Child Indicators Research
https://doi.org/10.1007/s12187-024-10184-9

Measuring Children's Developmental Status in China Using the ECDI2030: Comparing with Direct Assessment and Teacher Report

Zeyi Li[1] · Nirmala Rao[1]

Accepted: 1 October 2024

Abstract

This study compared different approaches for monitoring progress towards Sustainable Development Goal Target 4.2, which focuses on the proportion of children who are developmentally on track. UNICEF's Early Childhood Development Index 2030 (ECDI2030), a parent report measure, was compared with a corresponding direct assessment measure using a sample of children aged 3 and 5 ($N=309$; 154 girls) in China at two time points. In the second wave, the study also investigated the correlations and agreement between the ECDI2030 and a teacher report measure for children's development. Although Cronbach's alpha indicated that both adult report measures had lower reliability, McDonald's omega showed comparable reliability among the three measures when the assumption of tau-equivalence was relaxed. Moreover, both adult report measures tended to overestimate children's developmental levels, and were less effective in capturing the development of older children compared to the direct assessment measure. The correlations between parent report and direct assessment were significant for both girls and boys, urban children, and children from higher socioeconomic quartiles in both waves. Parents' education levels did not substantially moderate the correlations. Moreover, parent report may not predict children's development as effectively as direct assessment. Compared to parent report, teacher report was less effective in differentiating children's development across socioeconomic status and urbanicity. Parent and teacher judgements were more consistent on children's early learning competencies than on children's motor and psychosocial skills. Implications of the findings for population-based measurement of early childhood development are discussed.

Keywords SDG Target 4.2 · Direct Assessment · Parent Report · Teacher Report · The ECDI2030

Extended author information available on the last page of the article

Published online: 23 October 2024

1 Introduction

The early years, a critical period of development and a foundation for individual well-being, have been increasingly given global recognition. This is evident in the inclusion of an early childhood development (ECD) target, Target 4.2, in the UN Sustainable Development Goals (SDGs) (United Nations, 2015). This target, to be achieved by 2030, aims to ensure quality ECD, care, and education for all children. Two indicators are used to measure progress towards this target. The first indicator, SDG Target Indicator 4.2.1, measures the "proportion of children aged 24–59 months who are developmentally on track in health, learning, and psychosocial well-being, by sex". The other indicator, Indicator 4.2.2, focuses on children's participation rate in organised learning one year before primary school.

UNICEF, as the custodian agency for SDG Target Indicator 4.2.1, developed the Early Childhood Development Index 2030 (ECDI2030) to determine whether children aged 2 to under 5 years are "developmentally on track" (UNICEF, 2023). The ECDI2030 is a parent report (PR) measure which asks caregivers 20 questions about their child's development in the domains of Health, Learning, and Psychosocial Well-being. To be developmentally on track, children are expected to achieve at least seven, nine, 11, 13, and 15 milestones at ages 24 to 29 months, 30 to 35 months, 36 to 41 months, 42 to 47 months, and 48 to 59 months, respectively (UNICEF, 2023). The development of the ECDI2030 has involved extensive methodological work in more than 30 countries since 2016 (UNICEF, 2023). In field testing, items were piloted and screened based on their difficulty level and ability to discriminate competencies (UNICEF, 2023). Cognitive testing was also conducted to ensure items can be interpreted and understood in a comparable way by caregivers across cultural and linguistic backgrounds (Cappa et al., 2021). The ECDI2030 has demonstrated satisfactory reliability and validity (Halpin et al., 2024).

Nevertheless, some concerns have been raised regarding the ECDI2030's reliance on PR for data collection at the population level (Fernald et al., 2017; Raikes, 2017). These concerns include potential biases in parental judgment, such as confirmation bias, halo effect, and overestimations (Callan Stoiber, 1992; Fluck et al., 2005; Muntoni & Retelsdorf, 2019); the influence of parents' expectations and beliefs on the judgment of children's development, which varies by social-cultural contexts (Cintas, 1995; Phillipson & Phillipson, 2007); and whether PR alone can provide a comprehensive understanding of children's developmental status. Given these concerns, validating the ECDI2030 by comparing it with criterion measures or triangulating it with other information sources is necessary.

However, there is limited evidence regarding the associations between findings obtained through the administration of the ECDI2030 items and those obtained through different approaches assessing the same items. Therefore, this study addressed the existing literature gap by comparing different approaches for measuring children's development using the ECDI2030 items. Furthermore, although ECDI2030 is designed to monitor countries' progress towards the SDG Target, it has not been used to measure developmental changes within children over time. Hence, a longitudinal study was conducted to explore the effectiveness of the ECDI2030 in monitoring children's development by correlating developmental changes measured by different assessment approaches. The findings hold significant implications for future research and practice in selecting assessment approaches for measuring ECD at the population level.

1.1 Comparing Different Assessment Approaches to population-based Measurement of ECD

Progress towards SDG Target 4.2 necessitates collecting population-level data on child outcomes. Given the extent of the inequalities in global access to pre-primary education and the varying quality of services received by sub-groups of children within a country (UNESCO, 2022), population-based data on ECD can facilitate regional and global policy formation (McCoy et al., 2018) and help inform children's development across different sociodemographic backgrounds. Although data on child growth (e.g., height and weight) are routinely collected to document young children's nutritional and health status at the population level (de Onis & Blössner, 2003), achieving consensus on measures to track their development and learning remains a challenge (Raikes, 2017). The ECDI2030 was developed to provide a unified framework to monitor the global progress towards SDG Target 4.2. As a PR measure, it is easy to use at scale and can efficiently collect population-level data within existing household survey programs (e.g., Multiple Indicator Cluster Surveys and Demographic and Health Surveys).

Nevertheless, all assessment approaches have strengths and limitations. This study examined the methodological issues associated with using PR to measure children's developmental status by comparing the results with direct assessment (DA) and teacher report (TR). For DA, children were administered standardised tasks or activities by a trained assessor in individual sessions (Sabanathan et al., 2015). This is considered a standard approach for gauging children's "true" abilities and can be used to validate PR (Fernald et al., 2017). Comparing the ECDI2030 with a corresponding DA measure can help determine the sensitivity and accuracy of the ECDI2030 in gauging children's abilities across populations. However, DA can be prohibitive at the population level because of its high cost and resource-intensiveness, requiring trained assessors, specific assessment materials, and translated tools (Fernald et al., 2017). Moreover, DA may lack ecological validity when attempting to understand how children behave in real-life situations (Mccabe et al., 2000). For instance, children can exhibit executive function or self-regulation skills differently during standardised tasks compared to that observed by parents or teachers in the daily environment (Obradović et al., 2018). In such cases, multiple methods may be necessary to comprehensively understand children's holistic development.

Both PR and TR are examples of adult report (AR) measures, in which adults familiar with the children are asked questions to obtain information about the children's development. Using AR measures in large-scale assessments is feasible because it does not require extensive training to score and interpret the items (Lin et al., 2021; Snow & Van Hemel, 2008). Specifically, TR uses a frame of reference to assess a child's abilities relative to his or her peers (Martínez et al., 2009; Meissel et al., 2017; Ready & Wright, 2011), which potentially provides a more accurate prediction of early academic outcomes compared to PR (Blair & Razza, 2007; von

Suchodoletz et al., 2015). However, the accuracy and reliability of AR measures can be influenced by adult characteristics (e.g., education level) (Mashburn & Henry, 2004), psychological factors (e.g., maternal depressive symptoms and teacher's self-efficacy) (Alderman et al., 2021; Furnari et al., 2017), the availability and observability of clues for adults to make judgments (Bodnarchuk & Eaton, 2004), and how adults interpret and attribute the causes of their child's behaviours (De Los Reyes & Kazdin, 2005).

Given the strengths and limitations of different assessment approaches in measuring children's development, choosing the optimal assessment method for children's population-based outcomes poses challenges. One way to shed light on this matter is to examine the correlations among different approaches. Such comparison can help researchers and practitioners make informed decisions about which assessment methods to utilise in various contexts. The following sections discuss how the correlations between PR and DA can vary across populations and how PR and TR provide different information regarding children's development.

1.2 Comparing PR and DA Across Populations and Over Time

Numerous studies have compared PR and DA for measuring young children's learning and developmental competencies or for screening children's social-emotional or behavioural problems. These studies have identified biases in PR such as recall and social desirability biases (Bornstein et al., 2020). Some studies suggested that these biases can arise from differences in the sociodemographic background of children (Bennetts et al., 2016; Johnson et al., 2004). Examining potential biases across sub-groups is crucial for interpreting assessment outcomes relevant to specific populations. If undetected, differences in the assessment outcomes that are systematic across populations could lead to biased collection, misuse, or misinterpretation of data, which can be particularly detrimental when the data are used to inform policy-making.

Notably, PR findings can vary systematically by children's age and sex. For example, PR could be reliable in providing information on infants' gross motor development, as parents perceive such skill acquisitions as important signs of children's "typical development" (Bodnarchuk & Eaton, 2004). As children mature, parents may find it easier to observe and recollect cognitive and social skills acquired as children become more capable of manifesting these abilities in the environment (Snow & Van Hemel, 2008). Child sex also influences the PR of children's developmental outcomes. Gender stereotypes of parents related to children's reading and maths skills were shown to correlate with parents' beliefs in these abilities (Jacobs, 1991; Muntoni & Retelsdorf, 2019). Moreover, adults may overestimate the prevalence of behaviour problems in boys compared to girls (Chen, 2010; Walker, 2004).

Parents' socioeconomic status (SES) has been shown to be associated with their judgment of children's development, and parents' education plays an important role. Parents' education levels and accessibility to written materials were found to be positively associated with maternal knowledge (Benasich & Brooks-Gunn, 1996; Bornstein et al., 2010). Parents with more parenting knowledge were more likely to have appropriate expectations regarding their children's skill acquisition and developmental milestones (Stoiber & Houghton, 1993), thereby facilitating their judgment of their children's development. Parents' education and family income were also found to relate to the home learning environment of the children (Hart et al., 2016). Parents with a higher SES have more resources and provide more intellectually stimulating home activities for children than low SES parents (Kalil & Ryan, 2020; Rao et al., 2021), hence, the former can have more opportunities to observe specific skills in their children. Such experiences were found to influence parents' understanding and confidence when evaluating children's fine-grained skills (Zippert & Ramani, 2017). Lastly, it has been suggested that parents with high educational levels are more capable of comprehending survey questions (Dinnebeil et al., 2013), which can contribute to the increased accuracy of PR.

Conversely, some studies found no significant associations between parents' education level and the reliability of the evaluations (Alderman et al., 2021; Bedore et al., 2011). Likewise, Rao et al. (2021) observed only a marginally significant difference in the agreement between DA and PR for assessing children's developmental status in individuals with an "above secondary" education level compared to those with "secondary education or lower". These inconsistent findings suggest other explanations for the relation between parents' education level and the reliability of their reports. One possibility is that variations in parents' education levels may not be associated with the accuracy of the report of their children's developmental milestones when the overall variance in education levels of parents within a population is low, and there is not much difference in the children's attainment of a particular milestone. Additionally, there could be a non-linear relationship between parents' education level and report accuracy. Considering the mixed findings from previous literature, this study investigated differences in the correlations between DA and PR across different populations, specifically focusing on family SES.

In addition to examining differences in the correlations across groups, this study also compared the sensitivity of PR and DA in measuring children's developmental growth over time. Monitoring children's developmental status requires measures that can consistently predict children's development across various groups. Although previous studies have demonstrated PR's predictive power in measuring children's development (Feldman et al., 2005; Rubio-Codina & Grantham-Mcgregor, 2020; Stone et al., 2010), less attention has been paid to comparing between PR and DA in capturing children's growth across populations. To provide insights into this issue, the current study measured children's outcomes at two time points using both PR and DA to compare their predictive power and their consistency in measuring gains in children's development over time. Longitudinal comparisons between DA and AR can enhance our understanding of the trade-offs when choosing the assessment method for monitoring young children's development at the population level.

1.3 Comparing Between PR and TR

Home and school are two proximal environments that have significant influences on child development, as highlighted in Bronfenbrenner's ecological systems theory (1979). When assessing child development, it is important to consider reports from both parents and teachers, as they provide unique perspectives and represent different contexts of children’s behaviours, performance, and development (De Los Reyes et al., 2015; Bergold et al., 2019). Cross-informant discrepancies between teachers and parents have been observed in measuring children’s problem behaviours, social-emotional skills, psychosocial problems, and approaches to learning, with these discrepancies often attributed to the situational specificity of children’s behaviours across different settings (Berg-Nielsen et al., 2012; Bergold et al., 2019; Li et al., 2019; Winsler & Wallace, 2002).

