Research Article https://doi.org/10.12973/ijem.11.2.245

International Journal of Educational Methodology
Volume 11, Issue 2, 245 – 266.
ISSN: 2469-9632
https://www.ijem.com/

An Early Numeracy Digital Brief Assessment: Parametric and Non-parametric Item Response Theory Models

Cecilia Marconi*
Universidad de la República, URUGUAY

Dinorah de León
Universidad de la República, URUGUAY

Mario Luzardo
Universidad de la República, URUGUAY

Alejandro Maiche
Universidad de la República, URUGUAY

Received: December 26, 2024 ▪ Revised: March 8, 2025 ▪ Accepted: May 5, 2025

Abstract: Developing efficient and reliable tools for assessing early mathematical skills remains a critical priority in educational research. This study aimed to develop and validate a brief version of the Prueba Uruguaya de Matemática (Uruguayan Mathematics Test, PUMa), a digital tool to assess mathematical abilities in children aged 5 to 6. The original test included 144 items covering both symbolic (66%) and non-symbolic (34%) tasks, such as approximate number system, counting, numerical ordering (forward and backward), math fluency, composition and decomposition of numbers, and transcoding auditory-verbal stimuli into Arabic-visual symbols. Unlike most existing tools that require individual administration by trained professionals and lack cultural adaptation for Latin American contexts, PUMa is self-administered, culturally grounded, and suitable for large-scale assessments using tablets. Using a sample of 443 participants and applying parametric and non-parametric models within the framework of Item Response Theory (IRT), along with correlations with TEMA-3, preliminary evidence was generated showing that the brief version retained precision and validity. The resulting shortened tests included 69 and 73 items for the parametric and non-parametric versions, yielding a balanced representation of symbolic (56%) and non-symbolic (44%) tasks. Despite item reduction, ability scores remained highly correlated between original and brief versions ($r > .90$), and both brief versions demonstrated strong internal consistency ($\alpha = .94$). PUMa improves upon existing assessments by combining cultural relevance, group-based digital administration, and real-time data collection, offering a scalable solution for early identification and intervention. These features support personalized educational strategies that foster cognitive and academic development from the earliest stages.

Keywords: *Early numeracy assessment, item response theory, kernel smoothing IRT, parametric/non-parametric IRT models, symbolic/non-symbolic mathematics skills.*

To cite this article: Marconi, C., de León, D., Luzardo, M., & Maiche, A. (2025). An early numeracy digital brief assessment: Parametric and non-parametric item response theory models. *International Journal of Educational Methodology*, *11*(2), 245-266. https://doi.org/10.12973/ijem.11.2.245

Introduction

Early numeracy skills play a crucial role in shaping mathematical development and lay the foundation for long-term educational trajectories. Children who enter school with weaker math skills are more likely to be placed in lower-level tracks during high school, reducing their chances of pursuing higher education (Archbald & Farley-Ripple, 2012). Conversely, strong numerical and relational knowledge at the onset of formal education fosters sustained progress in mathematics throughout primary school (Aubrey et al., 2006). These findings highlight the importance of early numeracy skills, which research consistently links to later mathematical achievement (Duncan et al., 2007; Romano et al., 2010).

Many children face specific learning difficulties, such as dyscalculia, which often persist and exacerbate educational disparities over time (Davis-Kean et al., 2022). Early identification of mathematical difficulties is thus essential for timely intervention. Many existing assessments—the Basic Mathematical Competence Test of Early Mathematics Ability–Third Edition (TEMA-3) (Ginsburg et al., 2007) and PENS-B (Purpura et al., 2015) require one-on-one administration, trained personnel, and specialized materials. This restricts their scalability, especially in contexts with limited resources. However, in recent years, digital tools have emerged as a promising and scalable solution. Technology-based interventions—such as interactive applications and self-administered assessments—have demonstrated the potential to

* **Corresponding author:**
Cecilia Marconi, Universidad de la República, Uruguay. ✉ cecil29ster@gmail.com

support children with mathematical learning difficulties through personalized, adaptive, and cost-effective approaches (Benavides-Varela et al., 2020).

Despite this, the limited availability of culturally and linguistically adapted tools remains a barrier, especially in Spanish-speaking and Latin American contexts. Because many of these assessments were developed in different cultural contexts, their validity and applicability in diverse populations are limited. This challenge remains particularly acute in Latin America, where many children still lack access to assessment tools that are adapted to their linguistic, cultural, and educational contexts. As a result, there is a pressing need for robust and contextually relevant instruments that can generate reliable data to inform research and practice in education.

In response to these challenges, some recent efforts in Latin America have focused on developing more culturally relevant tools. For instance, Chile has implemented locally developed instruments (Fritz et al., 2017), and Uruguay has a longstanding tradition of research in mathematical cognition, which has led to the creation of several context-specific tools (Koleszar et al., 2020). One such tool developed a decade ago is the Arithmetic Calculation Efficiency Test (TECA) (Singer & Cuadro, 2014), a standardized assessment designed to evaluate arithmetic efficiency through basic operations such as addition, subtraction, multiplication, and division. While TECA is useful for identifying later learning difficulties, it is not suited for assessing early numeracy skills such as number sense or counting, leaving a critical gap in tools for evaluating foundational mathematical abilities in young children.

Despite these advancements, a notable gap remains in the availability of tools specifically designed to assess foundational and intuitive early numeracy skills—such as number sense, one-to-one correspondence—which precede formal instruction and are critical for later mathematical development. Addressing this gap is essential for the creation of comprehensive early assessment systems that promote educational equity from the initial stages of schooling.

Study Rationale

Building upon the need for culturally appropriate tools mentioned in the theoretical background, this study aims to respond to the persisting gap in early numeracy assessment instruments available in Spanish-speaking Latin American contexts. To this end, we developed the Prueba Uruguaya de Matemática (PUMa), a digital, self-administered screening tool specifically designed for children aged 5 to 6 (Maiche et al., 2020). PUMa provides a culturally relevant and scalable means of assessment for preschool-aged children, bridging a critical gap in early education by evaluating both symbolic and non-symbolic mathematical skills. Its digital format allows for rapid administration and real-time data collection, which enhances accessibility and scalability of early assessments. This is particularly relevant in the Uruguayan context, where a national policy for integrating technology into education—exemplified by the Ceibal initiative (Ceibal, n.d.)—ensures universal access to tablets and internet in public schools. As a result, PUMa can be administered simultaneously to entire classes without requiring trained personnel, offering educators a powerful tool to understand and support early mathematical learning at scale.

In this paper, we present the theoretical framework behind PUMa and detail the process of developing a brief version: a shorter and more practical alternative to the original test. Using both parametric and non-parametric Item Response Theory (IRT) models and correlations with the TEMA-3 test, we assess the psychometric performance of the shortened version, which includes about half the original items. This new brief version of PUMa is intended to bridge a critical gap in early-stage assessment, offering educators, researchers, and policymakers a robust and scalable resource to better understand and support young learners in Uruguay.

Our goal is to contribute an efficient and scalable tool for the early identification of numeracy challenges, facilitating timely and personalized educational interventions that promote equity from the earliest stages of formal education.

The PUMa Test: Construct and Measurement

PUMa is an innovative, online, self-administered screening tool designed to evaluate early mathematical abilities in preschool and first-grade children. Children assessed by PUMa receive audio instructions via headphones, which guide them through the assessment. This approach allows the simultaneous evaluation of an entire school class without requiring the presence of a specialized technician.

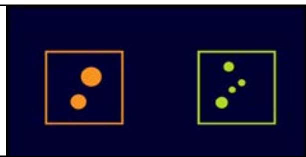
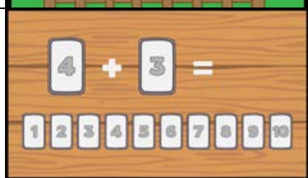
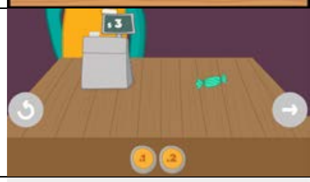
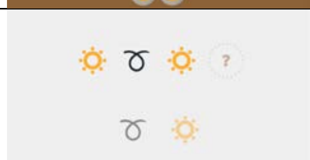
The initial version of PUMa comprises eight targeted tasks and aims to identify students at risk in foundational areas of mathematics (see Appendix A for additional task information). PUMa places particular emphasis on the non-symbolic component—an essential skill set typically developed prior to formal schooling. Unlike traditional assessments, which often prioritize symbolic knowledge, this test provides a comprehensive approach by encompassing both the symbolic and non-symbolic mathematical foundations essential for early math learning.

