of participants involved in a previous validation study using the Malay version. This new sample was specifically recruited to conduct a measurement invariance analysis across different demographic groups.

Procedure

Ethical approval was obtained before conducting the study (approval number: UM.TNC2/UMREC_1771), followed by permission letters from the Ministry of Education and the relevant school principals allowing the research to be conducted in the participating schools and kindergartens.

Parents were provided with detailed information about the study through a participant information sheet and informed consent was obtained before their children participated in the screening. Trained research assistants, under the supervision of a clinical psychologist, ensured that the screening instrument was administered systematically and consistently. The research assistants facilitated the session to ensure that the children followed the tasks correctly. They also closely monitored the child's behavior, including factors such as focus and mood (e.g., appearing distracted or irritable) to identify whether any underperformance was due to external factors. This allowed the team to interpret the results in context, ensuring that any challenges encountered during the session were not simply attributed to task performance, thereby enhancing the accuracy and reliability of the assessment outcomes. In addition to managing the tool, the research assistants were trained to build rapport with the children and identify those who did not meet the inclusion criteria. The TOY8 screening tool was designed to be user-friendly and accessible, enabling it to be administered in both schools and healthcare institutions by personnel with minimal training, rather than requiring professional developmental pediatricians or clinical psychologists.

In each session, children interacted with the TOY8 app along with physical materials (e.g., blocks and a stylus), completing tasks such as selecting the correct answer on the screen, stacking blocks, and drawing on sheets for approximately 15–20 min. The app tracked children's responses in real time, allowing for immediate feedback and analysis. All activities were guided by an animated character within the app that provided step-by-step instructions and demonstrations directly on the screen. While the child engaged with the app, the research assistant observed and documented the child's behavior, including temperament, learning methods, attention span, and pace, to supplement the AI-generated data with behavioral insights.

Table 2 Demographic information of the participating children and their parents

Characteristic	English (*n* = 1767) Age group		
	3–3.99 years old (*n* = 499)	4–4.99 years old (*n* = 648)	5–5.99 years old (*n* = 620)
Child's mean age (SD)	3.53 (0.29)	4.51 (0.29)	5.46 (0.28)
Child's gender			
Boy (%)	254 (50.8)	328 (50.6)	331 (53.4)
Girl (%)	245 (49.1)	320 (49.4)	289 (46.6)
Medical condition			
No (%)	479 (95.9)	618 (95.4)	697 (96.3)
Yes (%)	20 (4.1)	30 (4.6)	23 (3.7)
If yes, what is/are the condition(s)?			
Physical illness (eczema, asthma, etc.)	13	21	14
Speech delay	2	3	1
Autism Spectrum Disorder	4	3	6
ADHD	0	0	2
Developmental delay	1	3	0
Ethnicity			
Malay (%)	143 (29.7)	162 (25.0)	179 (28.9)
Chinese (%)	296 (59.3)	385 (59.4)	327 (52.7)
Indian (%)	43 (8.6)	81 (12.5)	87 (14.0)
Others (%)	17 (3.4)	20 (3.1)	27 (4.4)
Language used most often to communicate with child at home			
Malay (%)	11 (2.2)	25 (3.8)	14 (2.3)
English (%)	452 (90.5)	591 (91.2)	559 (90.2)
Chinese (%)	31 (6.2)	25 (3.9)	38 (6.1)
Tamil (%)	5 (1.0)	7 (1.1)	9 (1.4)
Religion			
Muslim (%)	146 (29.3)	167 (25.8)	189 (30.5)
Buddhist (%)	244 (8.9)	302 (46.6)	252 (40.6)
Christian (%)	49 (9.8)	79 (12.2)	67 (10.8)
Hindu	35 (7.0)	70 (10.8)	70 (11.3)
Others (%)	25 (5.0)	30 (4.6)	42 (6.9)
Parent's highest educational level			
Secondary education (%)	37 (7.5)	62 (9.6)	68 (11.0)
Certificate/Diploma (%)	111 (22.2)	111 (17.1)	117 (18.9)
Bachelor's Degree (%)	280 (56.1)	366 (56.5)	320 (51.6)
Postgraduate Degree	110 (14.2)	109 (16.8)	115 (18.5)
	Measurement invariance across groups (*n* = 1724)		
	3–3.99 years old (*n* = 958)	4–4.99 years old (*n* = 1260)	5–5.99 years old (*n* = 1273)
Language versions			
Malay (*n* = 1724)	459 (47.9)	612 (48.6)	653 (51.3)
English (*n* = 1767)	499 (52.1)	648 (51.4)	620 (48.7)
Gender			
Boy	484 (50.5)	623 (49.4)	632 (49.6)
Girl	474 (49.5)	637 (50.6)	641 (50.4)
Family monthly household income			
RM4000 and below ($1319.56 USD)	239 (24.0)	405 (32.1)	555 (43.3)
RM6000–RM9999 ($1759.34 USD)	434 (45.3)	481 (38.2)	383 (30.1)
RM10000 ($2200 USD) and above	241 (25.2)	309 (24.5)	282 (22.0)
Not reported	44 (4.6)	65 (0.2)	62 (4.7)

After the screening session, the parents received a parent report/proxy questionnaire via email or WhatsApp. They were asked to complete the questionnaire within seven days, with a reminder sent to those who did not respond within the given timeframe.

Data from both AI-based screening and parent-proxy questionnaires were integrated to provide a holistic view of the child's developmental progress. A simple developmental report was generated and shared with the parents (Appendix B), including recommended activities tailored to support the child's growth. Parents were informed that the report was not a diagnostic tool and that any concerns raised should be followed up with professional evaluation, if needed.

Data analysis

The validation process for the English version of the TOY8 involved CFA testing and split-half reliability. Measurement invariance was then analyzed using both the English and Malay versions of the TOY8.