Extant findings have mainly focused on children’s non-cognitive or social-emotional competencies, which are more subject to contextual variations upon assessment (Jones et al., 2016). Less evidence is available regarding the agreement between parents and teachers when measuring children’s development in other domains. The ECDI2030, which covers domains of Health, Learning, and Psychosocial well-being, is a valuable resource for comparing teachers’ and parents’ perceptions across these domains. For example, in the learning domain, it is possible that children’s learning competencies are exhibited more consistently across school and home environments, leading to a higher concordance between TR and PR. On the other hand, it is also plausible that teachers and parents possess discrepant knowledge about children’s development, particularly if their judgments are based on different types of learning activities and interactions with children in the home and school environments.

In the ECDI2030, interpretation is based on the total score of 20 items across the three domains; hence, comparing TR and PR at the domain level may be inappropriate (Halpin et al., 2024). In this study, we calculated the agreement between two AR measures at the item level to investigate whether agreements differed across domains. These insights can contribute to a more comprehensive understanding of the consistency and variation of perceptions between teachers and parents regarding children’s development in different contexts.

1.4 Methodological Considerations when Comparing Different Approaches

Understanding the associations among different assessment methods requires considering the alignment between assessment contexts, including assessment content, question/item formulation, and scoring. Without considering the degree of correspondence across assessments, interpreting the results generated by these different approaches would be challenging. The concordance between different approaches can vary depending on the assessment content. For example, there was a lower correlation between DA and AR when measuring children’s social-emotional development and behavioural problems compared to learning competencies or overall developmental status (Bergold et al., 2019; Pushparatnam et al., 2021; Waldman et al., 2021). Variations in the correlations between PR and TR were also present when measuring different subskills within the same domain (Li et al., 2019; Massa et al., 2008). Furthermore, characteristics, such as the wording of the question, response format, and scoring rubrics, can also affect adults’ judgment accuracy (Hoge & Coladarci, 1989). Considering the assessment context, the current study investigated the correlations among PR, DA, and TR using matched items, comparable response formats, and consistent scoring rules. This approach ensures the assessments are aligned and can better reveal the similarities and differences in children's outcomes measured across PR, DA, and TR.

2 The Current Study

This study investigated the psychometric robustness and contextual appropriateness of the ECDI2030 for monitoring progress towards SDG Target Indicator 4.2.1 in China. To achieve this, the ECDI2030 was compared with corresponding DA and TR measures. As a middle-income country with a relatively high enrolment rate of young children in preschool, China has been effectively monitoring the nation's progress towards SDG Target Indicator 4.2.2, which focuses on the proportion of children participating in organised learning before primary school. The Ministry of Education (MoE) collects annual administrative data on the gross enrolment rate (GER) of young children in preschool, which reached 89.7% in 2022 (Ministry of Education, 2023). However, there remains a lack of nationally representative data on children's developmental status to inform Early Childhood Care and Education (ECCE) policies (Li & Rao, 2023), which is an issue not specific to China (UNICEF & UNESCO, 2024).

Data on children's developmental status are helpful in evaluating whether ECCE provisions are effective in mitigating children's developmental gaps across populations. Studies have shown there are developmental inequalities across regions, urbanicity, and family SES (Rao et al., 2022; Su et al., 2021; Wang et al., 2022), which are linked to young children's differences in ECCE experiences.

On the one hand, urban-rural gaps in ECCE exist in various areas, such as higher ECCE quality, earlier starts, and more stable parental support and public funding available to urban children (Hu et al., 2016). In contrast, ECCE services in rural areas face challenges including lower quality, inadequate allocation of human and material resources, as well as lacking supervision and evaluation (Hong et al., 2015; Jin, 2008). Positive associations between structural quality (e.g., teacher qualification and teacher-child ratio) and process quality (e.g., teacher-child interaction) in ECCE and ECD have been found in China (Li et al., 2016; Wang et al., 2020).

On the other hand, rapid urbanisation and increased nationwide rural-to-urban mobility have raised concerns about the potential effect of parental migration on young children's participation in quality ECCE in urban areas. Specifically, due to place-based public resource distribution and management systems (Wen & Lin, 2012), migrant children (i.e., rural children who have moved to cities with their families) without urban household registration (*Hukou*) are unable to access public services and education in urban areas (Chen & Feng, 2013). Thus, they are less likely to receive a high-quality education in urban China. Moreover, low access to high-quality ECCE programs and unfavourable family environments are likely to occur concurrently for children living in socially disadvantaged conditions. Using the China Family Panel Studies data, studies have found that parental migration was negatively associated with children's ECCE participation as well as home environment quality (Gong & Rao, 2023; Xie et al., 2021).

To evaluate China's progress towards SDG Target 4.2, it is crucial to track children's developmental status to provide evidence of disparities in early development and inform policies to promote social justice. Although previous research has examined ECD in the Chinese context to inform policy, there has been limited systematic investigation of methodological issues in generating child outcome data at the population level. Given the large number of young children in China, adopting AR is more practicable than DA for measuring child development at the population level. Moreover, PR via electronic surveys is scalable due to the national high literacy rate and high internet penetration. Additionally, TR is another viable approach to monitor child development, considering the high preschool enrolment rate. Contextualising the measurement of SDG Target Indicator 4.2.1, this study compared different approaches for measuring children's developmental status in China. The findings can contribute to recommendations for psychometrically robust methods and indicators for the measurement of early development at the population level in China.

This study aims to examine the reliability and criterion validity of the ECDI2030 in measuring children's developmental status by comparing it with two corresponding measures (i.e., direct assessment and teacher report) in the Chinese context and to investigate whether the correlations among PR, DA, and TR would differ across children's sociodemographic background. There are four research questions: (1) Are there differences in the reliability and validity of DA, PR, and TR for measuring children's developmental status in the current sample? (2) What are the correlations between PR and DA for measuring children's development and developmental growth across child age, sex, urbanicity, SES, and parents' education levels? (3) What are the correlations among PR, TR, and DA across child age, sex, urbanicity, SES, and parents' education levels? and (4) Does the agreement between PR and TR vary across items in the Health, Learning, and Psychosocial Well-being domains? Data were collected on children's developmental status repeatedly using the ECDI2030 and a DA measure at two time points. Additionally, at Wave 2, TR data were collected and compared with ECDI2030.

3 Method

3.1 Study Setting

Data in the current study were collected from Beijing and Hebei Province in China between 2021 and 2022. Beijing is more economically and educationally advantaged than Hebei. According to the Seventh National Population Census in 2020, Beijing had the highest gross domestic product (GDP) per capita (¥164,900) in China, which was nearly 3.4 times that of Hebei Province. In terms of education, more than 42.9% (9.19 million) of Beijing's total population held at least a college diploma, whereas the corresponding figure for Hebei was 7% (5.24 million) (Beijing Municipal Bureau of Statistics, 2022; Hebei Provincial Bureau of Statistics, 2022).

3.2 Measures and Scoring

3.2.1 PR–ECDI2030

The ECDI2030 was used to assess children's overall developmental status. Specifically, the items in the ECDI2030 cover three domains of children's development: Health, Learning, and Psychosocial Well-being. Items 1 to 4 assess children's health, including gross and fine motor skills (e.g., Can (name) dress himself/herself, that is, put on pants and shirt without help?). Items 5 to 15 assess children's learning capabilities, including expressive language, literacy, prewriting, numeracy, and executive functioning (e.g., Can (name) say at least 10 or more words like 'Mama' or 'ball'?). Items 16 to 20 assess children's psychosocial well-being, including social skills, helpfulness, and internalising and externalising behaviours (e.g., Does (name) get along well with other children?).

All ECDI2030 item responses are "Yes", "No", and "Don't know", except for items 19 and 20 that require a frequency response. Responses for item 19, "How often does the child seem to be very sad or depressed?" were "Daily", "Week", "Monthly", "A few times a year", and "never". Reponses for item 20, "How much does the child kick, bite or hit other children", were "Not at all", "Less", "The same", "More", and "A lot more". The scoring in the ECDI2030 followed the UNICEF guidelines. A response of "Yes" was scored as "1", "No" was scored as "0", and "Don't Know" was scored as "0" (UNICEF, 2023). For item 19, a score of "0" was given if the child was reported to be "very sad or depressed daily". For item 20, a score of "1" was given if the child did not engage in kicking, biting, or hitting other children at all, or if their frequency of such behaviours was the same or less compared to their peers. Each item was individually assigned a binary score, and the mean score was calculated across all 20 items for each child, giving a continuous score ranging from 0 to 1.

Recognising the potential bias introduced by coding "Don't Know" responses as "0", we explored an alternative approach. In this study, we recoded the "Don't Know" responses as missing values. This allowed us to use pairwise deletions when calculating children's mean scores rather than scoring them as "0". Moreover, we examined the agreement between TR and PR at the item level under two scenarios: recoding "Don't Know" responses as "0" and retaining them as "Don't Know". This analysis enabled us to assess the impact of scoring "Don't Know" responses on the consensus between TR and PR.

3.2.2 TR–ECDI2030

The TR measure used the same 20 items as in the ECDI2030, except for item 16. In the PR measure, item 16 asked parents, "Does (*name*) ask about familiar people other than parents when they are not there, for example, 'Where is Grandma?'". To make this item relevant to a preschool context, it was reframed in the TR version as "Does the child ask about his/ her friends when they are not at school, for example, 'Where is John?' Or 'Where is Anna?'". The scoring guidelines for the TR version were identical to that of the PR version. Each item was coded as "0" or "1", and a total score was calculated as the mean score across the 20 items for each child.

3.2.3 DA–Early Childhood Development Assessment Scale (ECDAS) (iPad)

The Early Childhood Development Assessment Scale (ECDAS) was originally developed based on a pool of items from ECDI. The scale was designed to assess the Health, Learning and Psychosocial well-being of children aged 3 to 5 years (Rao et al., 2022). The ECDAS has been implemented in Bangladesh, China, India, and Myanmar. The scores are significantly related to children's sociodemographic factors (Rao et al., 2022). Moreover, Confirmatory Factor Analysis found ECDAS had acceptable reliability when measuring child development at the domain level (Richards et al., 2023).

In this study, we developed an iPad version of ECDAS, which incorporated the corresponding 20 items from the ECDI2030. During the DA session, trained assessors utilised iPads to administer ECDAS to the children. For example, when evaluating children's ability to name objects, various pictures of common objects were displayed on the iPad and the child was asked to name them in the given order. Some tasks were designed to assess children's development across multiple ECDI2030 items. For example, children's performance in expressive language skills, which involved saying different words and sentences, was evaluated through a storytelling task. The assessors scored the items based on the number of words and sentences used by the child during the task. In cases where it was challenging to implement the items on an iPad (e.g., Can (name) walk on an uneven surface, for example, a bumpy or steep road, without falling?), standardised one-on-one tasks were employed. These tasks utilised specially prepared physical test materials to assess the children's abilities.

To capture a broader range of children's capabilities, sub-items were devised within the ECDAS (iPad) items. For example, one of the ECDI2030 items asked parents whether their child can recognise at least five Chinese characters. In the ECDAS (iPad), children were rated on their abilities to recognise three Chinese characters as well as five Chinese characters, and the scores were calculated by averaging the scores across these two sub-items.

Scores for each directly assessed item ("No" = "0", "Yes" = "1", or "Reject") were recorded by the assessors and uploaded automatically through an iPad. When a child refused to give an answer, the "Reject" response was recorded by the assessor and retained for further analysis. All items were scaled from 0 to 1 despite the number of sub-items, and a total score for the DA was calculated as the mean score across the 20 items. The scale and length of PR and DA were the same, except that the DA aimed to capture more variability in children's development. Table 1 provides exemplar items and response formats for the PR, DA, and TR items of the ECDI2030.