Research underscores the interconnected roles of symbolic and non-symbolic processing in mathematical development, highlighting the importance of early and comprehensive assessments. Each task within PUMa evaluates a distinct aspect of numeracy, collectively encompassing 144 items (see Appendix B for item details). The items were developed in collaboration with psychologists and educational therapists specializing in mathematics, all of whom have extensive experience in clinical and educational settings, working with both typically developing children and those with specific mathematical difficulties. The design was informed by current literature on mathematical skills from a cognitive psychology perspective.

The tasks are framed within an engaging storytelling context where two characters, guided by their teacher, journey through diverse locations in Uruguay to solve math-related challenges. This narrative approach is designed to captivate and motivate young learners while maintaining a clear focus on the assessment goals.

Tasks are presented in a fixed sequence: each one starts automatically upon completion of the previous one, and tasks cannot be skipped. In turn, each task ends according to two independent criteria: either when the child completes all the trials or when 3 minutes pass. The group-administered format of PUMa includes an introductory briefing, with audio instructions provided at the outset of each session. The platform ensures a secure evaluation process, with sessions averaging 20.8 minutes in duration. Table 1 provides an overview of the tasks, including auditory prompts and visual elements presented to the children during the assessment (see Appendix A for additional task information). This innovative design offers a dynamic and child-friendly approach to identifying early numeracy challenges.

Table 1. PUMa Test Composition

Area	Audio says	Item
Approximate number system (ANS)	Touch the side where there are more fireflies	
Counting (CON)	Load the number of stones that the order indicates	
Numerical ordering forward (SNP)	Order the stones from smallest to largest	
Numerical ordering backward (SNR)	Order the stones from largest to smallest	
Transcoding auditory-verbal stimuli to Arabic-visual symbols (TRA)	Touch the sheep that has the number you heard	
Math fluency (CMV)	Add as fast as you can the total number of animals	
Composition and decomposition of numbers (CYD)	Choose the amount of coins to pay the exact price of the snack	
Patterns (PAT)	Complete the pattern by touching the missing symbol	

Approximate Number System

The approximate number system (ANS) is a component of number sense that involves the ability to estimate non-symbolic quantities in a non-exact way (Libertus et al., 2011). The ANS is present in many species as well as in humans and it is documented as early as birth. Methodologies used to evaluate ANS utilize comparison of sets where the participants should indicate which set is larger without counting or simply flashing dots on screen and asking how many dots were shown (Odic & Starr, 2018). In accordance with the characteristics of the stimulus, there are a number of variables that make the task easier or harder to do. One of them is the ratio between the sets of dots presented. For example, babies as young as six months old can discriminate 4 from 8 dots and 8 from 16, both differences with a 1:2 ratio (Halberda & Feigenson, 2008). Other visual features used in this type of task include the differences in size of the stimulus (large or small points) and the total area occupied by the whole set, that is, if the points are together or apart called convex hull and the congruent or incongruent combinations (DeWind & Brannon, 2016).

Several studies have shown the implication of ANS for learning symbolic math (Hyde et al., 2014; Szkudlarek & Brannon, 2017). For example, the acuity in ANS is positively correlated with math in school years (Libertus et al., 2011).

Numerical Ordering Forward

Counting is widely recognized in the literature as a key ability due to its significant role in mathematics. This competency involves understanding the sequential order of numbers—knowing which number comes before or after another in a progressive manner (Lyons & Beilock, 2011). To count accurately, children must not only know the number-word sequence but also understand that each item should be counted only once, following the correspondence principle. This foundational skill typically develops between the ages of 2 and 5 and has been identified as a strong predictor of later mathematical success (Aunio & Niemivirta, 2010; Morsanyi et al., 2024).

Numerical Ordering Backwards

Learning to count backwards correctly takes longer for children, likely because they must first master the forward sequence (Gilmore & Batchelor, 2021). Verbal count sequence knowledge underpins numeral order processing in children. This skill relies heavily on working memory and is influenced more by age than by formal schooling (Dellatolas et al., 2000).

Counting

We assess counting through one-to-one correspondence, a foundational skill and prerequisite for the ability to count and understand numbers (Clements & Sarama, 2015). It involves matching each item with a corresponding counterpart, a critical cognitive ability in early math development. For example, children might be tasked with matching a set of stones to a given number of points. Notably, this skill has been shown to predict children's future math performance (Seitz & Weinert, 2022).

Transcoding Auditory-Verbal Stimuli to Arabic-Visual Symbols

Number transcoding is the process whereby a number is spoken aloud, and children are required to represent it using symbols. Mastering the relationship between spoken and written multi-digit numerals lays the foundation for comprehending the numerical system (Habermann et al., 2020). The mental process behind transcoding involves converting information from a verbal numerical expression into a written representation, effectively changing its code.

Two cognitive models have been proposed to explain transcoding. Semantic models suggest that an abstract representation mediates the connection between understanding numerical input and producing numerical output. These models propose that the transcoding process is managed by an algorithmic mechanism (Moura et al., 2013)

Math Fluency

To solve basic math problems, children must be fluent in mentally performing number combinations. Therefore, foundational skills like number sense, subitization, number comparison, and counting are recognized as essential for developing mathematical fluency (Reigosa-Crespo & Estévez-Pérez, 2023). The development of mathematical fluency is a gradual process. Mastery of counting is typically a prerequisite, as it provides the foundation for future calculation strategies. For instance, children often use counting-on strategies, such as starting from the larger addend and counting up by the smaller one, to solve addition problems efficiently (Gliksman et al., 2022).

Patterns

Patterns are a fundamental part of mathematics and have long been associated with early mathematics learning. This ability involves visual skills and mathematical thinking, enabling children to understand relationships through abstraction and generalization (Mulligan et al., 2006). Repeating patterning knowledge draws on multiple cognitive skills, including relational reasoning, executive function, and spatial skills (Collins & Laski, 2015; Miller et al., 2016)

Composition and Decomposition

Numerical composition is the ability to determine changes in a number or quantity from the initial and final values (Purpura, 2010). In order to compose or decompose quantities, children must first master foundational skills, such as verbal or object counting—being able to say number words correctly and understanding the principles of counting (Baroody, 1987). Once children develop these abilities, typically between the ages of 4 and 6, they are then able to engage in problem-solving tasks (Olgan et al., 2017).

Methodology

Participants

The study sample consisted of 443 children, ages five to six, enrolled in preschool and first grade across eight private schools in Uruguay. The sample was collected in November 2020 and November 2022. Due to pandemic-related access constraints, a convenience sampling approach was employed through established connections with schools for participant recruitment. While this method facilitated access during a challenging period, it may have introduced sampling bias. To mitigate this risk, efforts were made to include schools from diverse socioeconomic backgrounds, thereby enhancing the representativeness of the sample. Additionally, complementary studies are currently underway employing randomized sampling strategies to further address this limitation. Approximately half of the participants were preschoolers (49.4%), with the remaining portion comprising first-grade students (50.6%). Gender distribution was balanced (49% female, 51% male), as confirmed by a chi-square test of homogeneity ($p = .6689$).

To support concurrent validity measures, 184 of the 443 participants were also assessed with the TEMA-3 within the same week. These 184 students were not randomly selected due to pandemic-related access constraints but rather comprised those who allowed evaluation under the prevailing health regulations. To minimize potential biases, the order of test administration was randomized, ensuring a balanced distribution of administration effects on the results.

Testing Procedure

The data collection protocol received approval from the Ethics Committee of the Universidad de la República. Approvals were obtained from school principals to administer the assessment within their institutions. The test was delivered through tablets, with each child provided with their own tablet and headphones. In the Uruguayan context, potential cultural biases related to digital test formats are mitigated by the widespread integration of technology in public education, particularly through the national Ceibal initiative (Ceibal, n.d.), which has provided students and schools with access to digital devices for over a decade. There is a well-established culture of using tablets in early childhood education, both in public and private institutions. Private schools, which composed our sample, also commonly incorporate digital tools into their pedagogical practices. As a result, tablet-based assessment is generally perceived as a natural and familiar activity for young children in Uruguay, reducing the likelihood of performance being affected by the digital format.

Before full implementation, a pilot study was conducted to refine the test. Feedback from teachers was gathered to improve instructions and the overall testing experience. The assessments were carried out in classroom settings by undergraduate psychology students who had received prior training. Some students participated through a practicum course where they learned theoretical aspects of mathematical cognition along with hands-on test administration practice. Others conducted their undergraduate or master's theses within this research project.