First, CFA using maximum likelihood estimation was conducted to verify whether the data from the English version of the TOY8 supported the factor structure across all domains of the Malay version of the TOY8. CFA is a critical step in validating the proposed model by testing whether the data fit the hypothesized structure. This analysis allowed us to assess the validity of the factor structure, ensuring that the items loaded appropriately onto their respective constructs and met the criteria for good model fit.

A model is considered to fit the data when the following values are obtained: chi-square/degrees of freedom (χ^2/df) = < 3.0, root mean square error of approximation (RMSEA) = < 0.08, and standardized root means square residual (SRMR) = < 0.06 [18–20]. The goodness of fit statistics exhibited a preference for sample bias: goodness-of-fit index (GFI) = > 0.90, adjusted (AGFI) > 0.80, Tucker−Lewis index (TLI) and comparative fit index (CFI) ≥ 0.90, and ≥ 0.95 considered a more ideal fit [25–27]. To identify the best-fitting model, we examined modification indices to identify the covariance to be drawn where the model could improve its fit. The final modified model demonstrated an improved fit as reflected in the key fit indices (CFI, RMSEA, and SRMR), indicating that it more accurately captured the underlying structure of the data. Although empirical statistics are significant when modifying a model, the contents of developmental milestones are of equal importance when making decisions to retain or remove an item [28].

Split-half reliability was used to assess the internal consistency and reliability of the tool. During this process, responses to the screening tool were randomly divided into two halves. Each half was treated as a separate set of items and their scores were compared. If the tool is internally consistent, then the scores of both halves should be highly correlated. A high correlation between the two halves indicates that the tool consistently measures the same underlying construct, demonstrating its reliability.

Measurement invariance analyses were conducted to investigate whether the Malay versus English version, children's gender and children from different income groups ascribed a different meaning to the same set of items in TOY8. This step was crucial to ensure that the tool was culturally appropriate and measured developmental milestones in a comparable manner across these subgroups. Establishing measurement invariance ensured that differences in scores reflected true differences in child development rather than the bias introduced by language, gender, or socioeconomic factors. For instance, if there were deviations between two languages, invariance analyses could pinpoint the differences.

First, configural invariance tests (equal-factor patterns) were conducted. Subsequently, metric, scalar, and residual invariance were tested by sequentially constraining the factor loadings, intercepts, and residual variances. These tests were conducted incrementally with key model fit indices (such as CFI, TLI, RMSEA, and SRMR—carefully monitored at each step to assess the impact of the constraints. Constraints were deemed acceptable if the model fit did not deteriorate significantly, ensuring that the model maintained an adequate fit across different groups. This approach ensured that the tool functions equivalently and fairly across various populations, minimizes bias, and supports reliable cross-group comparisons. As individual differences in the latent construct are often of interest, metric invariance (comparable factor loadings) is often a sufficient assumption [28]. In other words, when metric invariance is supported, it indicates that when there is an equal increase in raw scores, there is an equal increase in latent traits. Therefore, children from both groups interpreted the item in the same manner.

The criteria to support the assumption of measure invariance included a difference in the CFI value of ≤ 0.01 and an RMSEA value not greater than 0.015 [29]. Some studies used statistically insignificant models to support this assumption, whereas in the present study, chi-square tests were not used to test for differences in fit between models because chi-square tests can be significantly affected by the size of the sample. When the sample size is large, chi-square test can be overly sensitive to minor discrepancies between the observed data and the model, potentially leading to the rejection of models that fit reasonably well [30]. All analyses were performed using the IBM SPSS Statistics (SPSS) v.27 and AMOS version 27.

Results

CFA

The CFA results indicated that the model structure of the English version of the TOY8 matched that of the Malay version (Table 3).

Table 3 Confirmatory factor analysis of the English version of the TOY8 developmental screening tool

Domain	χ^2	df	χ^2/ df	GFA	AGFA	TLI	CFI	RMSEA	SRMR	*p* value
3-years-old* (*n* = *499)										
Gross motor (four-factor model: manipulation, stationary skills- balancing, locomotion- staircase, and locomotion)	64.291	38	1.692	0.978	0.961	0.958	0.957	0.037	0.369	< 0.001
Fine motor (one-factor model)	34.025	17	2.001	0.983	0.964	0.936	0.961	0.045	0.0565	< 0.001
Language (four- factor model: receptive, expressive, preposition, and color naming)	712.466	441	1.616	0.919	0.903	0.935	0.942	0.035	0.0470	< 0.001
Cognitive (three- factor model: sequence; matching size and color; memory)	117.532	83	1.416	0.969	0.956	0.953	0.963	0.029	0.0413	< 0.001
Personal social (two-factor model: personal hygiene and social skills)	78.617	48	1.638	0.971	0.953	0.932	0.951	0.039	0.0399	< 0.001
4-years-old (*n* = *648)*										
Gross motor (one-factor model: locomotion and balancing)	42.074	25	1.683	0.985	0.974	0.939	0.958	0.032	0.0336	< 0.001
Fine motor (one-factor model)	52.762	25	2.110	0.982	0.967	0.954	0.968	0.041	0.0353	< 0.001
Language: (five-factor model: expressive, prepositive, object function, color naming, and practical reasoning)	802.143	451	1.779	0.928	0.916	0.947	0.952	0.035	0.0432	< 0.001
Cognitive (six-factor mode: lego, arithmetic, identify alphabet, matching size, color and shape, spatial orientation and discrimination, and memory)	634.120	432	1.468	0.942	0.929	0.944	0.951	0.027	0.0397	< 0.001
Personal social (two-factor model: personal hygiene and social skill)	17.941	10	1.794	0.992	0.979	0.972	0.987	0.035	0.0215	< 0.001
5-years-old (*nn* = 620)										
Gross motor (one-factor model)	12.071	4	3.018	0.992	0.972	0.939	0.976	0.057	0.0289	< 0.001
Fine motor (one-factor model)	5.867	34	1.555	0.983	0.973	0.944	0.958	0.030	0.0336	< 0.001
Language (three-factor model: expressive, receptive, and analogy)	3780.647	219	1.692	0.950	0.937	0.947	0.954	0.033	0.0370	< 0.001
Cognitive (three-factor model: arithmetic, matching block design, and memory)	161.245	124	1.300	0.972	0.962	0.934	0.947	0.022	0.0354	< 0.001
Personal social (two-factor model: personal hygiene and social skills)	25.706	10	2.571	0.989	0.9668	0.938	0.971	0.050	0.0310	< 0.001