3.3 Procedure

The first wave of data collection took place between May and August 2021, targeting children between the ages of 3 and 4. The sampling process involved stratification based on age, sex, and urbanicity. Urban children were selected from kindergartens in Beijing, whereas their rural counterparts were sampled from kindergartens in Hebei Province. Prior to data collection, written informed consent was obtained from both kindergarten principals and parents. A local assessment team consisting of university

Table 1 Corresponding exemplar items of measures

Assessment names	Assessment domains	Exemplar items/sub-items	Response Formats
ECDI2030			
(Same items for PR and TR)	Health	ECD3. Can (name) dress (him/herself), that is, put on pants and a shirt without help?	Yes/No
	Learning	ECD6. Can (name) speak using sentences of three or more words	Yes/No
	Psychosocial well-being	ECD19. How often does (name) seem to be very sad or depressed?	Daily/ Weekly/ Monthly/A few times a year/ Never /Don't know
ECDAS (iPad) (DA)			
	Health (Through standardised task)	ECD3.1 Put on a T-shirt	Yes/No
		ECD3.2 Put on shorts	Yes/No
	Learning (Through iPad)	ECD6.1 Says at least one sentence with three or more words	Yes/No
		ECD6.2 Says at least two sentences with three or more words	Yes/No
		ECD6.3 Says at least three sentences with three or more words	Yes/No
	Psychosocial well-being (Through iPad)	ECD19. Resonate himself/ herself with being a happy child	Yes/No

students majoring in Early Childhood Education (ECE) was responsible for conducting the data collection. To ensure consistency and reliability, the research team provided training to the assessors via Zoom on how to use the measures. The inter-rater consistency was also examined.

During the first wave, children were assessed by ECDAS (iPad), and their parents were asked to rate their children's behaviours on the ECDI2030 items. Sociodemographic information, including child age, sex, residence (urban or rural area), maternal and paternal education levels, maternal and paternal occupations, and annual family income, was collected through parent questionnaires. Both the ECDI2030 survey and parent questionnaires were distributed through an online platform. At this stage, a TR measure was developed and piloted on kindergarten teachers in these districts.

The second wave of data collection took place a year later, between April and September 2022. Researchers conducted follow-up assessments with the same children who participated in the first wave. Similarly, children completed the ECDAS (iPad) measure and parents completed the ECDI2030 items relating to their child's development. In this wave, kindergarten teachers of the same children completed the developed TR survey. Teachers were also asked to fill out a background questionnaire for information on their age, sex, position, education level, and professional experience as early childhood educators.

3.4 Sample

In Wave 1, 447 children between the ages of 3 and 4 had both DA and PR data. In Wave 2, 365 children had DA data and 333 had PR data. Sample attrition for the DA data was not significantly related to child age, sex, urbanicity, or family income. However, in Wave 2, parents of urban children and those from higher-income families were more likely to complete the parent survey.

A final sample of 309 children (154 girls) aged 3 to 5 with data from both waves were used in the analysis. The average age of children in Wave 1 was 49.66 months (SD = 5.50) and in Wave 2 was 62.31 months (SD = 5.69). A total of 79 teachers completed the TR in Wave 2, and on average, each teacher rated 6.87 children. Importantly, the TR data were only collected in the second wave.

A small percentage of children were over the age of 4 in Wave 1 (0.01%) and were included in the 4-year-old group. Similarly, a few children were over the age of 5 in Wave 2 (0.02%) and were included in the 5-year-old group. However, these children were not included in the longitudinal analysis, which focused exclusively on children who were 3 years old in Wave 1 and 4 years old in Wave 2.

In this study, 90.44% of the PR respondents were mothers. Family SES index was calculated from the principal component analysis of family income, parents' education level, and parents' occupation in Wave 1, which was also used in the subsequent analyses. Analysis of maternal education levels revealed that mothers in the current sample were relatively privileged. Nearly half of the children had mothers with at least a bachelor's degree. For parents with education levels below a bachelor's degree, they were categorised as junior secondary education and below or senior secondary education and junior college, with equivalent percentages in each level.

Regarding teacher characteristics, 99% of them were female, and over 90% were the class teacher for the child. More than half of teachers had a junior college or a bachelor's degree. Descriptive information on the sample characteristics is provided in Table 2. The bivariate correlations between key variables are presented in Table 3.

3.5 Missing Data

Missingness was observed in the background questionnaires for parents and teachers. In Wave 2, the missingness ranged from 15 to 16% for parents' education level, occupations, and family income, which was higher than in Wave 1 (0.7–1%). Information on children's background questionnaires including SES and maternal education level from Wave 1 were used in the analyses. Missingness was also present in teacher

Table 2 Descriptive sample characteristics

	Mean (*SD*)	Range
Sample Size n(%)		
Total	309	
3 to 4-year-olds	138 (44.66%)	
4 to 5-year-olds	168 (54.37%)	
Child age (months) (Wave 1)	49.66(5.50)	36.1-68.27
Child age (months) (Wave 2)	62.31(5.69)	48.1-83.73
Child sex [a]	0.49(0.50)	
Urbanicity [b]	0.56 (0.50)	
SES *n* (%)		
Bottom quartile	81(26.30)	
2nd quartile	73(23.70)	
3rd quartile	77(25.00)	
Top quartile	77(25.00)	
Maternal Education *n* (%)		
Below bachelor's degree	163(52.75)	
Junior secondary and lower	72(23.30)	
Senior secondary and junior college	91(29.45)	
Bachelor's degree and above	146(47.25)	
Teacher age *n* (%)		
24 or under	8(10.13)	
25-29	30(37.97)	
30-34	16(20.25)	
35-39	13(16.46)	
40-44	7(8.86)	
45-49	3(3.80)	
50-54	2(2.53)	
Teacher sex[c]	0.99(0.11)	
Teacher education level *n* (%)		
Secondary school	9(11.54)	
Junior College	38(48.72)	
Bachelor's degree	31(39.74)	
Teacher major *n* (%)		
Early Childhood Education	52(67.53)	
Other	25(32.47)	
Teacher position *n* (%)		
Class teacher	74(94.87)	
Senior teacher	2(2.56)	
Vice Principal	1(1.28)	
Other	1(1.28)	
Teacher ECE work experience	7.91(5.93)	1-30

[ac] Male = 0 Female = 1. [b] Rural = 0 Urban = 1

background questions (1% each for position, education level, and major). Pairwise deletions were used in the calculations involving these variables.

No missingness was found for item and total scores in the PR and TR measures. However, the DA measure had missingness at the item level when children refused to respond. In Wave 1, items relating to storytelling had the highest percentage of "Reject" responses (26.86%). In Wave 2, "Reject" responses ranged from 0 to 5%.

Table 3 Bivariate correlations between key variables

	N	1	2	3	4	5	6	7	8	9	10
1 Child age	309	-									
2 Child sex[a]	309	–.02	-								
3 Urbanicity [b]	309	.08	–.03	-							
4 SES	308	.06	–.04	.84***	-						
5 MEdu	309	.10	–.05	.81***	.93***	-					
6 DA (Wave 1)	309	.60***	.15**	.37***	.30***	.33***	-				
7 DA (Wave 2)	309	.42***	.04	.50***	.47***	.49***	.57***	-			
8 PR (Wave 1)	309	.33***	.11*	.33***	.32***	.32***	.47***	.47***	-		
9 PR (Wave 2)	309	.26***	.10	.32***	.35***	.33***	.38***	.44***	.53***	-	
10 TR	309	.27***	.15**	.04	.11	.09	.25***	.28***	.20***	.19**	-

[a] Boys=0 Girls=1. [b] Rural=0 Urban=1. MEdu=Maternal Education. DA=Direct Assessment. PR=Parent Report. TR=Teacher Report. *$p<.05$. **$p<.01$. ***$p<.001$

We examined whether the reject responses were significantly associated with children's younger age, which might indicate that children were too young to accomplish the tasks. We also compared the proportion of "Reject" responses with the "Don't Know" responses from the two AR measures (Online Resource 1), as children might be shy or reluctant to respond due to unfamiliarity with the test setting. The findings indicated that some of the reject responses observed in the DA measure did not necessarily reflect the child's ability to complete the task.

Multiple imputations were conducted using the Predictive Mean Matching technique to retain as many cases as possible. This technique estimates missing values based on complete cases that have predicted values close to the missing cases (Woods et al., 2024). This method is more robust for non-normality data compared to other approaches (van Ginkel et al., 2020). Children's age, sex, and urbanicity were added as predictors in the chained equation used for imputing plausible values for the missing assessment outcomes.

Twenty datasets were imputed, and the mean and range were checked for reasonability. Regression analysis results were similar between datasets that used listwise deletions, pairwise deletions (Online Resource 2), and multiple imputations to handle missing data. For this study, pairwise deletions were used for descriptive analyses and correlations, while imputed data were used for regression analyses.

3.6 Analytical Methods

For the first research question, we assessed the internal consistency and criterion validity of the three measures. Cronbach's alpha and McDonald's omega were calculated to gauge internal consistency. Specifically, the alpha coefficient evaluates the extent of the inter-item correlations for measures assumed to be unidimensional (Cortina, 1993). A measure with a Cronbach's alpha larger than 0.70 is usually considered to have satisfactory internal consistency (Nunnally & Bernstein, 1994). The omega coefficient is a generalised form of the alpha coefficient and measures the proportions of the scale variance attributed to the primary factor (McDonald, 1999). Unlike the alpha coefficient, the omega coefficient does not assume tau-equivalence, which implies that each item on the scale measures the latent variable with the same degree of precision (Hayes & Coutts, 2020). Therefore, the omega coefficient is considered more accurate than the alpha coefficient when the assumption of tau-equivalence is not met (Trizano-Hermosilla & Alvarado, 2016).

For criterion validity, separate Ordinary Least Square (OLS) regressions were conducted to examine the associations between children's assessment outcomes and key sociodemographic variables. In the OLS regressions, child age and sex were controlled except for regressions, where they were used as independent variables. Based on previous studies that utilised ECDI to measure the level and inequality in ECD using global secondary data (Halpin et al., 2024; Lu et al., 2020; McCoy et al., 2016), we hypothesised that children's assessment scores generated by the three measures in the current sample would be positively associated with children's older age, females, urban residence, and higher SES. The analyses were repeated with "Don't Know" responses treated as missingness.

We used Spearman's rank correlation (rho) for the second research question to examine associations between assessment outcomes. Unlike Pearson's *r* correlation, which assumes linear associations between two variables on continuous scales, Spearman's rank correlation tests the monotonic trends in the correlations between the rank of two variables in a sample (Sedgwick, 2014). Moreover, Spearman's rho is also considered to be less sensitive to outliers (Zar, 2014). The Spearman's rho coefficient ranges from −1 to 1, with a larger value indicating a stronger correlation between the two variables. We calculated Spearman's rank correlations between DA and PR in each wave, respectively, with particular focus on the correlations within the age range targeted by SDG Indicator 4.2.1 (i.e., 2 to under 5 years of age).

For comparing across sociodemographic groups, we calculated the correlations between each assessment outcome after adjusting for the influence of children's age and sex. To this end, standardised residuals were generated from regression analyses of the child assessment outcomes on their age and sex. We then calculated correlations between these residuals for the different groups based on urbanicity, SES, and maternal education.

To examine the predictiveness of DA and PR on children's later development, we conducted regression analyses by regressing Wave 2 outcomes on Wave 1 outcomes measured by the same approach. Child age and sex were controlled for in these regressions. Next, the similarities between the gains in the children's development measured by PR and DA over time were investigated by correlating the standardised residuals obtained from the regression analyses. Finally, correlations between the gain scores across populations were computed and compared.

To answer the third research question, we examined whether the correlations between TR and DA differed from those between PR and DA. We further investigated whether correlations between DA and either of the AR measures would differ from those between the two AR measures. Similarly, correlations were generated by age group, child sex, and across populations. When generating the residuals from the regressions that predicted TR outcomes of children, the teacher factor was added as a fixed effect.

Cohen's Kappa coefficients were calculated to evaluate inter-rater agreement at the item level, addressing the fourth research question. Two different coding approaches were used for calculating the Kappa coefficients, coding "Don't know" responses as

"0" and keeping them as "Don't know". Cohen's Kappa coefficients are commonly used to evaluate inter-rater agreement when assessing subjects on categorical scales (Warrens, 2015). It is also useful for calculating the observed agreement corrected for chance (Warrens, 2015). The interpretations of the Kappa statistics in this study were based on the established benchmarks: values between 0.00 and 0.20, 0.21 to 0.40, 0.41 to 0.60, 0.61 to 0.80, and 0.80 to 1.00 indicated slight agreement, fair agreement, moderate agreement, substantial agreement, and perfect agreement between two measures on the same subject, respectively (Landis & Koch, 1977). All analyses were conducted in STATA 17 (StataCorp, 2021).

4 Results

4.1 Reliability and Validity of the Three Measures

Overall, the findings showed that PR tended to yield higher estimations on children's development compared to DA in terms of their "developmentally on track" status and mean assessment scores. Specifically, the percentage of children's "on track" status measured by DA and PR was 53.65% and 58.77% in Wave 1 and 94.08% and 95.61% in Wave 2, respectively. When "Don't know" responses were treated as missingness, the overall scores were even higher than when coded as "0" (Table 4). Figure 1 illustrates the raw scores of ECDI2030 and ECDAS across child age using the lpolyci regression function in STATA, which revealed a gap between the scores generated by DA and PR. Both AR measures were shown to be less sensitive than DA in reflecting the development of 4-year-olds.