To ensure standardized administration, a strict testing protocol was followed. First, the procedure was explained to the entire class, allowing both children and teachers to ask questions. Although teachers remained present, they were asked to minimize their intervention. Children were instructed to raise their hands if they needed assistance, at which point a test administrator would approach them. Each child completed the assessment on a tablet with their own set of headphones. Due to the nature of the test, which required the use of headphones, children remained focused on their tablets throughout the assessment, while seated in their usual spots within the classroom and preserving the natural arrangement of the learning environment.

Only children who were absent on the day of the group administration were tested separately. In these cases, they were taken to a dedicated space, often the school library or another available empty classroom, to complete the assessment in a quiet setting.

For the TEMA-3 administration, the starting point was determined by the child's age-appropriate entry item, assuming no developmental disorders. Testing continues until ceiling and floor criteria are established: a ceiling is defined as five consecutive incorrect responses, and a floor as five consecutive correct responses. The direct score was calculated based on the number of items completed up to the ceiling. This adaptive administration design allowed for difficulty levels to adjust to each child's performance, meaning not all children completed the same set of items, even within the same age group.

Parametric and Non-Parametric IRT models

IRT models are a family of mathematical models widely used in educational assessments to analyze the relationship between latent traits and item responses. Within this framework, two methodological approaches were applied to develop a brief version of the PUMa test: parametric IRT models, which rely on predefined mathematical functions, and non-parametric IRT techniques, which provide greater flexibility in estimating item characteristics without strict assumptions about their functional form.

The parametric models have been widely applied in educational assessment due to their mathematical simplicity and interpretability, and extensive literature supports their use in various psychometric contexts (e.g., Baker, 1985; Baker & Kim, 2004; Lord, 1980; Rasch, 1960; Reise & Moore, 2023). Traditional IRT models, such as the Rasch model and the one-, two-, and three-parameter logistic models, define item characteristic curves (ICCs) parametrically, using a fixed number of parameters to model the probability of a correct response as a function of the latent trait. However, such parametric models have several limitations. Specifically, they assume monotonic ICCs and a logistic functional form, which may not accurately capture the complexity of real-world data, especially when items exhibit non-monotonic patterns or diverge from logistic shapes (Douglas, 1997; Douglas & Cohen, 2001; Ramsay, 1991; Xu & Douglas, 2006). Furthermore, methods for estimating the ICC include joint maximum likelihood estimation, marginal likelihood estimation, and conditional maximum likelihood estimation; but if the assumptions are violated, estimates of item parameters and skill are poor.

In response to the limitations of parametric IRT models, non-parametric methods have been developed to provide greater flexibility in estimating ICCs. Mokken's models introduced foundational concepts based on monotonicity and double monotonicity, relaxing the strict assumptions of parametric models while maintaining essential ordering properties. Ramsay (1991) further advanced non-parametric IRT by introducing a regression approach based on kernel smoothing, enabling smooth, flexible ICC estimates without enforcing monotonicity constraints. The smoothing technique uses local averaging to estimate the relationship between the latent trait and the probability of choosing the correct response (Rajlic, 2020). In Ramsay's method, the kernel smoothing estimator computes ICCs as weighted averages of responses, applying Nadaraya-Watson weights (Nadaraya, 1964; Watson, 1964) to achieve adaptability in response patterns. This approach allows the ICC to vary continuously, accommodating diverse item response patterns beyond those defined by parametric models. This results in more flexible ICCs that provide a closer approximation to the true ICCs compared to those generated by parametric IRT models (van der Linden & Hambleton, 1997).

In kernel smoothing methods, bandwidth selection is essential, as it governs the trade-off between bias and variance in the resulting estimates. A smaller bandwidth produces estimates with lower bias but higher variance, while a larger bandwidth increases bias and decreases variance, affecting the smoothness of the ICC curve. Despite the importance of this parameter, no established method exists for identifying an optimal bandwidth in ICC contexts (Xu & Douglas, 2006). In this study, to optimize bandwidth selection with available computational resources, we employed least-squares cross-validation, allowing for adaptive refinement in ICC estimation to balance precision and stability across varying response patterns.

While Ramsay's kernel smoothing method provides advantages, including computational efficiency, ease of implementation, and effectiveness for moderate sample sizes, it does not enforce the monotonicity of ICCs, a standard assumption in IRT models. To address this, recent developments have introduced the non-parametric isotonic model, as proposed by Luzardo and Rodríguez (2015). This model incorporates an isotonic estimator for ICCs, grounded in the methodology established by Dette et al. (2006), ensuring that the monotonicity constraint is preserved without sacrificing the flexibility of non-parametric approaches. By maintaining the monotonicity of ICCs, this isotonic model overcomes one of the primary limitations associated with kernel smoothing in IRT applications, enhancing its applicability across psychometric assessments.

To enforce monotonicity in ICC estimation, Luzardo and Rodríguez (2015) introduced an innovative approach that estimates the inverse of the ICC in a monotonic manner. The final ICC estimator is then derived by reflecting this inverse function along the bisector of the first quadrant, ensuring that the resulting ICC adheres to the monotonicity constraint typically expected in IRT models. In this framework, the pseudo-difficulty parameter is defined as the ability level at which the ICC reaches 0.5, while pseudo-information is a function derived from the first derivative of the isotonic ICC, used to approximate Fisher information across the theta scale. The pseudo-difficulty parameter and the pseudo-information function are analogous to the difficulty parameter and the Fisher information in the parametric framework. Note that higher information implies lower error in the estimation of ability.

In summary, although traditional parametric IRT models offer a straightforward and interpretable structure for ICC estimation, non-parametric methods like Ramsay's kernel smoothing approach and the isotonic model by Luzardo and Rodríguez provide enhanced flexibility and adaptability.

Analytic Procedure

The analytic procedure for developing the brief version of the PUMa test was structured into distinct stages to ensure both validity and reliability. In Stage 1, the psychometric properties of the initial version of PUMa test were analyzed, forming the foundation for the creation of its brief version. Stage 2 involved three critical steps: first, item quality was evaluated using Classical Test Theory (CTT) to assess psychometric properties and identify items with high rates of missing data; second, ICCs were modeled using both parametric and non-parametric IRT methods, allowing a comparison of traditional and flexible estimation techniques; and third, a rigorous item selection process was conducted under both IRT approaches to retain items with high discriminative power and reliable measurement across latent ability scale. Finally, Stage 3 assessed criterion validity by comparing the brief PUMa test with TEMA-3 within the same participant group, providing evidence for concurrent validity and reinforcing its application for early numeracy assessment.

Stage 1 - Analysis of the Psychometric Properties of the Initial Version of PUMa

This stage focused on evaluating the internal consistency and correlation of the PUMa tasks. The correlation between the total PUMa sum score and each individual task sum score was as follows: ANS = 0.66, CON = 0.59, PAT = 0.49, CMV = 0.86, SNP = 0.76, SNR = 0.76, TRA =0.7, and CYD = 0.74, with all correlations statistically significant at *p-value* < .01. The correlation analysis reveals that symbolic skills (CMV, SNP, SNR, TRA, CYD) show stronger positive associations with the overall PUMa score, each with coefficients above 0.7, whereas non-symbolic skills (ANS, CON, PAT) exhibit moderate correlations, ranging from 0.49 to 0.66, highlighting the distinct roles of symbolic and non-symbolic abilities in contributing to total PUMa performance. As summarized in Table 2, non-symbolic tasks generally exhibit weaker correlations with other tasks, with the ANS task notably showing the lowest inter-task correlations, ranging from 0.26 to 0.35.

Table 2. Correlation Matrix - PUMa Tasks

	ANS	**CON**	**PAT**	**CMV**	**SNP**	**SNR**	**TRA**	**CYD**
PUMa	0,66	0,59	0,49	0,7	0,86	0,76	0,76	0,74
ANS		0,29	0,29	0,27	0,35	0,29	0,28	0,26
CON			0,33	0,53	0,5	0,51	0,49	0,43
PAT				0,38	0,37	0,43	0,22	0,36
CMV					0,71	0,68	0,61	0,74
SNP						0,72	0,58	0,64
SNR							0,61	0,6
TRA								0,48
CYD								

Evidence of internal consistency was measured using Cronbach's alpha, which yielded a coefficient of 0.98 for the complete PUMa scale, indicating high reliability. The Cronbach's alpha values calculated for each task within PUMa were as follows: ANS = 0.87, CON = 0.73, PAT = 0.71, CMV = 0.97, SNP = 0.92, SNR = 0.92, TRA = 0.87, and CYD = 0.96. These coefficients, all above the acceptable threshold of 0.7, indicate a robust internal consistency, reinforcing the scale's reliability in consistently assessing early numeracy skills across varied tasks.

Additionally, both the parametric and non-parametric brief versions demonstrated good internal consistency, with a Cronbach's alpha of .94 despite the reduced number of items.