Table 4 Spilt-half reliability coefficient of the English version of the Toy8 tool for each domain across age groups

Domain	Split-Half Reliability Coefficient		
	3-year-old	4-year-old	5-year-old
Gross motor	0.701	0.712	0.620
Fine motor	0.704	0.700	0.702
Language	0.759	0.828	0.842
Cognitive	0.716	0.736	0.701
Personal social	0.725	0.709	0.734

Reliability

The split-half reliability of the English version of the TOY8 was assessed using a random split of responses from parents and children in each age group ($n = 499$–648). All responses were randomized and split into first (Set A) and second (Set B) halves. The correlation coefficient was calculated using Pearson's r, and the total scores between sets A and B ranged from 0.419 to 0.707. This indicated a strong positive correlation between the two sets.

Subsequently, the Spearman-Brown prophecy formula was used to calculate the split-half reliability coefficient to estimate the reliability of the English version of the TOY8. The split-half reliability coefficient ranged from 0.620 to 0.828 (Table 4), indicating adequate-to-high reliability. These results are consistent with those of a previous Malay version of the TOY8.

Measurement invariance

Measurement invariance across language versions (Malay vs. English), gender (male vs. female), and income groups (B40, M40, and T20) was tested. Our results showed that all configural invariance models had a good fit to the data, demonstrating that multiple-group CFA was appropriate. The factors in the English version of the TOY8 could be measured with the same factor pattern as the Malay version for children aged 3–5 years. Further, equivalence analyses could be conducted. Furthermore, the difference in the CFI across all models with restriction of factor loading was not significant (ΔCFI = -0.004 to – 0.010), suggesting that the increase in the model was not substantial with the imposition of equality constraints, thus suggesting that all domains could be measured the same way across language versions, ages, and income groups. These results support metric invariance. Finally, several models supported scale invariance: three-year-old gross motor domain (gender) and language domain (language version) [see Table 5; four-year-old fine motor (gender), language domain (gender), and cognitive domain (gender) [see Table 6]; and five-year-old language domain (gender) [see Table 7].

Discussions and conclusions

Two samples of children aged 3–5 years were recruited to examine (1) the construct validity and reliability of the English version of the TOY8 based on its Malay version (a paper publication currently under review), and (2) the testing of measurement invariance across language versions, gender, and income groups to determine the cross-cultural applicability and validity of the TOY8 developmental screening tool. We also sought to ensure that TOY8 could accurately measure developmental milestones among children aged 3–5 years across diverse linguistic and demographic backgrounds in Malaysia. This tool is currently in the initial testing phase and is systematically evaluated as part of the early stage development process. Although this study provided valuable initial insights into its design and application, further data collection in diverse real-world settings should be conducted to establish its broader applicability.

The English version of the TOY8 developmental screening tool demonstrated construct validity and reliability as a screening tool in identifying developmental milestones in children. In addition, configural and metric invariances across language versions, gender and income groups in all three- to five-year-old were established; scalar invariance was established across gender in most three- to five-year-old. These findings provide preliminary evidence implies that our study aligns with the previous literature on child development, which indicates that there is a general similarity in gender and cross-cultural development domains in the first year of life [9, 31]. However, language, speech, and socioemotional skills are largely affected by a child's level of exposure and learning environment as they age [10, 16].

Another significant finding of this study is that establishing a stronger level of measurement invariance across all domains is challenging for children from different income groups. In this regard, only full metric measurement invariance can be achieved. This finding is consistent with the previous literature [32].

One possible reason for this is that children from different income groups often experience varying levels of exposure to resources [33]. For example, access to educational opportunities, material resources, parental involvement, healthcare, and community resources may vary significantly among income groups. High-income families typically have greater access to ECE programs. In contrast, lower-income families may face financial constraints that limit access to these resources, potentially affecting their developmental trajectories. Additionally, socioeconomic status (SES) can affect children's health (e.g., stunting issues) and access to better healthcare services. These factors further contribute to differences in language development, which are closely correlated with

Table 5 Measurement invariance across language versions, gender, and income groups of the English version of the TOY8 for the three-year-old subscale