The internal consistency of PR indicated by Cronbach's alpha was 0.69 in Wave 1 and 0.60 in Wave 2. The internal consistency for DA was better than for PR, with a Cronbach's alpha of 0.75 in Wave 1 and 0.68 in Wave 2. For TR, Cronbach's alpha was 0.66 in Wave 2. Specifically in children under 5 in Wave 2, Cronbach's alpha for DA, PR, and TR were 0.70, 0.66, and 0.67, respectively. When the "Don't know" response was treated as missing, the alpha coefficients were lower for PR in Wave 1 (alpha = 0.62), Wave 2 (alpha = 0.47), and for children under five (alpha = 0.57). The internal consistency for TR also slightly decreased both for the total sample (alpha = 0.65) and the 4-year-olds in Wave 2 (alpha = 0.62).

Next, the McDonald's omega coefficients for the three measures were calculated for the entire sample. For the DA measure, The omega coefficients were comparable to the alpha coefficient in Wave 1 (omega = 0.74), but slightly lower in Wave 2 (omega = 0.60). In the case of PR, the omega coefficients were higher in Wave 1 (omega = 0.72) and similar in Wave 2 (omega = 0.60) compared to their corresponding alpha coefficients. The omega coefficient for TR was similar to its alpha coefficient (omega = 0.69). When "Don't know" responses in PR and TR were recoded as missing values, the omega coefficients dropped to 0.65, 0.55, and 0.67 for PR in Wave 1, PR in Wave 2, and TR, respectively.

For criterion validity, regression analyses showed that all measures were significantly and positively associated with child age. However, inconsistencies were found in relation to child sex. In Wave 1, both DA and PR showed girls had higher scores

Table 4 Children’s “Developmentally on track” status and mean scores assessed by different measures

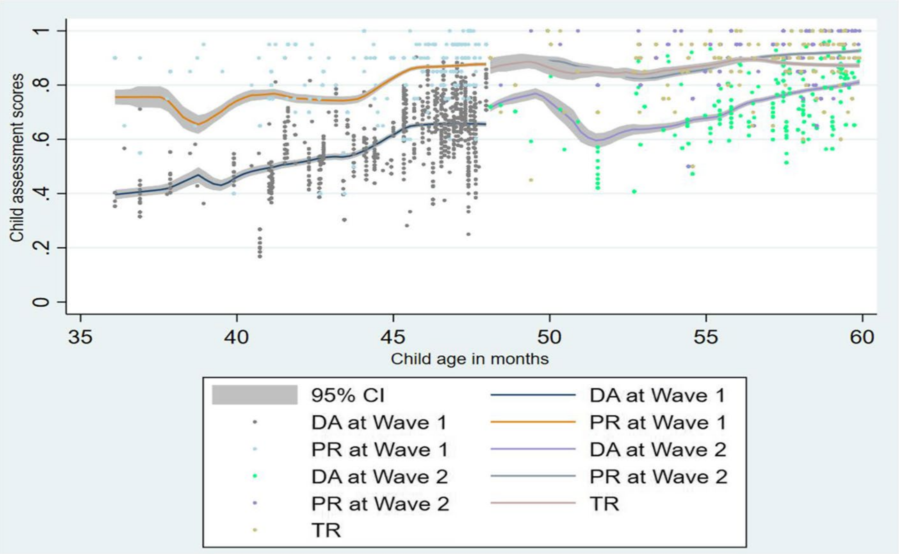
	Wave 1			Wave 2		
		M(*SD*)	Range		*M*(*SD*)	Range
Children’s “Developmentally on track” status (%)						
DA	36–42 months	40.74		48–60 months	58.77	
	42–48 months	63.06				
	48–60 months	57.14				
PR	36–42 months	96.30		48–60 months	95.61	
	42–48 months	91.89				
	48–60 months	94.05				
TR				48–60 months	87.72	
Assessment scores (Coded “don’t know” as “0”)						
DA	3-year-olds	0.62(0.15)	0.20–0.90	4-year-olds	0.79(0.12)	0.41-1
	4-year-olds	0.75(0.13)	0.23–0.99	5-year-olds	0.86(0.09)	0.56-1
PR	3-year-olds	0.83(0.13)	0.40-1	4-year-olds	0.92(0.09)	0.50-1
	4-year-olds	0.89(0.09)	0.45-1	5-year-olds	0.95(0.06)	0.65-1
TR				4-year-olds	0.87(0.10)	0.45-1
				5-year-olds	0.91(0.09)	0.50-1
(Coded “don’t know” as missingness and use pairwise deletions)						
PR	3-year-olds	0.86(0.11)	0.50-1	4-year-olds	0.93(0.07)	0.65-1
	4-year-olds	0.91(0.08)	0.56-1	5-year-olds	0.96(0.04)	0.75-1
TR				4-year-olds	0.89(0.10)	0.56-1
				5-year-olds	0.93(0.09)	0.53-1

than boys, but this pattern was not observed in Wave 2. However, TR outcomes for girls were significantly higher than those for boys in Wave 2. In terms of socioeconomic background, children from urban areas and with higher SES had better assessment results measured by DA and PR. However, there was no significant association between children’s urbanicity and SES with TR. Detailed regression results are presented in Table 5.

The regression analyses produced similar results when “Don’t know” responses from AR were coded as missing, except that TR outcomes were significantly and positively associated with children’s SES. Online Resource 3 provides more details of the regression results using a different scoring approach.

4.2 Correlations Between PR and DA Across Populations and Over Time

Overall, the correlations between the PR and DA were lower for older children compared to younger children. At Wave 1, the correlations between PR and DA for 3-year-olds and 4-year-olds were 0.50 ($n=138, p<.001$) and 0.34 ($n=168, p<.001$),



Note. The sample was a matched sample. N = 138.

Fig. 1 DA, PR, and TR scores of children between 3 and 4 years

respectively. In Wave 2, the correlation between PR and DA for 5-year-olds was 0.18 (n=168, p<.05), which was lower than the correlation for 4-year-olds (rho=0.55, n=138, p<.05).

When examining the correlations based on child sex, PR showed a slightly higher correlation with DA for girls (rho=0.50, n=154, p<.001) than for boys (rho=0.43, n=155, p<.001) in Wave (1) In Wave 2, the correlation was stronger for boys (rho=0.41, n=155, p<.001) than for girls (rho=0.36, n=154, p<.001). Nevertheless, all correlations had a medium strength. When considering urbanicity, the correlations between DA and PR were significant for urban children in both waves, with values of 0.41 (n=173, p<.001) in Wave 1 and 0.22 (n=136, p<.01) in Wave (2) For rural children, the correlation was only significant (rho=0.27, n=136, p<.01) as they grew older in Wave 2.

Correlations across SES and maternal education were also examined. The correlations did not display a linear pattern across SES. In Wave 1, correlations from all SES quartiles were significant, except for the bottom quartile, whereas PR from children in the third quartile (rho=0.42, n=77, p<.001) and the top quartile (rho=0.41, n=77, p<.001) showed stronger correlations with DA outcomes. In Wave 2, PR from children in the bottom quartile (rho=0.27, n=81, p<.05), the third quartile (rho=0.29, n=77, p<.05), and the top quartile (rho=0.28, n=77, p<.05) were significantly correlated with the DA results, but not for children in the second quartile.

When comparing the correlations across maternal education levels, a different pattern was observed from that in the SES analysis. In both waves, reports from mothers with all education levels were significantly correlated with DA, except for those with senior secondary education and junior college degrees. Overall, there was a higher correlation between PR and DA for parents with bachelor's degrees and above

Table 5 Associations between child assessment scores and sociodemographic variables

		Estimate	SE	95% CI		β	*p*	Adjusted R^2
				LL	UL			
DA (Wave 1)								
	Age in months	.02	.00	.01	.02	.57	.000	.33
	Sex [a]	.05	.01	.03	.08	.18	.000	.36
	Urbanicity [b]	.10	.01	.07	.12	.34	.000	.47
	SES	.02	.00	.01	.03	.29	.000	.44
DA (Wave 2)								
	Age in months	.01	.00	.01	.01	.42	.000	.17
	Sex [a]	.01	.01	-.00	.04	.06	.285	.17
	Urbanicity [b]	.11	.01	.09	.13	.47	.000	.39
	SES	.03	.00	.02	.03	.45	.000	.37
	DA (Wave 1)	.40	.05	.30	.49	.51	.000	.34
PR (Wave 1)								
	Age in months	.01	.00	.00	.01	.33	.000	.11
	Sex [a]	.03	.01	.00	.05	.12	.028	.12
	Urbanicity [b]	.07	.01	.05	.09	.31	.000	.21
	SES	.02	.00	.01	.02	.31	.000	.21
PR (Wave 2)								
	Age in months	.00	.00	.00	.01	.26	.000	.06
	Sex [a]	.02	.01	-.00	.03	.11	.056	.07
	Urbanicity [b]	.05	.01	.03	.06	.30	.000	.16
	SES	.01	.00	.01	.02	.34	.000	.19
	PR (Wave 1)	.34	.04	.27	.41	.49	.000	.28
TR								
	Age in months	.00	.00	.00	.01	.27	.000	.07
	Sex [a]	.03	.01	.01	.05	.16	.004	.09
	Urbanicity [b]	.00	.01	-.02	.02	.02	.726	.09
	SES	.01	.00	-.00	.01	.10	.063	.10

N = 309; [a] Boys = 0 Girls = 1; [b] Rural = 0 Urban = 1

(rho = 0.39, $n = 143$, $p < .001$) when compared to parents with lower education levels (rho = 0.25, $n = 163$, $p < .01$) in Wave 1. However, this difference diminished in Wave 2.

Next, we focused on changes in the correlations of the longitudinal sample, which included children aged 3 and 4 over time. The strength of correlations did not substantially change for child age, with coefficients of 0.50 ($n = 138$, $p < .001$) and 0.55 ($n = 138$, $p < .001$) for 3 and 4-year-olds, respectively. Interestingly, the correlation for 4-year-olds in Wave 2 (rho = 0.55, $n = 138$, $p < .001$), whose development had been tracked repeatedly, was higher than for 4-year-olds in Wave 1 (rho = 0.34, $n = 168$, $p < .001$).

Across child sex, PR in boys yielded a slightly higher correlation with the DA results compared to girls, and these correlations did not change substantially over time. In terms of urbanicity, urban children showed more correlations between DA and PR compared to their counterparts at 3 years of age, whereas the opposite was observed for rural children at 4 years of age. Moreover, the correlations between the two methods across SES quartiles revealed that PR of children in the third quartile exhibited stable correlations with DA over time. For different levels of maternal education, all groups of mothers reported similar results for DA of their children's development over time, except for mothers with senior secondary education and junior college degrees, whose reports did not align consistently with DA as children grew older.

Lastly, we compared the predictive power of PR and DA on children's development over time. Both DA (b = 0.40, SE = 0.05, $p < .001$, 95% CI [0.30, 0.49], β = 0.51, Adjusted $R^2 = 0.34$) and PR (b = 0.34, SE = 0.04, $p < .001$, 95% CI [0.27, 0.41], β = 0.49, Adjusted $R^2 = 0.28$) significantly predicted children's scores measured by the same assessment approach 1 year later (See Table 5). The predictive power of the DA results was stronger for children's development than with PR. Additionally, the correlations between the gain scores measured by DA and PR were significant (rho = 0.32, $n = 138$, $p < .001$), indicating that both methods captured some similarities in the growth of child development. Across different groups of children, PR showed more consistency with DA in measuring the growth of boys' development (rho = 0.37, $n = 72$, $p < .01$) compared to girls. In rural children, PR was more sensitive in measuring developmental change over time (rho = 0.30, $n = 62$, $p < .05$) compared to urban children. Across socioeconomic factors, PR for children in the third quartile (rho = 0.33, $n = 37$, $p < .05$) reflected children's development more effectively than that for other quartiles in terms of the correlations with DA. Lastly, when comparing parents' education levels, only PR from parents with junior secondary education and below was relatively more sensitive in reflecting children's developmental change over time (rho = 0.41, $n = 33$, $p < .05$).