Stage 2 - Development of a Brief Version of the Test

Step 1 - Analysis of Item Quality

Before estimating the ICCs, we conducted an analysis of item properties based on CTT. It is well-documented that items demonstrating robust psychometric properties within CTT are more likely to exhibit satisfactory fit within IRT models. This approach is commonly employed in large-scale international assessments, such as PISA and LLECE, to pre-screen items. Consistent with these established practices, we included only items that meet specific criteria: a CTT difficulty parameter (proportion of correct responses) between 0.1 and 0.9, and a biserial correlation of at least 0.3. Applying these criteria led to the elimination of 15 items. Additionally, items with more than 200 missing responses were excluded, resulting in the removal of 11 further items.

Step 2 - Modeling ICCs Using Parametric and Non-parametric Approaches

The second step focused on estimating the ICCs using both parametric and non-parametric approaches. Parametric IRT models were applied to simultaneously estimate item parameters and examinee abilities. This approach assumes that abilities follow a normal distribution with a mean of 0 and a standard deviation of 1. Various parametric IRT models, including the Rasch, one-, two-, and three-parameter logistic models (2PL, 3PL), were applied across different areas (e.g., ANS, Counting), followed by an evaluation of model fit and item fit. Items that did not conform to the selected model (*p-value* < 0.05) were excluded from further analysis. Once the ICCs were estimated, the parametric information function was also calculated to assess the precision of ability estimation at different levels of the latent trait.

Table 3 presents a summary of the selected parametric models (Rasch, 2PL, 3PL) along with model fit statistics. Model selection for each task was determined by examining the number of misfitting items and the overall goodness-of-fit of the model. Model-data fit was evaluated using several fit indices, including the M2 statistic (Maydeu-Olivares & Joe, 2006), the Root Mean Squared Error of Approximation (RMSEA) from model chi-square, the Standardized Root Mean Square Residual (SRMSR), the Tucker-Lewis Index (TLI), and the Comparative Fit Index (CFI), each of which provides insight into the adequacy of model fit.

As detailed in Table 3, the M2 statistics indicate a non-significant result (*p-value* > 0.05) for tasks CON, SNP, SNR, TRA, CMV, and CYD, suggesting an acceptable fit of the specified model to the observed data for these items. Conversely, for tasks ANS and PAT, the M2 statistic yielded a significant result (*p-value* < 0.05), indicating that the model did not adequately capture the underlying data structure for these particular items. Although the selected models represented the best fit among the Rasch, 2PL, and 3PL models for each respective task, the fit remained unsatisfactory for the ANS and PAT tasks within the parametric framework.

Following the parametric analysis, the estimation of ICCs continued using two non-parametric approaches: Ramsay's model and the Isotonic model. Ramsay's approach requires the construction of weights as outlined by Nadaraya and Watson (Nadaraya, 1964; Watson, 1964). In this procedure, the summed scores were defined as the statistic T. Initially, examinees *i*th were ranked based on Ti values, which were then transformed into quantiles of a chosen distribution—in this case, the normal distribution. Response patterns were subsequently ordered by the estimated ability rankings. The ICC estimation was performed by smoothing the relationship between each binary item response and the ability vector. A critical component in this technique is the selection of the bandwidth parameter (h-parameter). For this study, the optimal h value was determined using the *npregbw* function from the *np* package (Hayfield & Racine, 2008) in R, applying an Epanechnikov kernel with least-squares cross-validation. Following this, ICCs and the pseudo-information, as proposed by Luzardo (2019), were estimated using the isotonic non-parametric model.

Figure 1 provides a comparative visualization of the ICCs estimated under the three approaches (parametric, Ramsay's kernel smoothing, and Luzardo's isotonic) for selected sample items, with one item represented from each task. The ICC represents the probability of a correct response to an item as a function of the respondent's ability. This illustration highlights the differences in model behavior and fit across approaches, offering insights into the varying levels of flexibility and adherence to empirical data achieved by each method.

Step 3. Item Selection Process

Following the estimation of ICCs, the item selection process was conducted to construct a psychometrically robust brief version of the test using both parametric and non-parametric methods. The procedure was as follows: (i) Items were ordered in ascending difficulty based on the b-parameter from the parametric ICC estimation or the pseudo-difficulty parameter in the non-parametric isotonic model; (ii) a grid with an interval width of 0.2 was established across the range of b or pseudo-b values, and items were categorized within each interval accordingly; (iii) The number of items for each task in the brief version was determined by considering the relative weight of each task within the symbolic and non-symbolic domains. In addition, the initial number of items for each domain reflected their proportional representation in the original full test. Subsequently, the most informative items within each task were selected according to the criteria described as follows. Finally, each task in the brief version was reviewed, and items that did not contribute meaningfully to the construct were removed to obtain the final version. Our goal was for each scale to retain its psychometric properties, both individually and in relation to the total score. The process resulted in greater balance between symbolic and non-symbolic domains. Following this, items were selected based on the criterion of maximum information within each grid interval. In the case of the parametric approach, items with the highest discrimination parameters within each interval were chosen, as these provide more information. For the non-parametric approach, items with the highest pseudo-information were selected.

Stage 3 - Evidence of Criterion Validity

To establish criterion validity for the PUMa test, this study utilized the TEMA-3 assessment as an external benchmark or "gold standard." Initially, the constructs measured by both assessments were analyzed and compared to ensure conceptual alignment. ubsequently, using data from a sample of 184 students who completed both assessments, correlations between the PUMa and TEMA-3 scores were calculated, and Fisher's z-test was applied to assess whether differences in their correlation coefficients were significant.

Table 3. Parametric Estimations

	ANS	CON	PAT	SNP	SNR	TRA	CMV	CYD
N	376	435	443	439	384	432	436	437
Specified Model	2P	2P	3P	3P	2P	3P	2P	3P
M2 model fit statistic	748.7	13.8	62.9	3.5	25.97	144.96	189.5	63.32
Df	434	14	25	3	35	133	189	63
p-value	0	0.464	0	0.317	0.87	0.23	0.476	0.46
RMSEA	0.047	0	0.059	0.027	0	0.02	0.003	0
RMSEA_5	0.041	0	0.041	0	0	0	0	0
RMSEA_95	0.053	0.059	0.077	0.1145	0.026	0.03	0.028	0.036
SRMSR	0.104	0.061	0.055	0.173	0.115	0.052	0.181	0.071
TLI	0.868	1	0.923	0.998	1.003	0.996	1	1
CFI	0.877	1	0.957	0.999	1	0.997	1	1

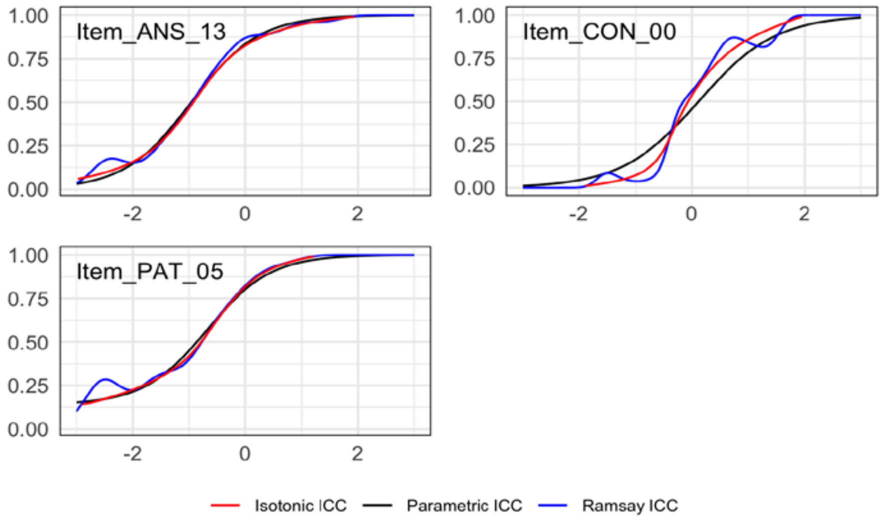


Figure 1. Comparison of Parametric and Non-parametric ICC estimation for a selection of sample items (one for each task) from the non-symbolic domain