	Model fit information					
Domain/Model	CFI	ΔCFI	TLI	RMSEA (90%CI)	ΔRMSEA	SRMR
Gross motor						
Language version						
1a. Configural	0.962		0.963	0.026 (0.019; 0.033)		0.0333
2a. Metric	0.958	-0.004	0.942	0.027 (0.020; 0.034)	-0.001	0.0314
3a. Scalar	0.938	-0.020	0.932	0.032 (0.026; 0.038)	-0.005	0.0309
Gender						
1b. Configural	0.952		0.930	0.030 (0.023; 0.036)		0.0438
2b. Metric	0.945	-0.007	0.927	0.030 (0.240; 0.305)	0.00	0.0474
3b. Scalar	0.938	-0.007	0.928	0.030 (0.024; 0.036)	0.00	0.0474
4b. Residual	0.916	-0.022	0.904	0.035 (0.043; 0.041)	-0.005	0.0709
Income						
1c. Configural	0.951		0.926	0.027 (0.022; 0.033)		0.0392
2c. Metric	0.941	-0.010	0.918	0.029 (0.023; 0.034)	-0.002	0.0442
3c. Scalar	0.912	-0.029	0.900	0.032 (0.021; 0.035)	-0.003	0.0489
Fine motor						
Language version						
1d. Configural	0.975		0.959	0.025 (0.014; 0.036)		0.0330
2d. Metric	0.972	-0.003	0.960	0.025 (0.015; 0.035)	0.00	0.0301
3d. Scalar	0.931	-0.041	0.914	0.037 (0.029; 0.045)	-0.012	0.0259
Gender						
1e. Configural	0.983		0.973	0.020 (0.006; 0.032)		0.0363
2e. Metric	0.983	0	0.977	0.019 (0.003; 0.029)	0.001	0.0404
3e. Scalar	0.911	-0.072	0.885	0.042 (0.035; 0.050)	-0.023	0.0418
Income						
1 f. Configural	0.989		0.982	0.014 (0.01; 0.025)		0.0532
2 f. Metric	0.989	0	0.986	0.012 (0.01; 0.022)	0.002	0.0531
3 f. Scalar	0.975	-0.014	0.974	0.017 (0.05; 0.025)	-0.005	0.0524
Language						
Language version						
1 g. Configural	0.952		0.946	0.021 (0.019; 0.023)		0.0433
2 g. Metric	0.955	0.003	0.950	0.020 (0.018; 0.022)	0.001	0.0433
3 g. Scalar	0.949	-0.006	0.946	0.021 (0.019; 0.023)	-0.001	0.0475
Gender						
1 h. Configural	0.951		0.946	0.021 (0.019; 0.023)		0.0465
2 h. Metric	0.950	-0.001	0.944	0.022 (0.020; 0.024)	0.001	0.0477
3 h. Scalar	0.917	-0.033	0.911	0.028 (0.026; 0.030)	-0.006	0.0538
Income						
1i. Configural	0.950		0.941	0.018 (0.017; 0.021)		0.0505
2i. Metric	0.947	-0.003	0.936	0.019 (0.017; 0.021)	-0.001	0.0611
3i. Scalar	0.909	-0.038	0.905	0.024 (0.022; 0.025)	-0.005	0.0657
Cognitive						
Language version						
1j. Configural	0.962		0.939	0.022 (0.019; 0.023)		0.0337
2j. Metric	0.956	-0.006	0.942	0.025 (0.018; 0.026)	-0.003	0.0403
3j. Scalar	0.932	-0.024	0.912	0.027 (0.019; 0.028)	-0.002	0.0467
Gender						
1k. Configural	0.959		0.950	0.021 (0.018; 0.022)		0.0465
2k. Metric	0.950	-0.009	0.942	0.023 (0.019; 0.025)	-0.002	0.0497
3k. Scalar	0.919	-0.031	0.909	0.028 (0.025; 0.031)	-0.005	0.0578
Income						
1 l. Configural	0.950		0.920	0.020 (0.019; 0.021)		0.0505

Table 5 (continued)

	Model fit information					
2 l. Metric	0.941	-0.009	0.907	0.024 (0.018; 0.025)	-0.004	0.0611
3 l. Scalar	0.907	-0.034	0.905	0.029 (0.021; 0.031)	-0.005	0.0657
Personal Social						
Language version						
1 m. Configural	0.953		0.931	0.027 (0.021; 0.034)		0.0277
2 m. Metric	0.943	-0.010	0.921	0.029 (0.023; 0.035)	-0.002	0.0327
3 m. Scalar	0.923	-0.020	0.912	0.031 (0.024; 0.037)	-0.002	0.0423
Gender						
1n. Configural	0.950		0.915	0.031 (0.025; 0.038)		0.0289
2n. Metric	0.943	-0.007	0.914	0.032 (0.025; 0.038)	-0.001	0.0308
3n. Scalar	0.917	-0.025	0.891	0.035 (0.030; 0.041)	-0.003	0.0456
Income						
1o. Configural	0.937		0.908	0.027 (0.021; 0.032)		0.0371
2o. Metric	0.927	-0.010	0.900	0.029 (0.024; 0.034)	-0.002	0.0542
3o. Scalar	0.910	-0.027	0.878	0.032 (0.022; 0.035)	-0.003	0.0594

Note: Model 1 = configural invariance (no constraint on all parameters); Model 2 = metric invariance (equally constrained for all factor loadings); Model 3 = scalar invariance (equally constrained factor loadings and intercepts); Model 4 = residual invariance (the sum of specific variance and error variance is similar). CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual

cognitive development and later academic achievement [34].

We acknowledge that factors such as access to educational resources, learning environments, and parental involvement can affect developmental outcomes across SES backgrounds. Therefore, we recommend that all stakeholders—parents, teachers, and healthcare professionals— consider these contextual factors when interpreting screening results. For instance, when children from lower SES backgrounds show delays in certain areas, it is important to explore whether these delays can be attributed to environmental factors rather than intrinsic developmental issues.

Because the app has been validated and shown to be reliable, it generates a comprehensive report of a child's outcomes, including recommendations for activities that support development in areas where improvement is required. These recommendations are tailored to leverage widely available resources. Additionally, the app provides referrals and additional support for families from lower SES groups, ensuring that they are connected with appropriate resources and interventions when necessary.

Finally, the screening tool can be used to increase parental awareness, highlighting potential developmental red flags. By focusing on areas of concern, the tool empowers parents to take proactive steps to support their child's developmental progress. If necessary, they are encouraged to seek interventions from qualified developmental providers or licensed psychologists. By increasing awareness of a child's developmental stage, this tool helps mitigate disparities in developmental opportunities across income groups.