4.3 Correlations among PR, DA, and TR

The correlation between TR and DA was consistently lower than the correlations between PR and DA across various factors such as age, sex, urbanicity, and SES quartiles. Specifically, for rural children and girls at 4 years of age, there was no significant correlation between TR and DA. The results also showed that the correlations within the two AR measures were smaller than those between DA and either AR measure across all groups of children. Notably, the correlation between TR and PR was only significant for children in the third SES quartile (rho = 0.23, $n = 77$, $p < .05$) and those whose parents had a bachelor's degree or higher education level (rho = 0.21, $n = 146$, $p < .05$). These correlations persisted when focusing on the subset of children at 4 years of age. The correlations are presented in Table 6.

4.4 Agreement Between PR and TR on the Different Items

First, we calculated Cohen's Kappa for the ECDI2030 items with all the "Don't know" responses from the AR measures coded as "0". Out of the 20 items, the only items to exhibit significant Cohen's Kappa values were item 1 (i.e., Can (name) walk on uneven surface), item 7 (i.e., Can (name) speak using sentences of five or more words), item 10 (i.e., Can (name) recognise at least five Chinese characters), item 11 (i.e., Can (name) write his/her name), item 13 (i.e., If you ask (name) to give you three objects, does he/she gives you the correct number), and item 17 (i.e., Does

Table 6 Correlations between DA and AR results across populations

	Wave 1		Wave 2				Longitudinal sample					Longitudinal sample	
		DA and PR		DA and PR	DA and TR	PR and TR	DA and PR			DA and TR	PR and TR	Gain scores measured by DA and PR	
	N		N				N	3-year-olds	4-year-olds			N	
All age groups	309	.46***	309	.39***	.24***	.18**							
3-year-olds	138	.50***											
4-year-olds	168	.34***	138	.55***	.21*	.21*	138	.50***	.55***	.21*	.21*	138	.32***
5-year-olds			168	.18*	.19*	.10							
Sex													
Girls	154	.50***	154	.36***	.26**	.20*	66	.45***	.44***	.10	.20	66	.23
Boys	155	.43***	155	.41***	.22**	.14	72	.54***	.60***	.28*	.22	72	.37**
Urbanicity													
Urban	173	.41***	173	.22**	.23**	.18*	76	.46***	.33**	.23*	.28*	76	.12
Rural	136	.15	136	.27**	.10	.04	62	.30*	.46***	.01	.03	62	.30*
SES Quartiles													
Bottom quartile	81	.18	81	.27*	.12	.09	33	.30	.44**	-.04	-.04	33	.26
2nd quartile	73	.26*	73	.18	.06	-.07	34	.36*	.33	.09	.02	34	.11
3rdquartile	77	.42***	77	.29*	.23*	.23*	37	.51**	.49**	.30	.40*	37	.33*
Top quartile	77	.41***	77	.28*	.09	.15	33	.35*	.27	.03	.25	33	.10
Maternal Education													
Below bachelor's degree	163	.25**	163	.29***		.11	75	.36**	.45***		.01	75	.24*
Junior Secondary and below	72	.37**	72	.24*		-.14	33	.51**	.45**		-.14	33	.41*
Senior Secondary and junior college	91	.17	91	.20		.17	42	.28	.29		.28	42	**.08**
Bachelor's degree and above	146	.39***	146	.30***		.21*	63	.45***	.40**		.30*	63	.17

All coefficients are Spearman's rho correlations coefficients. Correlations between DA and PR across urbanicity, SES quartiles, and maternal education were correlated between residuals after accounting for the influence of child age and sex. Correlations that involve teacher report across urbanicity, SES quartiles, and maternal education were calculated after controlling for child age, sex, and teacher effects. $^{*}p<.05$ $^{**}p<.01$ $^{***}p<.001$

(name) offer help to someone who seems to need help). This indicated that the ratings from teachers and parents on these items were not statistically independent. Items 1, 7, 13, and 17 all had Kappa values between 0 and 0.2, suggesting minor agreement between TR and PR. In contrast, items 10 and 11 had a Kappa value of 0.27, indicating fair agreement between TR and PR on children's early literacy competencies. For other items, the agreement between parents and teachers could be due to guessing or chance.

Second, we investigated whether the agreement between TR and PR differed when "Don't know" responses were retained instead of being recoded as "0". The results demonstrated that the corrected agreement remained stable for items 1, 7, and 13. However, the corrected agreement decreased for items 10, 11, and 17. Notably, there were relatively more "Don't know" responses from teachers for item 10 (14.61%), which assesses literacy, and item 11 (10.68%), which evaluates pre-writing skills. On the other hand, parents were more likely to respond to item 17 with "Don't know" (13.27% in Wave 1 and 8.09% in Wave 2, Online Resource 1). The "Don't know" responses lowered the agreement between teachers and parents in measuring children's Learning domain and the Psychosocial Well-being domain became non-significant. Therefore, the original agreement rate might have been masked by some actual "Don't know" responses from PR and TR, which were originally coded as "0". The agreement analysis is presented in Table 7.

5 Discussion

This study addressed some methodological concerns about using PR to monitor children's early development at the population level. As the official measure designed to monitor the progress towards SDG Target 4.2, findings from ECDI2030 have significant implications for policymaking on ECD by providing evidence of children's development across populations. However, questions have been raised about the psychometric robustness of PR scores and the comparability of PR scores across sociocultural contexts. Therefore, this study focused on investigating the reliability and validity of the ECDI2030 within the context of a middle-income country. Specifically, the correlations between the ECDI2030 and two corresponding measures, DA and TR, were examined.

5.1 The Reliability and Validity of Different Approaches for Measuring Children's Developmental Status in China

The reliability of the three assessment measures, as indicated by Cronbach's alpha, was only satisfactory, with DA demonstrating the highest reliability. This implies that AR might capture more measurement error than DA when using the observed scores (Nunnally & Bernstein, 1994). When the restriction of tau-equivalence was relaxed, omega coefficients of PR were close to those of DA in both waves. Specifically, the reliability of PR in Wave 1 increased when there was no assumption of tau-equivalence. TR appeared to have the highest reliability among the three measures in

Table 7 Agreement between PR and TR by ECDI2030 items

Domains	Items	Scores		Coded "Don't know" as "0"			Keeping the "Don't know" response		
		PR	TR	Expected agreement (%)	Corrected agreement (%)	Kappa	Expected agreement (%)	Corrected agreement (%)	Kappa
Health	1	0.93	0.93	87.18	89.13	0.15*	87.18	89.13	0.15*
	2	0.99	0.98	97.13	97.10	-0.01	97.13	97.10	-0.01
	3	0.96	0.96	91.68	91.30	-0.05	91.65	91.30	-0.04
	4	0.91	0.78	72.75	71.74	-0.04	72.53	71.74	-0.03
Learning	5	0.99	0.99	97.85	97.83	-0.01	97.85	97.83	-0.01
	6	1.00	0.99	98.55	98.55	0.00	98.55	98.55	0.00
	7	0.99	0.90	88.70	89.86	0.10*	88.70	89.86	0.10**
	8	1.00	0.99	98.55	98.55	0.00	98.55	98.55	0.00
	9	0.98	0.91	89.51	89.13	-0.04	89.39	89.13	-0.02
	10	0.82	0.57	54.62	66.67	0.27**	52.18	60.58	0.18**
	11	0.43	0.20	54.41	66.67	0.27**	50.19	63.77	0.27**
	12	0.97	0.99	96.42	96.38	-0.01	96.40	96.38	-0.01
	13	0.98	0.96	94.36	95.65	0.23**	94.33	95.65	0.23**
	14	0.95	0.94	89.72	90.58	0.08	89.62	89.86	0.02
	15	0.89	0.88	79.49	78.26	-0.06	79.00	78.26	-0.04
Psychosocial Well-being	16	0.95	0.86	82.56	82.61	0.00	82.26	82.61	0.02
	17	0.87	0.85	75.71	80.43	0.19*	74.39	76.81	0.09
	18	0.98	0.95	92.97	92.75	-0.03	92.90	92.75	-0.02
	19	0.80	0.91	75.14	77.54	0.10	21.06	24.64	0.05
	20	0.94	0.91	86.52	86.96	0.03	46.90	47.83	0.02

N = 138. $^{*}p$ <.05. $^{**}p$ <.01

Wave 2. These findings suggested that although DA could be more reliable than AR, Cronbach's alpha could have underestimated the reliability of AR.

It is important to note that the non-specific nature and brevity of population-based measures like the ECDI2030 can decrease the value of Cronbach's alpha. Moreover, the reliability of the measure is influenced by both its length and internal consistency (Tavakol & Dennick, 2011). The ECDI2030 is a single index with only 20 items that target various aspects of children's development, so its reliability may be compromised. Additionally, the relatively low alpha could indicate a lack of inter-relatedness between the items (Tavakol & Dennick, 2011), suggesting the need for further examination of the potential multidimensional construct of the measure. Based on the findings of Cronbach's alpha and McDonald's omega, it would be difficult to determine which measure produced the most reliable assessment results. Consequently, multiple sources of information on early childhood development should be obtained wherever possible.

Furthermore, when comparing the assessment scores generated from different approaches, both AR measures tended to generate higher scores for children's development. The PR scores also exhibited less variation as children matured, indicating a potential ceiling effect in the PR measures. This ceiling effect could be attributed to differences in the way items are asked and scored between PR and DA. Adopting appropriate wording and response formats for AR scales is particularly important for population-based measures, which usually have a limited number of items, and therefore, the presentation of items becomes critical. The simplified binary response format used in PR (scored as either “0” or “1”) can introduce challenges. Parents may mistakenly assume that their children possess the skills being asked about, leading to overestimation. Alternatively, they may inaccurately answer “Yes” even if the child has only demonstrated the target skills once or twice rather than consistently (Cappa et al., 2021). A simplified response format can reduce variations in assessment results and potentially introduce biases when evaluating children’s developmental status.

We also examined the association between the ECDI2030 outcomes and child age, sex, urbanicity, and SES. Different patterns emerged when analysing the three measures. Both PR and DA outcomes effectively differentiated children’s development across urbanicity and SES in both waves. This aligns with previous studies conducted in China, which showed developmental disparities among children based on urbanicity and SES backgrounds during early childhood (Rao et al., 2022; Zhao et al., 2020).

However, contrary to previous studies in China that suggested sex disparities in ECD (Li et al., 2020; Zhao et al., 2020), the DA and PR results in the current study showed significant developmental differences between girls and boys in the first wave, but not in the second wave. This suggests that boys may catch up with girls in terms of the assessment outcomes by age 4 and 5. Girls outperform boys in academic capabilities measured by the ECDI2030, but this advantage diminishes over time as boys eventually acquire the same skills (Etchell et al., 2018; Rinaldi et al., 2023). Notably, the sex differences measured by PR at Wave 2 were marginally insignificant ($p=.056$). Further studies are needed to investigate sex disparities in ECD across children ages at the population level using psychometrically robust measures and a larger sample in China (Weber et al., 2017).

In contrast, TR failed to show differences in children’s assessment outcomes across urbanicity and SES. This may be due to several reasons. First, teachers typically assess and judge a child’s development in relation to their peers within the classroom (Martínez et al., 2009; Meissel et al., 2017); hence, the range of abilities of children within a specific classroom matter. A child who is considered advanced in a rural classroom may be seen to be falling behind compared to their peers in an urban classroom. Similarly, the composition of students’ SES background within a classroom can influence a teacher’s judgment (Ready & Wright, 2011). Teachers tend to perceive children in higher-SES or higher-achieving classrooms as more advanced in terms of academic skills, regardless of individual background (Ready & Wright, 2011). These factors make it challenging for TR to differentiate children’s development across urbanicity and SES. Furthermore, variations in teachers’ familiarity with the sampled children, influenced by factors such as class size and timing of the report (Fernald et al., 2017; Russo et al., 2019), can affect the reliability and validity of TR. Therefore, unlike PR or DA, the accuracy of TR can be affected by factors at the classroom or school level.

5.2 Correlations between PR and DA across Populations and over time

The average correlation between PR and DA in the current sample was moderate. This suggests that DA and PR should not be used interchangeably when measuring the same item. When implementing assessments at scale or in situations where collecting DA data is challenging, it is important to validate PR measures against other criterion measures. It is also valuable to explore factors that may influence the accuracy of PR assessments. For example, investigating whether children from higher SES backgrounds or those with mothers who have higher education levels can be assessed with less bias using PR would be beneficial.