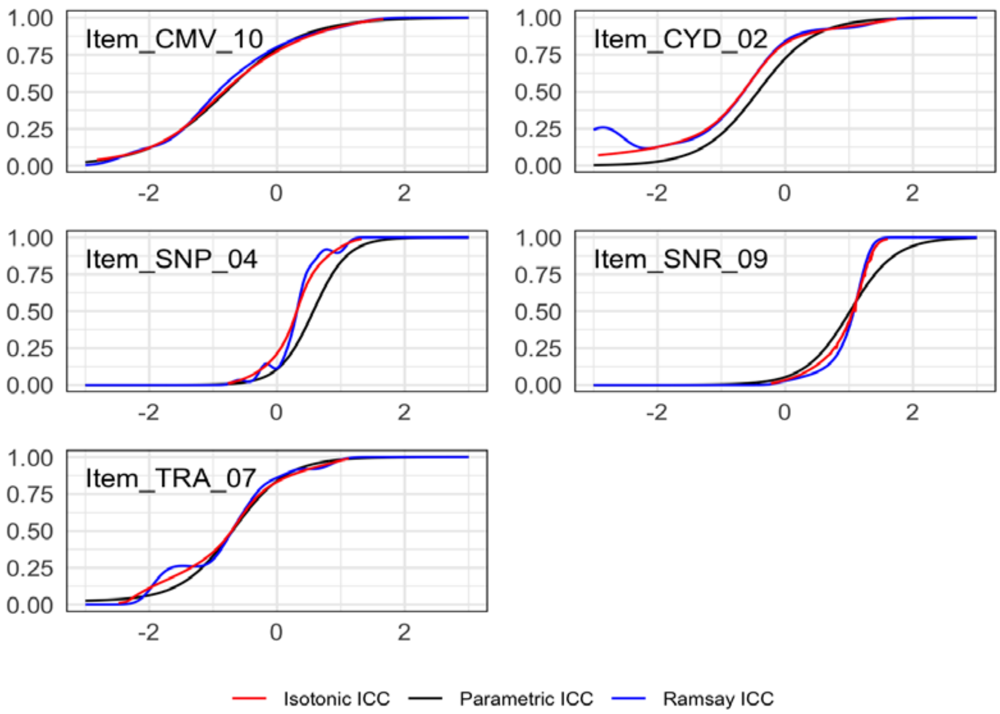


Figure 2. Comparison of Parametric and Non-parametric ICC estimation for a selection of sample items (one for each task) from the symbolic domain

Results

The Brief Test - Parametric Approach

The development of the brief test under the parametric approach was guided by the estimations of parametric ICCs, following the item selection method outlined in the analytic procedure. Table 4 summarizes the initial item pool before the exclusion process, which was subsequently refined by removing items that: (1) did not satisfy the psychometric criteria specified by CTT, (2) contained substantial missing data, or (3) demonstrated poor fit to the parametric model.

The parametric selection process reduced the item pool from 144 to 69 (52% reduction). For Counting (CON) and Progressive Numeric Series (SNP), no item selection was performed due to their limited number of items (6 and 4, respectively).

The initial version of the PUMa test exhibited a greater representation of symbolic skill items, while non-symbolic skills were less represented. The brief parametric version achieved a more balanced distribution between symbolic and non-symbolic skills compared to the original test.

Table 4. Number of Items Retained and Removed for Each Task under Parametric Approach

	Initial- Original Test		Removed: CTT item psychometrics properties not met (1)	Removed high level NAs (2)	Removed parametric model non-fit (3)	Initial- Before brief procedure	Retained: Brief Test	%
ANS	32	22%	1	0	0	31	21	30%
CON	7	5%	0	0	1	6	6	9%
PAT	10	7%	0	0	1	9	5	7%
SNP	10	7%	3	1	2	4	4	6%
SNR	10	7%	0	0	0	10	7	10%
TRA	20	14%	1	0	2	17	10	14%
CMV	34	24%	5	8	5	16	8	12%
CYD	21	15%	5	2	0	14	8	12%
Total	**144**	**100%**	**15**	**11**	**11**	**107**	**69**	**100%**

After defining the items included in the parametric brief test, ability estimates were calculated using both IRT and CTT, with the latter based on the total sum score. Table 5 presents the correlations between IRT-based ability estimates for the initial version of PUMa versus the brief test, as well as the correlations between IRT-based and CTT-based ability estimates.

All correlations were positive and strong (r > .90); CON and SNP reached r = 1.0, as they retained the same items.. These findings under the parametric approach are promising, indicating that with approximately half the original number of items, the brief test achieves comparable precision in ability estimation.

Table 5. Correlation Coefficients Using the Parametric Approach

Correlations: Abilities estimated from the initial and brief test versions under IRT Parametric Models								
	ANS	**CON**	**PAT**	**SNP**	**SNR**	**TRA**	**CMV**	**CYD**
	0.979	1	0.934	1	0.974	0.949	0.917	0.965
Correlations: Abilities estimated under IRT Parametric Models and the Total Sum Score								
	ANS	**CON**	**PAT**	**SNP**	**SNR**	**TRA**	**CMV**	**CYD**
Initial	0.977	0.934	0.967	0.941	0.972	0.952	0.931	0.965
Brief	0.951	0.934	0.884	0.941	0.948	0.911	0.865	0.95

The following section evaluates the non-parametric approach to determine if it achieves similar performance to that of the parametric method.

The Brief Test - Non-parametric Isotonic Approach

Similar to the parametric brief test, the non-parametric version was created by removing items that failed to meet CTT psychometric standards or exhibited a high number of missing responses. Unlike the parametric approach, items identified through ICC estimation were not eliminated beforehand. Item selection was based on estimated pseudo-difficulty parameters, with higher pseudo-information items retained within predefined grid intervals (see Analytic Approach). Table 6 summarizes the item selection results.

Table 6. Number of Items Retained and Removed for Each Task under Non-Parametric Approach

	Initial- Original Test		Removed: CTT item psychometrics properties not met (1)	Removed high level NAs (2)	Initial-Before brief procedure	Retained: Brief Test	%
ANS	32	22%	1	0	31	17	23%
CON	7	5%	0	0	7	7	10%
PAT	10	7%	0	0	10	8	11%
SNP	10	7%	3	1	6	6	8%
SNR	10	7%	0	0	10	7	10%
TRA	20	14%	1	0	19	10	14%
CMV	34	24%	5	8	21	11	15%
CYD	21	15%	5	2	14	7	10%
Total	**144**	**100%**	**15**	**11**	**118**	**73**	**100%**

The non-parametric approach, using isotonic ICC estimates, produced a 73-item brief test (49% reduction), closely matching the parametric method. The final distribution of symbolic (56%) and non-symbolic (44%) items reflects a balanced structure similar to the parametric version. Ability estimates for the non-parametric brief test were calculated using IRT and CTT, and correlations were computed. As shown in Table 7, most correlations exceeded 0.9, except CYD (r = .884). For CON and SNP tasks, correlations reached 1.0, as no brief version was developed for these tasks due to limited item availability.

Table 7. Correlation Coefficients Using the Non-Parametric Approach

Correlations: Abilities estimated from the initial and the brief test versions under IRT Non-Parametric Models								
	ANS	**CON**	**PAT**	**SNP**	**SNR**	**TRA**	**CMV**	**CYD**
	0.94	1	0.942	1	0.919	0.903	0.916	0.884
Correlations: Abilities estimated under IRT NonParametric Models and the Total Sum Score								
	ANS	**CON**	**PAT**	**SNP**	**SNR**	**TRA**	**CMV**	**CYD**
Initial	0.957	0.973	0.967	0.921	0.939	0.968	0.948	0.948
Brief	0.915	0.973	0.94	0.921	0.912	0.908	0.921	0.898

Both approaches achieved comparable psychometric performance, reducing test length by approximately 50% while maintaining high correlations with ability estimates from the full test.

Preliminary Evidence of Criterion Validity

In this analysis, a sample of 184 students who completed both the TEMA-3 and PUMa assessments was examined. Although the TEMA-3 and PUMa tests differ slightly in structure, as shown in Table 8, both evaluate constructs closely related to early numeracy skills. To explore the criterion validity of the PUMa test, the correlation between the total scores of the initial version of the PUMa test and TEMA-3 was calculated, yielding a substantial positive correlation of 0.77. This result aligns with findings that strong correlations between similar numeracy assessments support criterion validity by demonstrating that the test effectively measures targeted early mathematical skills.

Table 8. Composition of the PUMa test and TEMA-3 by Numeracy Domains, Assessed Tasks and Number of Items

Domain	Task	TEMA-3	PUMa
Numeracy	Counting (CON)	13	7
	Progressive Numeric Series (SNP)	10	10
	Regressive Numeric Series (SNR)	2	10
Comparison	Approximate Number System (ANS)	1	32
	Number comparison	2	0
Arithmetic	Visual Mental Calculation (CMV)	25	34
Concepts	Composition and Decomposition (CYD)	0	21
Conventions	Transcoding(TRA)	9	20
Visuospatial Skills	Patterns(PAT)	0	10
Visuospatial Skills	Number conservation	1	0
Mental manipulation of quantities*	mental number line, equivalent distribution of quantities, tens and hundreds, commutativity rule	12	
Total		72	144

*Note: These skills present in the last 25 items of the topic-3 test are designed for 8-year-old children.