Limitations and future research

One limitation of this study was that the tool was not designed to comprehensively screen children with moderate-to-severe disabilities. Currently, this tool is intended as a developmental screening tool to identify children at risk of delays or in need of further evaluation and support. This is not meant to replace formal assessments of children with moderate-to-severe disabilities. For children in this category, developmental concerns are often identified earlier and addressed through specialized assessments rather than general screening tools, particularly for children aged 3–5 years. For example, tools such as the Modified Checklist for Autism in Toddlers are used as early as 18 months of age to screen for autism spectrum disorder and related conditions. Therefore, our tool is specifically aimed at detecting children who may otherwise go undetected without a systematic screening process. Importantly, children who were unable to complete the screening tasks were not excluded. Instead, they were referred for further assessment using teacher or research assistant reports and observational data. This approach ensures that children requiring additional support are not overlooked and receive appropriate follow-ups. Future iterations of the tool may explore ways to accommodate children with moderate-to-severe disabilities better, enhance inclusivity, and broaden the scope of its application.

Additionally, although it is important to note that AI assessment has promising opportunities for evaluating child development, it should not be viewed as a substitute for in-person assessment by trained professionals. This limitation arises from the inability of AI systems to capture subtle distinctions in behavioral cues or

Table 6 Measurement invariance across language versions, gender, and income groups of the English version of the TOY8 for the four-year-old subscale

	Model fit information					
Model	CFI	ΔCFI	TLI	RMSEA (90%CI)	Δ RMSEA	SRMR
Gross motor						
Language version						
1a. Configural	0.958		0.928	0.029 (0.023; 0.035)		0.0269
2a. Metric	0.954	-0.004	0.920	0.031 (0.025; 0.037)	-0.002	0.0222
3a. Scalar	0.785	-1.690	0.697	0.060 (0.55; 0.065)	-0.029	0.0308
Gender						
1b. Configural	0.957		0.926	0.029 (0.023; 0.035)		0.0235
2b. Metric	0.952	-0.005	0.931	0.028 (0.022; 0.033)	0.001	0.0286
3b. Scalar	0.941	-0.011	0.926	0.029 (0.024; 0.034)	-0.001	0.0271
Income						
1c. Configural	0.952		0.925	0.023 (0.018; 0.028)		0.0269
2c. Metric	0.947	-0.005	0.930	0.022 (0.018; 0.027)	0.001	0.0285
3c. Scalar	0.863	-0.084	0.852	0.033 (0.029; 0.036)	-0.011	0.0315
Fine motor						
Language version						
1d. Configural	0.970		0.957	0.028 (0.022; 0.033)		0.0293
2d. Metric	0.966	-0.004	0.958	0.027 (0.022; 0.033)	0.001	0.0305
3d. Scalar	0.901	-0.065	0.894	0.043 (0.039; 0.048)	-0.016	0.0314
Gender						
1e. Configural	0.977		0.965	0.025 (0.021; 0.029)		
2e. Metric	0.978	0.001	0.972	0.023 (0.019; 0.026)	0.002	0.0259
3e. Scalar	0.980	0.002	0.978	0.020 (0.017; 0.024)	0.003	0.0288
4e. Residual	0.945	-0.035	0.939	0.033 (0.029; 0.038)	-0.013	0.0374
Income						
1 f. Configural	0.981		0.971	0.019 (0.015; 0.022)		0.0247
2 f. Metric	0.979	-0.002	0.975	0.017 (0.014; 0.021)	0.002	0.0268
3 f. Scalar	0.965	-0.014	0.964	0.021 (0.018; 0.024)	-0.004	0.0277
Language						
Language version						
1 g. Configural	0.952		0.944	0.023 (0.021; 0.024)		0.0399
2 g. Metric	0.943	-0.009	0.935	0.024 (0.023; 0.026)	-0.001	0.0420
3 g. Scalar	0.912	-0.031	0.921	0.028 (0.024; 0.031)	-0.004	0.0591
Gender						
1 h. Configural	0.950		0.942	0.023 (0.021; 0.024)		0.0522
2 h. Metric	0.950	0.00	0.944	0.022 (0.021; 0.024)	0.001	0.0528
3 h. Scalar	0.946	-0.004	0.941	0.023 (0.022; 0.024)	0.00	0.0529
4 h. Residual	0.912	-0.034	0.923	0.029 (0.023; 0.031)	-0.006	0.0601
Income						
1i. Configural	0.950		0.942	0.019 (0.017; 0.020)		0.0433
2i. Metric	0.944	-0.006	0.937	0.019 (0.018; 0.021)	0.00	0.0445
3i. Scalar	0.918	-0.026	0.913	0.023 (0.022; 0.024)	-0.004	0.0440
Cognitive						
Language version						
1j. Configural	0.953		0.947	0.020 (0.018; 0.021)		0.0410
2j. Metric	0.949	-0.004	0.944	0.020 (0.019; 0.022)	0.00	0.0463
3j. Scalar	0.899		0.893	0.028 (0.027; 0.029)	-0.008	0.0672
Gender						
1k. Configural	0.953		0.947	0.021 (0.020; 0.022)		0.0359
2k. Metric	0.952	-0.001	0.948	0.021 (0.019; 0.022)	0.00	0.0360
3k. Scalar	0.952	0.00	0.949	0.021 (0.019; 0.022)	0.00	0.0360
4k. Residual	0.951	-0.001	0.949	0.021 (0.019; 0.022)	0.00	0.0379

Table 6 (continued)

	Model fit information					
Income						
1 l. Configural	0.951		0.944	0.017 (0.016; 0.019)		0.0310
2 l. Metric	0.943	-0.008	0.938	0.018 (0.017; 0.019)	-0.001	0.0307
3 l. Scalar	0.924	-0.019	0.921	0.020 (0.019; 0.022)	-0.002	0.0316
Personal Social						
Language version						
1 m. Configural	0.983		0.965	0.027 (0.019; 0.036)		0.0212
2 m. Metric	0.982	-0.001	0.968	0.026 (0.018; 0.034)	0.001	0.0293
3 m. Scalar	0.890	-0.092	0.850	0.056 (0.050; 0.063)	-0.030	0.0297
Gender						
1n. Configural	0.981		0.960	0.029 (0.021; 0.038)		0.0229
2n. Metric	0.979	-0.002	0.964	0.028 (0.020; 0.036)	0.001	0.0269
3n. Scalar	0.945	-0.034	0.927	0.040 (0.033; 0.046)	-0.012	0.0263
Income						
1o. Configural	0.981		0.960	0.025 (0.017; 0.032)		0.0214
2o. Metric	0.972	-0.009	0.956	0.026 (0.020, 0.032)	-0.001	0.0286
3o. Scalar	0.925	-0.047	0.913	0.036 (0.031; 0.042)	-0.010	0.0293