Our findings indicated that PR may be less sensitive than DA when measuring the development of older children in our sample. The PR scores increased at a slower rate compared to DA scores as children grew older, which resulted in a decreasing correlation between PR and DA. Nevertheless, the correlation between PR and DA remained stable across child sex, suggesting that PR may not exhibit sex-based bias in assessing the holistic development of young children. An interesting observation in children assessed in both waves is that the correlation between their PR and DA results did not decrease as they matured. This could be due to parents becoming more attentive to the specific abilities being assessed and providing children with more opportunities for practising those skills or engaging with them in related learning activities. These experiences might also lead parents to develop a more accurate understanding of their children's developmental progress, enabling them to make more objective evaluations (Zippert & Ramani, 2017). It would be worthwhile to explore the potential effect of "repeated report" on the accuracy of PR in future studies using a larger and more representative sample to examine if the accuracy of PR is related to parents' familiarity with the assessment items.

After controlling for child age, there was a slightly different pattern in the correlations between PR and DA across urbanicity. We would expect that PR would be more accurate for both age groups when evaluating rural children than urban children compared to DA. Contrary to expectations, the correlations between PR and DA were only significant for 4-year-old rural children. This suggests that rural parents of 3-year-olds may face challenges in accurately reporting their children's development. Considering that urbanicity was highly correlated with children's SES in our study sample, rural parents' lower access to resources and knowledge about their children's development compared to their urban counterparts might be an important factor hindering the reliable judgment of their children's development (Bornstein et al., 2010).

The moderating effect of SES on the correlations between the PR and DA results was also examined. Although higher SES children tended to achieve higher scores, potentially leading to smaller correlations due to a ceiling effect, parents with higher SES still made relatively accurate judgments about their children's development. Specifically, parents from the third SES quartile consistently provided reports over time that aligned with the DA measurements, indicating reliable assessment of their children's development.

We particularly looked at the correlations between PR and DA across maternal education, a key component of children's SES. Consistent with the findings from earlier work, we found that maternal education level did not strongly correlate with the accuracy of PR (Bedore et al., 2011; Guiberson et al., 2011; Thal et al., 2000). The results showed that parents with or without a bachelor's degree did not show a substantial difference in the correlations between PR and DA when evaluating children's development. A more nuanced analysis focusing on parents without bachelor's degrees revealed that mothers with lower education levels (i.e., junior secondary and below) evaluated their children's development more in line with the DA results than those with higher education levels (i.e., senior secondary and junior college). Surprisingly, PR from mothers in the latter group did not significantly correlate with DA.

It is unclear how family SES and parent's education level relate to the accuracy of PR. The parent's SES and accuracy of PR could be associated with the home learning environment that parents create for their children (Zippert & Ramani, 2017), which varies based on their income and education level (Kalil & Ryan, 2020; Rao et al., 2021). Nevertheless, factors such as maternal employment, caregiving arrangements, and whether the child was an only child can also influence the amount and quality of the time and interactions between mothers and their children, influencing their understanding of their child's development. For example, our findings showed that parents with education levels of junior secondary and below and those with senior secondary and junior college were mainly from rural areas, with the former more likely to work close to home and the latter tending to look for jobs in suburban or urban areas (e.g., Beijing in the current study) with higher salaries. This could result in less time spent at home and limited exposure to their children's daily performance and behaviours for the latter. Additionally, they may lack sufficient support and resources to provide high-quality learning activities and interactions for their children compared to parents with bachelor's degrees and above. These factors could contribute to the unexpected lower correlations between DA and PR from parents with senior secondary education and junior college degrees.

The above findings emphasise a nuanced interpretation of the PR results from parents with varying education levels in different contexts. A previous study investigating parents' perceptions of the ECDI2030 items found that parents with primary school or lower education levels had difficulty understanding some items (Cappa et al., 2021). However, this may not be contextually relevant in countries like China, where the majority of mothers have completed lower secondary education. Hence, the ability to interpret assessment items might not vary that much across caregivers in China, while factors such as occupation, working arrangements, and immigration status may become important.

Lastly, when comparing the sensitivity of PR and DA for measuring change in their children's development over time, we found that DA was more effective in predicting children's development 1 year later. On the other hand, PR showed only a moderate correlation with DA for measuring the growth of child development between 3 and 4 years of age, and the correlations between these gains varied across populations.

PR was better at capturing developmental changes in specific subgroups, including boys, rural children, children from the third-SES quartile, and children with mothers with junior secondary degrees and below. This finding could be because the correlations between PR and DA for these groups of children were higher than for their counterparts, and the changes in their assessment scores were more significant over time. Studies have suggested that an indirect approach such as PR may not be as sensitive as DA when evaluating the effectiveness of intervention programmes or policies on children's development and learning (Fernald & Pitchik, 2019). The preliminary evidence from this study suggests that caution should be exercised when relying solely on the ECDI2030 to measure children's developmental change at the population level in an upper-middle-income country, wherein the majority of the population has completed junior secondary education. It is necessary to conduct further studies in diverse contexts with larger sample sizes to explore the predictive validity of the ECDI2030 in middle-income contexts.

5.3 Comparing Different Reporters of the ECDI2030 when Measuring Children's Developmental Status

Overall, PR was found to be more sensitive in capturing children's development across ages compared to TR, as evidenced by a higher correlation between PR and DA than between TR and DA. Consistent with the results obtained in the validation of TR, the correlations between TR and DA were not significant across all SES quartiles, which again indicated there were limitations in using TR to detect SES-based developmental gaps among young children in China.

The correlations between TR and PR were also explored. Given that PR and TR were found to exhibit less variations than DA in measuring children's developmental status, the correlations between the two AR measures were similar or smaller compared to the correlations involving DA. However, it is worth noting that significant correlations were observed between TR and PR for children from urban areas, in the third SES quartile or with mothers holding bachelor's degrees and above. These results indicated that teachers and mothers of these children provided consistent evaluations, and teachers demonstrated relatively higher accuracy when making judgments about the children's development compared to other groups. One possible explanation for these findings is that teachers are more likely to make accurate judgments for children whose development is close to the classroom mean (Vitiello & Williford, 2021). Children from medium-to-high SES quartiles would typically not be at the extremes of the developmental spectrum of a classroom. Besides, parents with higher education levels or from higher SES quartiles may be more involved in their children's school activities and have more opportunities to observe their children's performance and behaviours in different environments. This could contribute to a better understanding of their children's developmental status and align more closely with teachers' assessments than parents from other backgrounds. Further investigation into the correlation between teachers and parents across children's SES would be valuable in shedding light on the factors influencing the variations in the correlation of their judgments.

5.3.1 Agreement between TR and PR across Items

At the item level, there was greater agreement between parents' and teachers' reports of children's learning across all three domains. Among the six items with significant agreement on parents' and teachers' judgment, four were from the Learning domain

(36%), one was from the Health domain (25%), and one was from the Psychosocial Well-being domain (20%).

The higher agreement between parents' and teachers' evaluation of children's learning could be due to the accessibility and observability of these skills (Bodnarchuk & Eaton, 2004; Mashburn & Henry, 2004). As children grow older and start attending preschool, their pre-academic learning competencies, such as expressive language, recognition of Chinese characters, writing, and counting, become more discernible to adults. Parents and teachers would have increased opportunities to observe children's demonstrations of these skills, allowing for more aligned judgments. The observable nature of these learning skills based on concrete indicators makes it easier for both parents and teachers to evaluate. While children's social-emotional competencies, such as social cognition and empathy, also developed through interactions with their environments and the acquisition of linguistic and communication skills over time, it can be challenging to conceptualise and measure these skills precisely (Jones et al., 2016). Defining what it means to "be helpful" or "get along well with other children" can vary across different scenarios, and using a simple "Yes" or "No" response to evaluate children's social-emotional skills may introduce subjectivity.

We also found that parents and teachers had different tendencies to respond with "Don't know" to certain items, which influenced the agreement between TR and PR. Specifically, when considering parents' "Don't know" response to item 17 (i.e., Does (name) offer help to someone who seems to need help), the agreement between PR and TR regarding children's Psychosocial Well-being domain was no longer significant after correcting for chance. This finding aligns with previous research indicating discrepancies between informants when evaluating children's social-emotional competencies or problem behaviours (Achenbach et al., 1987; Dinnebeil et al., 2013). As parents and teachers may provide inconsistent judgment when evaluating the psychosocial well-being domain, relying solely on either TR or PR may not obtain a holistic view of children's social-emotional development. This also highlights the importance of exploring the domain structure of a population-based assessment for selecting the appropriate AR assessment approach for children's holistic development.

Finally, it is worth noting that teachers were more likely to respond with "Don't know" to items related to children's literacy skills, such as recognising Chinese characters and writing. This finding may seem unexpected, as these pre-academic skills are expected to be easily observed by teachers in a school or classroom setting. We posit that this is due to the ECE policy forbidding formal teaching of literacy skills (e.g., Chinese character recognition) in Chinese kindergartens and the ECCE directives that decrease "schoolification" (Chen et al., 2022; Ministry of Education 2018). As a result, teachers may avoid making definitive choices on these items. Given the differences in parents' and teachers' "Don't know" responses, further research is needed to delve into the motivations behind these responses in both groups. This emphasises the need for a nuanced scoring approach for the ECDI2030 items when using AR measures, which can help unpack the unique information from different reporters when interpreting the assessment outcomes of children's developmental status.

6 Conclusions and Limitations

To the best of our knowledge, this study is the first attempt to examine the psychometric properties of the ECDI2030 by comparing it with two common early childhood assessment methods (i.e., DA and TR) using corresponding items. This study revealed potential issues to consider when using PR data on children's developmental status to inform progress towards SDG Target 4.2 in specific contexts and had several implications for other societies. First, the ECDI2030 has demonstrated reasonable reliability and validity, with scores showing positive associations with child age, sex, urbanicity, and SES. However, compared to DA, the PR approach exhibited lower sensitivity and discrimination across child age. Future studies measuring SDG Target Indicator 4.2.1 should prioritise cross-method validation of child development data within and across cultures. Second, since there could be systematic differences in the reliability or accuracy of PR across populations, countries need to consider the appropriateness and relevance of adopting a PR measure when collecting child development data and explore the potential factors that could moderate the reliability and validity of PR measures in their own contexts. Third, this study offers insights for future studies and practices when selecting assessment approaches to inform the population-based development of young children. It highlights the importance of a comprehensive understanding of child assessment outcomes across information sources and cautions against over-reliance on a single measure to inform early child development at the population level.

Despite the valuable insights gained from this study, several limitations should be acknowledged. First, the study sample consisted of children attending preschools in two neighbouring areas in China, which may limit the generalisability of the results to other groups, particularly socially disadvantaged children such as left-behind children or migrant children. Therefore, it is crucial to investigate the psychometric robustness of current measures with a more representative sample. Second, although DA was considered to be the most reliable and accurate among the three assessment approaches, it is important to acknowledge that it also has limitations. Factors such as children's short attention span, non-compliance, and unfamiliarity with the test settings can influence DA results (Nordahl-Hansen et al., 2014). These factors should be considered when interpreting the findings related to DA. Third, this study primarily focused on internal consistency and criterion-related validity when comparing the psychometric properties of the three measures. Further studies can employ other indicators to examine their reliability and validity. For example, studies can investigate their associations with established criterion measures, not necessarily with the same items. This examination could add to the current findings in terms of the accuracy of the different approaches in measuring the ECDI2030 items. Fourth, more waves of longitudinal data are necessary to investigate the stability of AR in tracking children's developmental status over time. Previous studies suggested that different assessment measures can uniquely predict various learning and developmental competencies of young children (Ahmed et al., 2022; Obradović et al., 2022). Comparing the predictive validity of PR and TR may enable a comprehensive understanding of the strengths and limitations of different reporters when capturing changes in children's development. Fifth, this study operationalised the similarities between different methods based on their correlations. Although correlations help describe the relationships between different assessment outcomes, they may not fully capture the extent of overestimations or underestimations of specific measures. Further research is needed to investigate the concordance among the approaches assessing children's development by adopting different analytical methods. Finally, the current study focused on the impact of structural factors, such as family SES and education levels, when comparing AR against DA. It would be beneficial to gather more information about the quality of home learning environments and school activities that parents and teachers engage in with children. Also, teacher characteristics, such as their education levels and work experience, were not examined in current analyses but may potentially moderate the correlations between TR and the other methods. Incorporating other possible moderators in future analyses could help us better understand how to utilise and interpret AR outcomes more appropriately.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12187-024-10184-9.