Correlations between the TEMA-3 scores and the total sum scores of the brief versions of PUMa were calculated, revealing a correlation of 0.68 for the parametric brief version and 0.73 for the non-parametric brief version. Additionally, correlations were calculated for each task between the total sum scores of PUMa (in both its initial and brief versions) and the TEMA-3 scores, as well as between the PUMa ability estimates derived from IRT (initial and brief versions) and the TEMA-3 scores. These correlations were evaluated using both parametric (Table 9) and non-parametric approaches (Table 10) for comparative analysis.

A key finding is that all correlations were positive across both methodological approaches and in both versions of the PUMa test (initial version and brief); see results in Tables 9 and 10. Additionally, the correlation magnitudes were notably consistent between the parametric and non-parametric methods. It is also evident that correlations for tasks assessing non-symbolic abilities (ANS, CON, PAT) were markedly lower than those associated with the symbolic components of PUMa. This pattern likely reflects the lower weighting of non-symbolic items within TEMA-3 and, as mentioned earlier, the adaptive nature of the test's administration.

Table 9. Correlations under the Parametric Approach

Correlations: PUMA Total Sum Score & TEMA-3 Scores								
	ANS	**CON**	**PAT**	**SNP**	**SNR**	**TRA**	**CMV**	**CYD**
Initial	0.366	0.473	0.37	0.707	0.737	0.653	0.68	0.734
Brief	0.362	0.431	0.313	0.687	0.729	0.57	0.651	0.713
Correlations: Abilities estimated under IRT Parametric Models & TEMA-3 Scores								
	ANS	**CON**	**PAT***	**SNP**	**SNR**	**TRA***	**CMV***	**CYD**
Initial	0.387	0.395	0.396	0.673	0.731	0.646	0.647	0.704
Brief	0.385	0.395	0.345	0.673	0.714	0.586	0.569	0.687

(*) Correlations are statistically different according to Fisher´s z-tests

Table 10. Correlations under Non-Parametric Approach

Correlations: PUMA Total Sum Score and TEMA-3 Scores								
	ANS	**CON**	**PAT**	**SNP**	**SNR**	**TRA**	**CMV**	**CYD**
Initial	0.366	0.498	0.416	0.723	0.693	0.653	0.651	0.723
Brief	0.353	0.498	0.416	0.723	0.69	0.586	0.634	0.738
Correlations: Abilities estimated under IRT NonParametric Models and TEMA-3 Scores								
	ANS	**CON**	**PAT**	**SNP**	**SNR**	**TRA***	**CMV**	**CYD**
Initial	0.371	0.478	0.406	0.645	0.647	0.701	0.65	0.645
Brief	0.337	0.478	0.411	0.645	0.664	0.612	0.617	0.695

(*) Correlations are statistically different according to Fisher´s z-tests

Discussion

Developing efficient and reliable tools for assessing early mathematical skills remains a crucial priority in educational research. This study contributes to this effort by providing evidence of the psychometric validity of the initial digital version of the PUMa test, which originally included 144 items, while also addressing the limitations associated with its considerable length. To achieve this goal, we developed and evaluated a brief test of early mathematical skills tailored for Uruguayan children. The digital format allows for quick administration and real-time data collection, enhancing the accessibility and scalability of early assessments. This is particularly relevant in Uruguay, where a national policy for integrating technology into education, exemplified by the Ceibal initiative (Ceibal, n.d.), ensures universal access to digital tools. Through this program, every student is provided with their own tablet, and all schools are equipped with internet connectivity. This fosters an environment conducive to the implementation of innovative educational assessments that ultimately support personalized interventions to enhance children's cognitive and academic growth from an early age.

To ensure consistency in results, two methodological approaches within the IRT framework—parametric and non-parametric—were employed. Both approaches demonstrated high correlations between the abilities estimated from the initial version of PUMa and the brief tests, highlighting that a reduced test with nearly half the number of items can estimate abilities with the same level of precision. This level of efficiency provides substantial benefits, including cutting test administration times in half, reduced participant fatigue, and improved engagement levels for young learners.

Both brief test versions achieved a balance between symbolic and non-symbolic tasks, despite variations in the specific items included in each. This balance, as research highlights that combining symbolic and non-symbolic components enhances the validity of early math assessments by capturing a broader range of foundational numeracy skills (Outhwaite et al., 2024). These results align with discussions in the literature about the relationship between symbolic and non-symbolic number processing. For example, symbolic and non-symbolic number processing tend to be more strongly related within the subitizing range, small quantities easily recognized without counting, suggesting a developmental link between the two components. Non-symbolic processing may influence symbolic number abilities, particularly for small quantities, underscoring the importance of both types of processing in early numeracy development (Hutchison et al., 2020).

In terms of criterion validity, the initial version of PUMa and both brief versions showed moderate to high positive correlations with TEMA-3 scores, a well-established benchmark in early numeracy assessment. As expected, symbolic skills exhibited stronger correlations with TEMA-3, while non-symbolic skills (such as ANS or patterning tasks) showed weaker associations. This pattern likely reflects the stronger emphasis of TEMA-3 on formal and symbolic tasks, as well as the greater variability and perceptual demands inherent in non-symbolic assessments. Additionally, it is possible that non-symbolic tasks engage cognitive processes that are partially distinct—such as visual estimation or attentional filtering—that are not as directly targeted by traditional school-based assessments.. These conceptual and methodological differences may partly explain the lower correlations observed.

This research has practical implications for educational evaluations. By providing a digital screening tool that is effective and efficient, educators and researchers can identify children at risk for mathematical difficulties early on. The ability to collect real-time data offers valuable insights at the classroom level, enabling earlier and more targeted interventions. Additionally, the digital nature of the test allows for greater scalability, ensuring that it can be implemented in diverse educational settings. Importantly, the PUMa platform generates automated reports with task-level breakdowns of student performance. These reports are designed for ease of interpretation and allow educators to identify specific areas of difficulty—such as challenges with numerical ordering or counting. Teachers can use this information to design targeted small-group interventions, adjust instructional pacing, or provide additional activities focusing on particular skills. For example, if a group of students shows low performance in transcoding verbal numbers into written digits, the teacher might implement a set of games or manipulatives that reinforce this specific concept.

While the digital format facilitates rapid and scalable administration, its broader applicability depends on contextual factors such as infrastructure reliability, internet connectivity, and teacher readiness. In Uruguay, these challenges are partially mitigated by the national Ceibal initiative, which provides universal digital access in public schools. Still, successful implementation of tools like PUMa requires basic teacher training, reliable technical support, and continued investment in digital equity—particularly in rural or under-resourced areas.

Finally, the brief format also supports repeated administrations over time, enabling ongoing progress monitoring without inducing fatigue or disengagement. This feature enhances the tool's practicality for both classroom use and research purposes, as it allows educators to track individual learning trajectories and adjust instruction accordingly.

Conclusion

This study provides preliminary validation for a brief, digital version of the PUMa test, demonstrating that test length can be reduced without compromising psychometric integrity or conceptual coverage. By integrating both symbolic and non-symbolic tasks, the shortened version captures a broad spectrum of early numeracy skills and exhibits meaningful correlations with established assessments such as the TEMA-3. The digital format enhances scalability and supports real-time data collection, increasing the tool's value for early screening and intervention in educational contexts with sufficient technological infrastructure.

While these findings are encouraging, future research should continue to examine the predictive validity of the test and assess its adaptability across diverse educational contexts. In this regard, a norming study of the PUMa instrument is currently underway, employing a representative two-stage sampling design—first stratified by department and socioeconomic level, then clustered by school. The study includes 934 students assessed between 2023 and 2024. This ongoing effort is expected to improve the interpretability and utility of the test in varied socio-educational settings and support the development of national systems for monitoring educational trajectories from early childhood.

Recent initiatives have also focused on the cross-cultural adaptation of PUMa. The instrument has been translated into Portuguese and implemented in Brazilian educational settings as part of a comparative study with Uruguay, focused on gender differences in symbolic mathematics performance during early education. Additionally, ongoing collaborations with researchers in France and El Salvador reflect growing international interest in adapting and deploying the instrument across diverse linguistic and cultural contexts. These efforts underscore the potential of PUMa to support early numeracy assessment in a wide range of educational environments.

Recommendations

Based on the findings of this study, several key recommendations are proposed for future policymakers and researchers in the field of early childhood education and assessment.