Note: Model 1 = configural invariance (no constraint on all parameters); Model 2 = metric invariance (equally constrained for all factor loadings); Model 3 = scalar invariance (equally constrained factor loadings and intercepts); Model 4 = residual invariance (the sum of specific variance and error variance is similar). CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual

interpersonal interactions, which may be crucial for a comprehensive assessment.

Furthermore, there are concerns regarding data privacy and the potential for bias in AI assessments, which should be carefully considered and addressed in any implementation, particularly regarding the confidentiality and security of sensitive personal information. Although the TOY8 screening tool is beneficial for child development, it is important to weigh the potential benefits against potential risks and challenges and to ensure that it is used in an ethical and responsible manner. There may be a risk of bias in AI algorithms, which may inadvertently perpetuate inequalities or misrepresent certain groups if not properly mitigated.

This study focused solely on the Klang Valley region, a metropolitan region in Malaysia that includes Kuala Lumpur, the national capital, and several surrounding areas in the state of Selangor, particularly during and after the pandemic; these areas may not adequately represent the diverse demographics and experiences of children across all states in Malaysia, although those living in the Klang Valley region are a diverse group. Furthermore, statistical analyses such as EFA, CFA, and measurement invariance are essential for understanding the psychometric properties of the assessment and identifying the underlying constructs based on established developmental domains and milestones. These statistical analyses represent the first step in establishing construct validity and provide a foundational framework for evaluating how well the tool measures the intended constructs. We acknowledge that the core of a comprehensive developmental assessment lies in its real-world application to children and families, gathering meaningful feedback from parents, and evaluating its practical utility through methods such as inter-rater and test-retest reliability. These processes will be the next steps in further examining the reliability of the tool.

Finally, to further enhance the convergent validity of the TOY8, another study is underway to establish the clinical confirmations of TOY8 based on professional assessments using the Griffiths Scales of Child Development version III assessment tool (gold standard) have been conducted in all states in Malaysia. These efforts ensure that the tool reliably and validly measures the developmental progress of children aged 3–6 years in Malaysia and will contribute to establishing standardized norms specific to the Malaysian context.

Practical implications

The use of the TOY8 developmental screening tool to measure the developmental progress of children aged 3–5 years in Malaysia is potentially valuable to both the government and ECE educators, as it could help identify areas where children may need additional support or intervention in a more efficient and objective way.

The TOY8 could provide data-driven insights into the educational needs and progress of children across the country for Malaysian government. This could help policymakers make informed decisions regarding resource allocation, curriculum development, intervention programs, and other areas related to education.

For ECE educators, the TOY8 can be used to help formulate teaching plans and identify areas in which individual children may need additional support. For

Table 7 Measurement invariance across language version, gender and income groups of the English version of the TOY8 for the five-year-old subscale

	Model fit information					
Model	CFI	ΔCFI	TLI	RMSEA (90%CI)	ΔRMSEA	SRMR
Gross motor						
Language version						
1a. Configural	0.977		0.957	0.026 (0.019; 0.034)		0.0183
2a. Metric	0.974	-0.003	0.959	0.025 (0.018; 0.032)	0.001	0.0154
3a. Scalar	0.890	-0.084	0.867	0.046 (0.040; 0.051)	-0.021	0.0189
Gender						
1b. Configural	0.969		0.942	0.031 (0.024; 0.038)		0.0209
2b. Metric	0.968	-0.001	0.952	0.027 (0.021; 0.034)	0.004	0.0228
3b. Scalar	0.940	-0.028	0.933	0.033 (0.028; 0.039)	-0.006	0.0229
Income						
1c. Configural	0.969		0.949	0.024 (0.018; 0.030)		0.0252
2c. Metric	0.966	-0.003	0.955	0.022 (0.017; 0.028)	-0.002	0.0224
3c. Scalar	0.849	-0.117	0.855	0.040 (0.036; 0.044)	-0.018	0.0239
Fine motor						
Language version						
1d. Configural	0.952		0.930	0.027 (0.023; 0.031)		
2d. Metric	0.946	-0.006	0.931	0.027 (0.023; 0.031)	0.00	0.0220
3d. Scalar	0.853	-0.093	0.837	0.042 (0.039; 0.045)	-0.015	0.0294
Gender						
1e. Configural	0.950		0.928	0.028 (0.024; 0.032)		0.0279
2e. Metric	0.949	-0.001	0.936	0.026 (0.022; 0.030)	0.002	0.0291
3e. Scalar	0.879	-0.070	0.866	0.038 (0.035; 0.041)	0.012	0.0300
Income						
1 f. Configural	0.943		0.917	0.024 (0.021; 0.028)		0.0278
2 f. Metric	0.940	-0.004	0.927	0.023 (0.020; 0.026)	0.001	0.0288
3 f. Scalar	0.871	-0.069	0.868	0.031 (0.028; 0.034)	-0.008	0.0284
Language						
Language version						
1 g. Configural	0.953		0.947	0.020 (0.018; 0.022)		0.0257
2 g. Metric	0.945	-0.008	0.939	0.021 (0.020; 0.023)	0.001	0.0268
3 g. Scalar	0.835	-0.110	0.827	0.036 (0.035; 0.037)	-0.015	0.0288
Gender						
1 h. Configural	0.950		0.943	0.021 (0.019; 0.022)		0.0325
2 h. Metric	0.949	-0.001	0.942	0.022 (0.020; 0.023)	-0.001	0.0326
3 h. Scalar	0.946	-0.003	0.944	0.021 (0.019; 0.022)	0.001	0.0330
Income						
1i. Configural	0.953		0.946	0.016 (0.015; 0.018)		0.0259
2i. Metric	0.945	-0.008	0.940	0.017 (0.016; 0.018)	-0.001	0.0265
3i. Scalar	0.874	-0.071	0.871	0.025 (0.024; 0.026)	-0.008	0.0265
Cognitive						
Language version						
1j. Configural	0.982		0.972	0.015 (0.019; 0.020)		0.0212
2j. Metric	0.979	-0.003	0.972	0.015 (0.010; 0.020)	0.00	0.0254
3j. Scalar	0.876	-0.103	0.858	0.034 (0.030; 0.038)	-0.019	0.0210
Gender						
1k. Configural	0.977		0.965	0.017 (0.012; 0.022)		0.0245
2k. Metric	0.981	0.004	0.975	0.014 (0.009; 0.019)	0.003	0.0217
3k. Scalar	0.974	-0.007	0.9698	0.016 (0.011; 0.021)	-0.002	0.0231
Income						
1 l. Configural	0.979		0.969	0.013 (0.008; 0.018)		0.0216
2 l. Metric	0.981	0.003	0.0975	0.012 (0.007; 0.016)	0.001	0.0222