Acknowledgements We extend our heartfelt gratitude to Xiaohui Xu and her team at Capital Normal University for their invaluable assistance in data collection, which significantly contributed to the study's success. We are also deeply appreciative of the feedback provided by members of the labs of Professor Aisha Yousafzai and Professor Dana McCoy during the first author's appointment as a Visiting Student Fellow at Harvard Graduate School of Education. Furthermore, our sincere thanks go to the ECDE lab at the Faculty of Education, The University of Hong Kong, for their resources and expertise throughout the research process.

Author Contributions NR and ZL conceptualised and designed the study. ZL performed the formal analysis and wrote the first draft with the support of NR. Both authors critically reviewed this draft and approved the final draft for submission, with final responsibility for publication.

Funding The work was supported by a grant to Nirmala Rao from the Research Grants Council, Hong Kong SAR China. Project No. HKU-17602519.

Declarations

Competing Interests We have no conflicts of interest to declare.

References

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*(2), 213.

Ahmed, I., Steyer, L., Suntheimer, N. M., Wolf, S., & Obradović, J. (2022). Directly assessed and adult-reported executive functions: Associations with academic skills in Ghana. *Journal of Applied Developmental Psychology, 81*, 101437. https://doi.org/10.1016/j.appdev.2022.101437

Alderman, H., Friedman, J., Ganga, P., Kak, M., & Rubio-Codina, M. (2021). Assessing the performance of the Caregiver Reported Early Development Instruments (CREDI) in rural India. *Annals of the New York Academy of Sciences, 1492*(1), 58–72.

Bedore, L. M., Peña, E. D., Joyner, D., & Macken, C. (2011). Parent and teacher rating of bilingual language proficiency and language development concerns. *International Journal of Bilingual Education and Bilingualism, 14*(5), 489–511.

Beijing Municipal Bureau of Statistics. (2022). *The per capita disposable* income *of residents in Beijing increased by 3.2%.* Retrieved from https://www.beijing.gov.cn/gongkai/shuju/sjjd/202301/t20230119_2905639.html

Benasich, A. A., & Brooks-Gunn, J. (1996). Maternal attitudes and knowledge of child-rearing: Associations with family and child outcomes. *Child Development, 67*(3), 1186–1205.

Bennetts, S. K., Mensah, F. K., Westrupp, E. M., Hackworth, N. J., & Reilly, S. (2016). The Agreement between parent-reported and directly measured child language and parenting behaviors. *Frontiers in Psychology*, *7*(1710). https://doi.org/10.3389/fpsyg.2016.01710

Berg-Nielsen, T. S., Solheim, E., Belsky, J., & Wichstrom, L. (2012). Preschoolers' psychosocial problems: In the eyes of the beholder? Adding teacher characteristics as determinants of discrepant parent–teacher reports. *Child Psychiatry & Human Development, 43*(3), 393–413. https://doi.org/10.1007/s10578-011-0271-0

Bergold, S., Christiansen, H., & Steinmayr, R. (2019). Interrater agreement and discrepancy when assessing problem behaviors, social-emotional skills, and developmental status of kindergarten children. *Journal of Clinical Psychology, 75*(12), 2210–2232.

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development, 78*(2), 647–663. https://doi.org/10.1111/j.1467-8624.2007.01019.x

Bodnarchuk, J. L., & Eaton, W. O. (2004). Can parent reports be trusted?: Validity of daily checklists of gross motor milestone attainment. *Journal of Applied Developmental Psychology, 25*(4), 481–490. https://doi.org/10.1016/j.appdev.2004.06.005

Bornstein, M. H., Cote, L. R., Haynes, O. M., Hahn, C.-S., & Park, Y. (2010). Parenting knowledge: Experiential and sociodemographic factors in European American mothers of young children. *Developmental Psychology, 46*(6), 1677.

Bornstein, M. H., Putnick, D. L., Costlow, K. M., & Suwalsky, J. T. D. (2020). Retrospective report revisited: Long-term recall in European American mothers moderated by developmental domain, child age, person, and metric of agreement. *Applied Developmental Science, 24*(3), 242–262. https://doi.org/10.1080/10888691.2018.1462090

Bronfenbrenner, U. (1979). Contexts of child rearing: Problems and prospects. *American psychologist, 34*(10), 844.

Callan Stoiber, K. (1992). Parents' beliefs about their children's cognitive, social, and motor functioning. *Early Education & Development, 3*(3), 244–259. https://doi.org/10.1207/s15566935eed0303_4

Cappa, C., Petrowski, N., De Castro, E. F., Geisen, E., LeBaron, P., Allen-Leigh, B., Place, J. M., & Scanlon, P. J. (2021). Identifying and minimizing errors in the measurement of early childhood development: Lessons learned from the cognitive testing of the ECDI2030. *International Journal of Environmental Research and Public Health, 18*(22), 12181. https://www.mdpi.com/1660-4601/18/22/12181

Chen, J. J. L. (2010). Gender differences in externalising problems among preschool children: Implications for early childhood educators. *Early Child Development and Care, 180*(4), 463–474. https://doi.org/10.1080/03004430802041011

Chen, S., Chen, C., & Wen, P. (2022). Parental anxiety, endorsement of literacy learning, and home literacy practices among Chinese parents of young children. *Reading and Writing, 35*(4), 825–852. https://doi.org/10.1007/s11145-021-10220-y

Chen, Y., & Feng, S. (2013). Access to public schools and the education of migrant children in China. *China Economic Review, 26*, 75–88.

Cintas, H. L. (1995). Cross-cultural similarities and differences in development and the impact of parental expectations on motor behavior. *Pediatric Physical Therapy*, *7*(3). https://journals.lww.com/pedpt/fulltext/1995/00730/cross_cultural_similarities_and_differences_in.4.aspx

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology, 78*(1), 98.

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological bulletin, 131*(4), 483.

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin, 141*(4), 858–900. https://doi.org/10.1037/a0038498

de Onis, M., & Blössner, M. (2003). The World Health Organization global database on child growth and malnutrition: Methodology and applications. *International Journal of Epidemiology, 32*(4), 518–526.

Dinnebeil, L. A., Sawyer, B. E., Logan, J., Dynia, J. M., Cancio, E., & Justice, L. M. (2013). Influences on the congruence between parents' and teachers' ratings of young children's social skills and problem behaviors. *Early Childhood Research Quarterly, 28*(1), 144–152.

Etchell, A., Adhikari, A., Weinberg, L. S., Choo, A. L., Garnett, E. O., Chow, H. M., & Chang, S.-E. (2018). A systematic literature review of sex differences in childhood language and brain development. *Neuropsychologia, 114*, 19–31.

Feldman, H. M., Dale, P. S., Campbell, T. F., Colborn, D. K., Kurs-Lasky, M., Rockette, H. E., & Paradise, J. L. (2005). Concurrent and predictive validity of parent reports of child language at ages 2 and 3 years. *Child Development, 76*(4), 856–868. https://doi.org/10.1111/j.1467-8624.2005.00882.x

Fernald, L. C., & Pitchik, H. O. (2019). The necessity of using direct measures of child development. *The Lancet Global Health, 7*(10), e1300–e1301.

Fernald, L. C. H., Prado, E. L, Kariger,P. K., & Raikes, A. (2017). *A toolkit for measuring early* childhood *development in low and middle income countries.* World Bank Group. Washington, D.C. Retrieved from http://documents.worldbank.org/curated/en/384681513101293811/A-toolkit-for-measuring-early-childhood-development-in-low-and-middle-income-countries

Fluck, M., Linnell, M., & Holgate, M. (2005). Does counting count for 3-to 4-year-olds? Parental assumptions about preschool children's understanding of counting and cardinality. *Social Development, 14*(3), 496–513.

Furnari, E. C., Whittaker, J., Kinzie, M., & DeCoster, J. (2017). Factors associated with accuracy in prekindergarten teacher ratings of students' mathematics skills. *Journal of Psychoeducational Assessment, 35*(4), 410–423.

Gong, J., & Rao, N. (2023). Early learning opportunities of preschool children affected by migration in China. *Early Childhood Research Quarterly, 63*, 228–239.

Guiberson, M., Rodriguez, B. L., & Dale, P. S. (2011). Classification accuracy of Brief Parent Report Measures of Language Development in Spanish-speaking toddlers. *Language, Speech & Hearing Services in Schools, 42*(4), 536–549. https://doi.org/10.1044/0161-1461(2011/10-0076)

Halpin, P. F., de Castro, E. F., Petrowski, N., & Cappa, C. (2024). Monitoring early childhood development at the population level: The ECDI2030. *Early Childhood Research Quarterly, 67*, 1–12.

Hart, S. A., Ganley, C. M., & Purpura, D. J. (2016). Understanding the home math environment and its role in predicting parent report of children's math skills. *PLOS ONE, 11*(12), e0168227. https://doi.org/10.1371/journal.pone.0168227

Hayes, A. F., & Coutts, J. J. (2020). Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But…. Communication *Methods and Measures*, *14*(1), 1–24. https://doi-org.eproxy.lib.hku.hk/10.1080/19312458.2020.1718629

Hebei Provincial Bureau of Statistics. (2022). *The per capita disposable income of residents in Hebei* reached *30,867 yuan.* Retrieved from http://he.people.com.cn/n2/2023/0120/c192235-40275224.html

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of educational research, 59*(3), 297–313.

Hong, X., Liu, P., Ma, Q., & Luo, X. (2015). The way to early childhood education equity–policies to tackle the urban-rural disparities in China. *International Journal of Child Care and Education Policy, 9*(1).

Hu, B. Y., Roberts, S. K., Leng Ieong, S. S., & Guo, H. (2016). Challenges to early childhood education in rural China: lessons from the Hebei province. *Early Child Development and Care, 186*(5), 815–831.

Jacobs, J. E. (1991). Influence of gender stereotypes on parent and child mathematics attitudes. *Journal of Educational Psychology, 83*(4), 518–527. https://doi.org/10.1037/0022-0663.83.4.518

Jin, J. (2008). Early childhood education administration in the new period: Challenges and opportunities. *Chinese Education & Society, 41*(2), 77–90.

Johnson, S., Marlow, N., Wolke, D., Davidson, L., Marston, L., O'Hare, A., Peacock, J., & Schulte, J. (2004). Validation of a parent report measure of cognitive development in very preterm infants. *Developmental Medicine & Child Neurology, 46*(6), 389–397. https://doi.org/10.1017/s0012162204000635

Jones, S. M., Zaslow, M., Darling-Churchill, K. E., & Halle, T. G. (2016). Assessing early childhood social and emotional development: Key conceptual and measurement issues. *Journal of Applied Developmental Psychology, 45*, 42–48.

Kalil, A., & Ryan, R. (2020). Parenting practices and socioeconomic gaps in childhood outcomes. *The Future of Children, 30*(2020), 29–54.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Li, K., Pan, Y., Hu, B., Burchinal, M., De Marco, A., Fan, X., & Qin, J. (2016). Early childhood education quality and child outcomes in China: Evidence from Zhejiang Province. *Early Childhood Research Quarterly, 36*, 427–438.

Li, L., Fan, J., & Jin, Z. (2019). Comparing multimethod assessment of approaches to learning among preschool children: Direct measure, teacher report, and parent report. *Psychology in the Schools, 56*(8), 1271–1286.

Li, Y., Tang, L., Bai, Y., Zhao, S., & Shi, Y. (2020). Reliability and validity of the caregiver reported early development instruments (CREDI) in impoverished regions of China. *BMC Pediatrics, 20*(1), 475–516. https://doi.org/10.1186/s12887-020-02367-4

Li, Z., & Rao, N. (2023). Advancing equity in early childhood development in China: Progress in meeting Sustainable Development Goal Target 4.2. *International Journal of Chinese Education, 12*(2), 2212585X231175499. https://doi.org/10.1177/2212585x231175499

Lin, J., Napoli, A. R., Schmitt, S. A., & Purpura, D. J. (2021). The relation between parent ratings and direct assessments of preschoolers' numeracy skills. *Learning and Instruction, 71*, 101375. https://doi.org/10.1016/j.learninstruc.2020.101375

Lu, C., Cuartas, J., Fink, G., McCoy, D., Liu, K., Li, Z., Daelmans, B., & Richter, L. (2020). Inequalities in early childhood care and development in low/middle-income countries: 2010–2018. *BMJ global health, 5*(2), e002314.

Martínez, J. F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment, 14*(2), 78–102.

Mashburn, A. J., & Henry, G. T. (2004). Assessing school readiness: Validity and bias in preschool and kindergarten teachers' ratings. *Educational Measurement: Issues and Practice, 23*(4), 16–30.