Table 7 (continued)

	Model fit information					
3 l. Scalar	0.877	-0.104	0.870	0.026 (0.023; 0.030)	-0.014	0.0225
Personal social						
Language version						
1 m. Configural	0.972		0.951	0.032 (0.026; 0.038)		0.0203
2 m. Metric	0.967	-0.005	0.949	0.033 (0.029; 0.039)	-0.001	0.0171
3 m. Scalar	0.902	-0.065	0.879	0.051 (0.046; 0.056)	-0.018	0.0183
Gender						
1n. Configural	0.960		0.930	0.039 (0.033; 0.045)		0.0303
2n. Metric	0.960	0.000	0.942	0.035 (0.030; 0.040)	0.004	0.0310
3n. Scalar	0.945	-0.015	0.936	0.037 (0.032; 0.042)	-0.002	0.0311
Income						
1o. Configural	0.956		0.923	0.034 (0.029; 0.039)		0.0261
2o. Metric	0.946	-0.010	0.926	0.034 (0.029; 0.038)	0.00	0.0266
3o. Scalar	0.925	-0.021	0.907	0.032 (0.027; 0.0038)	-0.002	0.0302

Note: Model 1 = configural invariance (no constraint on all parameters); Model 2 = metric invariance (equally constrained for all factor loadings); Model 3 = scalar invariance (equally constrained factor loadings and intercepts); Model 4 = residual invariance (the sum of specific variance and error variance is similar). CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual

example, AI assessments can be used to identify children struggling with certain concepts or skills, or those who may benefit from more challenging work.

In addition, AI-based screening tools can help streamline the assessment process and reduce the amount of time teachers must spend on administrative tasks. By improving the accuracy and consistency of assessments, AI-based tools can help eliminate errors that may occur when recording data manually. In addition, digital tools can provide objective data that are less susceptible to personal biases or subjectivity.

AI-based screening tools also allow educators to easily track children's progress over time, which can help identify areas where additional support or intervention may be needed. This could also provide a more comprehensive view of children's development, which could inform curriculum planning and individualized instruction.

Overall, AI-based screening tools could be a valuable addition to the educational landscape in Malaysia. However, their use should be carefully considered based on evidence of their effectiveness, with appropriate safeguards in place to protect the privacy and well-being of children. Educators should receive appropriate training and support to interpret and use the results of the TOY8 effectively.

Abbreviations

AI	Artificial intelligence
CFA	Confirmatory factor analysis
CFI	Comparative fit index
ECE	Early childhood education
EFA	Exploratory factor analysis
RMSEA	Root mean square error of approximation
SES	Socioeconomic status
SRMR	Standardized root mean square residual
TLI	Tucker-Lewis Index
TOY8	Toy Eight Developmental Screening Tool

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40359-025-02489-3.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

The authors wish to express their appreciation to Toybox Creations and Technology Sdn Bhd for their contribution, assistance and bearing the cost of the data collection in developing the Toy8 developmental screening tool.

Author contributions

SWW (first author): Involved in data analysis and interpretation of data, manuscript writing, and approved the submitted version PA (corresponding author): Involved in developing the questionnaire, research design, and manuscript writing, and approved the submitted version ANY: Involved in developing the questionnaire, manuscript writing, and approved the submitted version PJW: Involved in developing AI items, and approved the submitted version.

Funding

The research did not receive any specific funding.

Data availability

The data that support the findings of this study are available from Toybox Creations and Technology Sdn Bhd but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Toybox Creations and Technology Sdn Bhd. All materials are copyrighted and therefore will not be available.

Declarations

Consent for publication

Not applicable as no personal data or identifying images were included in the current study.

Human ethics and consent to participate

The researchers-maintained data privacy and security while adhering to the applicable ethical guidelines (approval no. UM). TNC2/UMREC_1771 was approved by the University of Malaya Research Ethics Committee. Informed consent was obtained from the parents or guardians and did not include children. Additionally, permission to recruit participants from preschools

and kindergartens was obtained from the Ministry of Education and school principals.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Psychology, School of Medical and Life Sciences, Sunway University, No. 5 Jalan Universiti, Bandar Sunway, Selangor Darul Ehsan, Petaling Jaya 47500, Malaysia
[2]Department of Management and Marketing, Faculty of Business and Economics, Universiti Malaya, Kuala Lumpur 50603, Malaysia
[3]Department of Educational Psychology and Counselling, Faculty of Education, Universiti Malaya, Kuala Lumpur 50603, Malaysia

Received: 9 April 2024 / Accepted: 14 February 2025
Published online: 07 March 2025

References

1. Institute for Public Health (IPH). National Health and Morbidity Survey 2022 (NHMS 2022): Maternal and Child Health – Key Findings; 2023.
2. UNICEF Malaysia. UNICEF Malaysia Annual Report. 2021. 2022.
3. Arumugam S, Hock KE. The symptomatic behaviour screening tool (symbest) for early identification of developmental delays among children age 3–4. J Pendidikan Bitara UPSI. 2019;12:1–19.
4. Ip P, Li SL, Rao N, Ng SSN, Lau WWS, Chow CB. Validation study of the Chinese early development instrument (CEDI). BMC Pediatr. 2013;13:146.
5. Shekhawat DS, Gupta T, Singh P, Sharma P, Singh K. Monitoring tools for early identification of children with developmental delay in India: an update. Child Neuropsychol. 2022;28:814–30.
6. United Nations. Goal 4 Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all. 2012. https://sdgs.un.org/goals/goal4
7. Goldfeld S, Yousafzai A. Monitoring tools for child development: an opportunity for action. Lancet Glob Health. 2018;6:e232–3.
8. McCoy DC, Sudfeld CR, Bellinger DC, Muhihi A, Ashery G, Weary TE, et al. Development and validation of an early childhood development scale for use in low-resourced settings. Popul Health Metr. 2017;15:3.
9. Ertem IO, Krishnamurthy V, Mulaudzi MC, Sguassero Y, Balta H, Gulumser O, et al. Similarities and differences in child development from birth to age 3 years by sex and across four countries: a cross-sectional, observational study. Lancet Glob Health. 2018;6:e279–91.
10. Kirchhoff C, Desmarais EE, Putnam SP, Gartstein MA. Similarities and differences between western cultures: toddler temperament and parent-child interactions in the United States (US) and Germany. Infant Behav Dev. 2019;57:101366.
11. Chu MN, Le PTN. Language policy strategies of Malaysia, Singapore and Indonesia. J Ind Asn Stds. 2020;1:2050009.
12. Yamat H, Umar NFM, Mahmood MI. Upholding the malay language and strengthening the English language policy: an education reform. Int Educ Stud. 2014;7:197–205.
13. How SY, Chan SH, Abdullah AN. Language vitality of Malaysian languages and its relation to identity. Gema Online J Lang Stud. 2015;15:119–39.
14. WHO. Developmental difficulties in early childhood: Prevention, Early Identification, Assessment and intervention in low- and Middle-Income Countries. Geneva, Switzerland: World Health Organization; 2012.
15. Sabanathan S, Wills B, Gladstone M. Child development assessment tools in low-income and middle-income countries: how can we use them more appropriately? Arch Dis Child. 2015;100:482–8.
16. Doennecke N, Brandenburg J, Maehler C. Cross-cultural measurement invariance of a developmental assessment tool in a small-scale intervention study. Infant Behav Dev. 2023;73:101888.
17. Ismail HIHM, Ng HP, Thomas T. Paediatric protocols for Malaysian hospitals. 4 ed. Malaysian Paediatric Association; 2019.
18. Fernald LCH, Prado E, Kariger P, Raikes A. A toolkit for measuring early childhood development in low and middle-income countries. Washington, DC: World Bank; 2017.
19. Singapore Government Health Promotion Board. Health booklet 2014. https://chapi.healthhub.sg/api/public/content/30de6c1e56d34868afb5fa6df399e082?v=35bba801
20. Centers for Disease Control and Prevention. CDC's developmental milestones; 2020. https://www.cdc.gov/ncbddd/actearly/milestones/index.html
21. Mullen EM. Mullen scales of early learning. Circle Pines, MN: AGS; 1995.
22. Griffiths R, Huntley M. Griffiths mental development scales-revised: Birth to 2 years; 1996.
23. Faust T, Mullis S, Solomon K. Malaysian Development Language Assessment Kit (MDLAK). Kuala Lumpur: Malaysian Care; 1992.
24. Cruchinho P, López-Franco MD, Capelas ML, Almeida S, Bennett PM, Miranda da Silva M, et al. Translation, cross-cultural adaptation, and validation of measurement instruments: a practical guideline for novice researchers. J Multidiscip Healthc. 2024;17:2701–28.
25. Bentler PM, Bonett DG. Significance tests and goodness of fit in the analysis of covariance structures. Psychol Bull. 1980;88:588–606.
26. Lt H, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Model Multidiscip J. 1999;6:1–55.
27. Maydeu-Olivares A. Assessing the size of model misfit in structural equation models. Psychometrika. 2017;82:533–58.
28. Bollen KA. Structural equations with latent variables. Wiley; 2014.
29. Dimitrov DM. Testing for factorial invariance in the context of construct validation. Meas Eval Couns Dev. 2010;43:121–49.
30. Brown TA. Confirmatory factor analysis for applied research. Guilford; 2015.
31. Fernald LC, Kariger P, Engle P, Raikes A. Examining early child development in low-income countries: a toolkit for the assessment of children in the first five years of life. World Bank; 2009.
32. Ertem IO, Atay G, Dogan DG, Bayhan A, Bingoler BE, Gok CG, et al. Mothers' knowledge of young child development in a developing country. Child Care Health Dev. 2007;33:728–37.
33. Hill HD, Morris P, Gennetian LA, Wolf S, Tubbs C. The consequences of income instability for children's well-being. Child Dev Perspect. 2013;7:85–90.
34. Soliman A, De Sanctis V, Alaaraj N, Ahmed S, Alyafei F, Hamed N et al. Early and long-term consequences of nutritional stunting: from childhood to adulthood. Acta Biomed Atenei Parmensis. 2021;92.

Publisher's note