Massa, J., Gomes, H., Tartter, V., Wolfson, V., & Halperin, J. M. (2008). Concordance rates between parent and teacher clinical evaluation of language fundamentals observational rating scale. *International Journal of Language & Communication Disorders, 43*(1), 99–110. https://doi.org/10.1080/13682820701261827

Mccabe, L. A., Hernandez, M., Lara, S. L., & Brooks-Gunn, J. (2000). Assessing preschoolers' self-regulation in homes and classrooms: Lessons from the field. *Behavioral Disorders, 26*(1), 53–69. https://doi.org/10.1177/019874290002600106

McCoy, D. C., Peet, E. D., Ezzati, M., Danaei, G., Black, M. M., Sudfeld, C. R., Fawzi, W., & Fink, G. (2016). Early childhood developmental status in low- and middle-income countries: National, regional, and global prevalence estimates using predictive modeling. *PLOS Medicine, 13*(6), e1002034. https://doi.org/10.1371/journal.pmed.1002034

McCoy, D. C., Waldman, M., Team, C. F., & Fink, G. (2018). Measuring early childhood development at a global scale: Evidence from the Caregiver-Reported early development instruments. *Early Childhood Research Quarterly, 45*, 58–68.

McDonald, R. P. (1999). Test *theory: A unified treatment*. Lawrence Erlbaum.

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. *Teaching and teacher education, 65*, 48–60.

Ministry of Education of the People's Republic of China. (2018). A notice of operating a special action to governance the "elementary-schoolization" in kindergartens and preschools. Retrieved from http://www.moe.gov.cn/srcsite/A06/s3327/201807/t20180713_342997.html

Ministry of Education of the People's Republic of China. (2023). *The Main Results of 2022 National Education Statistical Bulletin.* Retrieved from http://www.moe.gov.cn/jyb_sjzl/sjzl_fztjgb/202307/t20230705_1067278.html

Muntoni, F., & Retelsdorf, J. (2019). At their children's expense: How parents' gender stereotypes affect their children's reading outcomes. *Learning and Instruction, 60*, 95–103. https://doi.org/10.1016/j.learninstruc.2018.12.002

Nordahl-Hansen, A., Kaale, A., & Ulvund, S. E. (2014). Language assessment in children with autism spectrum disorder: Concurrent validity between report-based assessments and direct tests. *Research in Autism Spectrum Disorders, 8*(9), 1100–1106. https://doi.org/10.1016/j.rasd.2014.05.017

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory.* New York. NY: McGraw-Hill

Obradović, J., Finch, J. E., Connolly, C., Siyal, S., & Yousafzai, A. K. (2022). The unique relevance of executive functions and self-regulation behaviors for understanding early childhood experiences and preschoolers' outcomes in rural Pakistan. *Developmental Science, 25*(6), e13271. https://doi.org/10.1111/desc.13271

Obradović, J., Sulik, M. J., Finch, J. E., & Tirado-Strayer, N. (2018). Assessing students' executive functions in the classroom: Validating a scalable group-based procedure. *Journal of Applied Developmental Psychology, 55*, 4–13. https://doi.org/10.1016/j.appdev.2017.03.003

Phillipson, S., & Phillipson, S. N. (2007). Academic Expectations, Belief of ability, and involvement by parents as predictors of child achievement: A cross-cultural comparison. *Educational Psychology, 27*(3), 329–348. https://doi.org/10.1080/01443410601104130

Pushparatnam, A., Luna Bazaldua, D. A., Holla, A., Azevedo, J. P., Clarke, M., & Devercelli, A. (2021). Measuring early childhood development among 4–6 year olds: The identification of psychometrically robust items across diverse contexts. *Frontiers in public health, 9*, 17.

Raikes, A. (2017). Measuring child development and learning. *European Journal of Education, 52*(4), 511–522.

Rao, N., Chan, S. W. Y., Su, Y., Richards, B., Cappa, C., De Castro, E. F., & Petrowski, N. (2022). Measuring being "Developmentally on Track": Comparing direct assessment and caregiver report of early childhood development in Bangladesh, China, India and Myanmar. *Early Education and Development, 33*(6), 1013–1035. https://doi.org/10.1080/10409289.2021.1928446

Rao, N., Cohrssen, C., Sun, J., Su, Y., & Perlman, M. (2021). Early child development in low-and middle-income countries: Is it what mothers have or what they do that makes a difference to child outcomes? *Advances in Child Development and Behavior, 61*, 255–277.

Rao, N., Su, Y., & Gong, J. (2022). Persistent urban–rural disparities in early childhood development in China: The roles of maternal education, home learning environments, and early childhood education. *International Journal of Early Childhood, 54*(3), 445–472. https://doi.org/10.1007/s13158-022-00326-x

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48*(2), 335–360.

Richards, B., Rao, N., & Chan, S. W. Y. (2023). Measuring indicators of sustainable development goal target 4.2.1: factor structure of a direct assessment tool in four Asian countries. *Oxford Review of Education, 49*(1), 69–92. https://doi.org/10.1080/03054985.2022.2093844

Rinaldi, P., Pasqualetti, P., Volterra, V., & Caselli, M. C. (2023). Gender differences in early stages of language development: Some evidence and possible explanations. *Journal of Neuroscience Research, 101*(5), 643–653.

Rubio-Codina, M., & Grantham-Mcgregor, S. (2020). Predictive validity in middle childhood of short tests of early childhood development used in large scale studies compared to the Bayley-III, the Family Care Indicators, height-for-age, and stunting: A longitudinal study in Bogota. *Colombia. PLOS ONE, 15*(4), e0231317. https://doi.org/10.1371/journal.pone.0231317

Russo, J. M., Williford, A. P., Markowitz, A. J., Vitiello, V. E., & Bassok, D. (2019). Examining the validity of a widely-used school readiness assessment: Implications for teachers and early childhood programs. *Early Childhood Research Quarterly, 48*, 14–25.

Sabanathan, S., Wills, B., & Gladstone, M. (2015). Child development assessment tools in low-income and middle-income countries: How can we use them more appropriately? *Archives of Disease in Childhood, 100*(5), 482–488. https://doi.org/10.1136/archdischild-2014-308114

Sedgwick, P. (2014). Spearman's rank correlation coefficient. *BMJ, 349*(nov28 1), g7327. https://doi.org/10.1136/bmj.g7327

Snow, C. E., & Van Hemel, S. B. (2008). *Early childhood assessment: Why, what, and how*. The National Academies Press.

StataCorp. (2021). *Stata Statistical Software: Release 18*. StataCorp LLC.

Stoiber, K. C., & Houghton, T. G. (1993). The relationship of adolescent mothers' expectations, knowledge, and beliefs to their young children's coping behavior. *Infant Mental Health Journal, 14*(1), 61–79.

Stone, L. L., Otten, R., Engels, R. C. M. E., Vermulst, A. A., & Janssens, J. M. A. M. (2010). Psychometric properties of the parent and teacher versions of the Strengths and Difficulties Questionnaire for 4- to 12-Year-Olds: A review. *Clinical Child and Family Psychology Review, 13*(3), 254–274. https://doi.org/10.1007/s10567-010-0071-2

Su, Y., Rao, N., Sun, J., & Zhang, L. (2021). Preschool quality and child development in China. *Early Childhood Research Quarterly, 56*, 15–26.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education, 2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Thal, D., Jackson-Maldonado, D., & Acosta, D. (2000). Validity of a parent-report measure of vocabulary and grammar for Spanish-speaking toddlers. *Journal of Speech, Language, and Hearing Research, 43*(5), 1087–1100.

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best Alternatives to Cronbach's Alpha Reliability in Realistic Conditions: Congeneric and Asymmetrical Measurements. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00769

UNESCO. (2022). *Education starts early: Progress, challenges, and opportunities*. Retrieved from. https://unesdoc.unesco.org/ark:/48223/pf0000383668

UNICEF. (2023). The *Early Childhood Development Index 2030: A new measure of early childhood development. New York.* Retrieved from https://data.unicef.org/resources/early-childhood-development-index-2030-ecdi2030/

UNICEF & UNESCO. (2024). *Global Report on Early Childhood Care and Education: The right to a strong foundation*. Retrieved from https://www.unicef.org/reports/global-report-early-childhood-care-and-education-right-strong-foundation

United Nations. (2015). *Sustainable Development Goals*. Department of Economic and Social Affair. Retrieved from https://sdgs.un.org/goals

van Ginkel, J. R., Linting, M., Rippe, R. C. A., & van der Voort, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment, 102*(3), 297–308. https://doi.org/10.1080/00223891.2018.1530680

Vitiello, V. E., & Williford, A. P. (2021). Alignment of teacher ratings and child direct assessments in preschool: A closer look at teaching strategies GOLD. *Early Childhood Research Quarterly, 56*, 114–123. https://doi.org/10.1016/j.ecresq.2021.03.004

von Suchodoletz, A., Uka, F., & Larsen, R. A. A. A. (2015). Self-regulation across different contexts: Findings in young Albanian children. *Early Education and Development, 26*(5–6), 829–846. https://doi.org/10.1080/10409289.2015.1012189

Waldman, M., McCoy, D. C., Seiden, J., Cuartas, J., Team, C. F., & Fink, G. (2021). Validation of motor, cognitive, language, and socio-emotional subscales using the caregiver reported early development instruments: An application of multidimensional item factor analysis. *International Journal of Behavioral Development, 45*(4), 368–377.

Walker, S. (2004). Teacher reports of social behaviour and peer acceptance in early childhood: Sex and social status differences. *Child Study Journal, 34*(1), 13–28.

Wang, B., Luo, X., Yue, A., Tang, L., & Shi, Y. (2022). Family environment in rural China and the link with early childhood development. *Early Child Development and Care, 192*(4), 617–630.

Wang, L., Dang, R., Bai, Y., Zhang, S., Liu, B., Zheng, L., Yang, N., & Song, C. (2020). Teacher qualifications and development outcomes of preschool children in rural China. *Early Childhood Research Quarterly, 53*, 355–369.

Warrens, M. J. (2015). Five ways to look at Cohen's Kappa. *Journal of Psychology & Psychotherapy, 05*(04). https://doi.org/10.4172/2161-0487.1000197

Weber, A., Darmstadt, G. L., & Rao, N. (2017). Gender disparities in child development in the east Asia-Pacific region: A cross-sectional, population-based, multicountry observational study. *The Lancet Child & Adolescent Health, 1*(3), 213–224.

Wen, M., & Lin, D. (2012). Child development in rural China: Children left behind by their migrant parents and children of nonmigrant families. *Child development, 83*(1), 120–136.

Winsler, A., & Wallace, G. L. (2002). Behavior problems and social skills in preschool children: Parent-teacher agreement and relations with classroom observations. *Early Education & Development, 13*(1), 41–58. https://doi.org/10.1207/s15566935eed1301_3

Woods, A. D., Gerasimova, D., Van Dusen, B., Nissen, J., Bainter, S., Uzdavines, A., Davis-Kean, P. E., Halvorson, M., King, K. M., Logan, J. A. R., Xu, M., Vasilev, M. R., Clay, J. M., Moreau, D., Joyal-Desmarais, K., Cruz, R. A., Brown, D. M. Y., Schmidt, K., & Elsherif, M. M. (2024). Best practices for addressing missing data through multiple imputation. *Infant and Child Development, 33*(1). https://doi.org/10.1002/icd.2407

Xie, W., Sandberg, J., Uretsky, E., Hao, Y., & Huang, C. (2021). Parental Migration and Children's Early Childhood Development: A Prospective Cohort Study of Chinese Children. *Population Research and Policy Review*, 1–30.

Zar, J. H. (2014). Spearman rank correlation: Overview. *Wiley StatsRef: Statistics Reference Online*.

Zhao, J., Brinkman, S. A., Zhang, Y., Song, Y., Lu, C., Young, M. E., Zhang, Y., Ip, P., Shan, W., & Jiang, F. (2020). Measuring early childhood development with The Early Human Capability Index (eHCI): A reliability and validity study in China. *BMC Pediatrics, 20*(1). https://doi.org/10.1186/s12887-020-02210-w

Zippert, E. L., & Ramani, G. B. (2017). Parents' estimations of preschoolers' number skills relate to at-home number-related activity engagement. *Infant and Child Development, 26*(2), e1968. https://doi.org/10.1002/icd.1968

Authors and Affiliations

Zeyi Li[1] · **Nirmala Rao**[1]

✉ Zeyi Li
zeyili@connect.hku.hk

✉ Nirmala Rao
nrao@hku.hk

[1] Faculty of Education, The University of Hong Kong, PokFuLam Road,
999077 Hong Kong Island, Hong Kong SAR, China