Educational systems are encouraged to adopt digital tools similar to the brief version of the PUMa test. These tools offer significant advantages, including enhanced accessibility, efficiency, and the ability to collect real-time data. Their implementation can benefit both classroom-level interventions and large-scale educational evaluations. To facilitate their integration into existing curricula, brief digital assessments like PUMa can be embedded into the diagnostic stage at the beginning of the school year. Results may be used to identify skill gaps aligned with national learning progressions, thereby allowing teachers to tailor instruction from the outset. Furthermore, digital tools support curriculum-based progress monitoring through repeated administrations, providing a dynamic link between formative assessment and instructional planning.

Sustained research is needed to further explore the structure and development of foundational mathematical skills in early childhood. Particular attention should be given to identifying the optimal combination of symbolic and non-symbolic tasks in assessments. The current study demonstrated that a balanced integration of these components yields psychometrically strong and conceptually rich evaluations. Future research should expand on this by examining how different task types contribute to the development of foundational competencies and interact with instructional strategies and developmental trajectories.

Longitudinal studies are particularly important for understanding how early numeracy skills evolve and for evaluating the predictive validity of brief assessments. The PUMa test provides a promising tool for this purpose. Future investigations could examine how early scores forecast later academic outcomes and assess the test's sensitivity to the effects of instructional interventions. Additional evidence of criterion validity would also support the early identification of students at risk for mathematical learning difficulties.

Another promising avenue lies in the exploration of innovative technologies that support self-administration. These tools not only allow young learners to engage independently with mathematical content but also enable efficient progress tracking. Research should explore how self-administered formats influence student engagement, performance accuracy, and assessment autonomy in early education settings.

Finally, the cross-cultural applicability of the PUMa test deserves further exploration. Current international collaborations highlight its potential for broader implementation, but further validation is needed across diverse linguistic and educational contexts. Understanding how the instrument performs in different systems will be key to establishing its robustness as a universal tool for assessing early mathematical skills.

By integrating these recommendations, educational systems can enhance early identification of mathematical difficulties, support timely interventions, and ultimately foster the cognitive and academic growth of young learners.

Limitations

The sample used in this study was limited to Uruguayan children, which may constrain the generalizability of the findings to populations with different cultural or educational backgrounds. Additionally, the reliance on a single dataset restricts the capacity to validate the results across diverse demographic or age groups. Future research should aim to include a wider range of populations and multiple datasets to enhance the robustness and external validity of the assessment, thereby ensuring its applicability in varied educational settings.

Although data collection through this method was particularly valuable during a challenging period, it may have introduced sampling bias. To mitigate this issue, ongoing studies are being conducted with more rigorous and randomized sampling designs. In particular, the authors are currently implementing the *National Standardized Test for the Early Assessment of Language and Mathematical Competencies: A Key for the Monitoring of Educational Trajectories*. This study employs a stratified and clustered sampling framework, which controls for the sociocultural context of schools. Such design ensures that the standardization process for the brief version of the PUMa test accounts for socioeconomic diversity and avoids systematic biases.

While both shortened versions of the test demonstrated strong psychometric properties, the parametric version was selected for implementation due to its mathematical simplicity and greater ease of interpretation. This choice aims to facilitate broader use and practical adoption of the instrument. Moreover, a longitudinal follow-up study has been initiated to trace students' academic progress over time. This research will offer further insights into the predictive validity of early numeracy skills and the long-term impact of early assessment on academic achievement.

Despite ongoing efforts to expand the test's applicability and validate its predictive utility, certain limitations stem from its digital format. Although digital administration enhances efficiency and scalability, it restricts the direct assessment of specific mathematical skills. For example, the tool does not support verbal counting, limiting its evaluation to the subskill of one-to-one correspondence. Furthermore, the lack of manipulable materials hinders the exploration of spatial and geometric reasoning, highlighting the need for complementary instruments to capture a broader spectrum of mathematical abilities.

Test validation is inherently a continuous process. While the current study presents initial evidence of the test's reliability and utility, further validity evidence is necessary—particularly regarding test-retest reliability. Assessing the temporal stability of the instrument involves re-administering the test after an appropriate interval—long enough to prevent item recall, but short enough to avoid the confounding effects of developmental changes, which are especially rapid during early childhood. The development of this brief version of the PUMa test provides a solid foundation for conducting such reliability studies in the future.

In conclusion, the present study offers evidence supporting the development and application of a digital tool that is efficient, scalable, and psychometrically sound for the assessment of early numeracy. The brief versions of the PUMa test demonstrate that it is feasible to achieve a balance between precision and feasibility without compromising validity. These findings have significant implications for educational research and practice, particularly in informing the design of personalized interventions that support children's learning trajectories from an early age.

Ethics Statements

The research protocol was reviewed and approved by the Ethics Committee of Universidad de la República. Prior to data collection, participants were fully informed about the study's purpose, procedures, and voluntary nature, and written informed consent was obtained from each school principal. They were explicitly advised of their right to withdraw at any time without facing any consequences. Participants were also assured of their anonymity, and all information collected was treated with strict confidentiality.

Acknowledgements

We would like to express our gratitude to all the participating schools for their collaboration and support throughout this study.

Funding

This project was funded by the National Agency for Research and Innovation of Uruguay (ANII) under the Sectoral Fund "Digital Inclusion: Education with New Horizons" - 2019, Grant No. FSED_2_2019_1_156716.

Generative AI Statement

As the authors of this work, we did not use generative AI in the development of this work. We take full responsibility for the content of the published manuscript.

Authorship Contribution Statement

Marconi: Data processing, analysis, results, writing; De León: Theoretical framework, writing, instrument design; Luzardo: Statistical supervision; Maiche: Theoretical supervision, instrument design, writing.

References

Archbald, D., & Farley-Ripple, E. N. (2012). Predictors of placement in lower level versus higher level high school mathematics. *The High School Journal, 96*(1), 33-51. https://doi.org/10.1353/hsj.2012.0014

Aubrey, C., Godfrey, R., & Dahl, S. (2006). Early mathematics development and later achievement: Further evidence. *Mathematics Education Research Journal*, *18*, 27-46. https://doi.org/10.1007/BF03217428

Aunio, P., & Niemivirta, M. (2010). Predicting children's mathematical performance in grade one by early numeracy. *Learning and Individual Differences*, *20*(5), 427-435. https://doi.org/10.1016/j.lindif.2010.06.003

Baker, F. B. (1985). *The basics of item response theory*. Heinemann.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). CRC Press. https://doi.org/10.1201/9781482276725

Baroody, A. J. (1987). *Children's mathematical thinking: A developmental framework for preschool, primary, and special education teachers*. Teachers College Press.

Benavides-Varela, S., Callegher, C. Z., Fagiolini, B., Leo, I., Altoè, G., & Lucangeli, D. (2020). Effectiveness of digital-based interventions for children with mathematical learning difficulties: A meta-analysis. *Computers and Education, 157*, Article 103953. https://doi.org/10.1016/j.compedu.2020.103953

Ceibal. (n.d.). *Ceibal homepage*. https://ceibal.edu.uy

Clements, D. H., & Sarama, J. (2015). Developing young children's mathematical thinking and understanding. In S. Robson, & S. Flannery Quinn (Eds.) *The Routledge international handbook of young children's thinking and understanding* (pp. 331-344). Routledge.

Collins, M. A., & Laski, E. V. (2015). Preschoolers' strategies for solving visual pattern tasks. *Early Childhood Research Quarterly, 32*, 204-214. https://doi.org/10.1016/j.ecresq.2015.04.004

Davis-Kean, P. E., Domina, T., Kuhfeld, M., Ellis, A., & Gershoff, E. T. (2022). It matters how you start: Early numeracy mastery predicts high school math course-taking and college attendance. *Infant and Child Development, 31*(2), Article e2281. https://doi.org/10.1002/icd.2281

Dellatolas, G., Von Aster, M., Willadino-Braga, L., Meier, M., & Deloche, G. (2000). Number processing and mental calculation in school children aged 7 to 10 years: A transcultural comparison. *European Child and Adolescent Psychiatry, 9*(Suppl 2), S102-S110. https://doi.org/10.1007/s007870070003

Dette, H., Neumeyer, N., & Pilz, K. F. (2006). A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli, 12*(3), 469-490. https://doi.org/10.3150/bj/1151525131

DeWind, N. K., & Brannon, E. M. (2016). Significant inter-test reliability across approximate number system assessments. *Frontiers in Psychology, 7,* Article 310. https://doi.org/10.3389/fpsyg.2016.00310

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika, 62*(1), 7-28. https://doi.org/10.1007/BF02294778

Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, *25*(3), 234-243. https://doi.org/10.1177/01466210122032046

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428-1446. https://doi.org/10.1037/0012-1649.43.6.1428

Fritz, A., Ehlert, A., Ricken, G., & Balzer, L. (2017). *Mathematik- und Rechenkonzepte bei Kindern der ersten Klassenstufe–Diagnose* [Mathematics and arithmetic concepts in first-grade children – Diagnosis]. Hogrefe. https://bit.ly/4iwAVZl

Gilmore, C., & Batchelor, S. (2021). Verbal count sequence knowledge underpins numeral order processing in children. *Acta Psychologica, 216*, Article 103294. https://doi.org/10.1016/j.actpsy.2021.103294

Ginsburg, H., & Baroody, A. J. (2007). *Tema-3: Test de competencia matemática básica* [Tema-3: Test of basic mathematical competence]. Tea & Hogrefe. https://bit.ly/3RYFCQp

Gliksman, Y., Berebbi, S., & Henik, A. (2022). Math fluency during primary school. *Brain Sciences, 12*(3), Article 371. https://doi.org/10.3390/brainsci12030371

Habermann, S., Donlan, C., Göbel, S. M., & Hulme, C. (2020). The critical role of Arabic numeral knowledge as a longitudinal predictor of arithmetic development. *Journal of Experimental Child Psychology, 193*, Article 104794. https://doi.org/10.1016/j.jecp.2019.104794

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the "Number Sense": The Approximate Number System in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology, 44*(5), 1457-1465. https://doi.org/10.1037/a0012682

Hayfield, T., & Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of statistical software*, *27*(5), 1-32. https://doi.org/10.18637/jss.v027.i05

Hutchison, J. E., Ansari, D., Zheng, S., De Jesus, S., & Lyons, I. M. (2020). The relation between subitizable symbolic and non-symbolic number processing over the kindergarten school year. *Developmental Science, 23*(2), Article e12884. https://doi.org/10.1111/desc.12884

Hyde, D. C., Khanum, S., & Spelke, E. S. (2014). Brief non-symbolic, approximate number practice enhances subsequent exact symbolic arithmetic in children. *Cognition, 131*(1), 92-107. https://doi.org/10.1016/j.cognition.2013.12.007

Koleszar, V., de León, D., Díaz-Simón, N., Fitipalde, D., Cervieri, I., & Maiche, A. (2020). Cognición numérica en Uruguay: De la clínica y los laboratorios al aula [Numerical cognition in Uruguay: From clinics and laboratories to the classroom]. *Studies in Psychology, 41*(2), 294-318. https://doi.org/10.1080/02109395.2020.1749000

Libertus, M. E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental Science, 14*(6), 1292-1300. https://doi.org/10.1111/j.1467-7687.2011.01080.x

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Routledge.

Luzardo, M. (2019). Item selection algorithms in computerized adaptive test comparison using items modeled with nonparametric isotonic model. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology* (Vol. 265, pp. 95-105). Springer. https://doi.org/10.1007/978-3-030-01310-3_9

Luzardo, M., & Rodríguez, P. (2015). A nonparametric estimator of a monotone item characteristic curve. In L. van der Ark, D. Bolt, W. C. Wang, J. Douglas, & S. M. Chow (Eds.), *Quantitative psychology research* (Vol. 140, pp. 99-108). Springer. https://doi.org/10.1007/978-3-319-19977-1_8

Lyons, I. M., & Beilock, S. L. (2011). Numerical ordering ability mediates the relation between number-sense and arithmetic competence. *Cognition, 121*(2), 256-261. https://doi.org/10.1016/j.cognition.2011.07.009

Maiche, A., González, M., Puyol, L., López, F., & De León, D. (2020). *Prueba Uruguaya de Matemática: PUMa* (Versión digital) [Software]. Registro Ley N.º 9.739 – Libro 41/835. http://puma.cicea.uy

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*(4), 713-732. https://doi.org/10.1007/s11336-005-1295-9

Miller, M. R., Rittle-Johnson, B., Loehr, A. M., & Fyfe, E. R. (2016). The influence of relational knowledge and executive function on preschoolers' repeating pattern knowledge. *Journal of Cognition and Development, 17*(1), 85-104. https://doi.org/10.1080/15248372.2015.1023307

Morsanyi, K., Peters, J., Battaglia, E., Sasanguie, D., & Reynvoet, B. (2024). The causal role of numerical and non-numerical order processing abilities in the early development of mathematics skills: Evidence from an intervention study. *Current Research in Behavioral Sciences, 6*, Article 100144. https://doi.org/10.1016/j.crbeha.2023.100144

Moura, R., Wood, G., Pinheiro-Chagas, P., Lonnemann, J., Krinzinger, H., Willmes, K., & Haase, V. G. (2013). Transcoding abilities in typical and atypical mathematics achievers: The role of working memory and procedural and lexical competencies. Journal of *Experimental Child Psychology*, *116*(3), 707-727. https://doi.org/10.1016/j.jecp.2013.07.008

Mulligan, J., Papic, M., Prescott, A. E., & Mitchelmore, M. (2006). *Improving early numeracy through a pattern and structure mathematics awareness program (PASMAP).* In *Conference of the Mathematics Education Research Group of Australasia,* (pp. 376-383). MERGA. https://bit.ly/41xIFU5

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications, 9*(1), 141-142. https://doi.org/10.1137/1109020

Odic, D., & Starr, A. (2018). An introduction to the approximate number system. *Child Development Perspectives, 12*(4), 223-229. https://doi.org/10.1111/cdep.12288

Olgan, R., Cengizoğlu, S., & Solmaz, G. (2017, April 20-23). *Preschool children's views on unsolvable math problems* [Conference presentation]. 26th International Conference on Educational Sciences, Antalya, Turkey. https://open.metu.edu.tr/handle/11511/75178

Outhwaite, L. A., Aunio, P., Leung, J. K. Y., & Van Herwegen, J. (2024). Measuring mathematical skills in early childhood: A systematic review of the psychometric properties of early maths assessments and screeners. *Educational Psychology Review, 36*, Article 110. https://doi.org/10.1007/s10648-024-09950-6

Purpura, D. J. (2010). *Informal number-related mathematics skills: An examination of the structure of and relations between these skills in preschool* (Publication No. 3462344) [Doctoral dissertation, The Florida State University]. ProQuest Dissertations Publishing. http://bit.ly/4lNlizd

Purpura, D. J., Reid, E. E., Eiland, M. D., & Baroody, A. J. (2015). Using a brief preschool early numeracy skills screener to identify young children with mathematics difficulties. *School Psychology Review, 44*(1), 41-59. https://doi.org/10.17105/SPR44-1.41-59

Rajlic, G. (2020). Visualizing items and measures: An overview and demonstration of the kernel smoothing item response theory technique. *The Quantitative Methods for Psychology*, 16(4), 363-375. https://doi.org/10.20982/tqmp.16.4.p363

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*(4), 611-630. https://doi.org/10.1007/BF02294494

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Nielsen and Lydiche.

Reigosa-Crespo, V., & Estévez-Pérez, N. (2023). Conceptual foundations of early numeracy: Evidence from infant brain data. In M. Gomides, I. Starling-Alves, F. H. Santos (Eds.), *Progress in brain research*, (Vol.282, pp. 1-15). Elsevier. https://doi.org/10.1016/bs.pbr.2023.10.002

Reise, S. P., & Moore, T. M. (2023). Item response theory. In H. Cooper, M. N. Coutanche, L. M. McMullen, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Foundations, planning, measures, and psychometrics* (2nd ed., pp. 809-835). American Psychological Association. https://doi.org/10.1037/0000318-037

Romano, E., Babchishin, L., Pagani, L. S., & Kohen, D. (2010). School readiness and later achievement: Replication and extension using a nationwide Canadian survey. *Developmental Psychology*, *46*(5), 995-1007. https://doi.org/10.1037/a0018880

Seitz, M., & Weinert, S. (2022). Numeracy skills in young children as predictors of mathematical competence. *British Journal of Developmental Psychology, 40*(2), 224-241. https://doi.org/10.1111/bjdp.12408

Singer, V., & Cuadro, A. (2014). Psychometric properties of an experimental test for the assessment of basic arithmetic calculation efficiency. *Studies in Psychology, 35*(1), 183-192. https://doi.org/10.1080/02109395.2014.893657

Szkudlarek, E., & Brannon, E. M. (2017). Does the approximate number system serve as a foundation for symbolic mathematics? *Language Learning and Development, 13*(2), 171-190. https://doi.org/10.1080/15475441.2016.1263573

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. Springer https://doi.org/10.1007/978-1-4757-2691-6

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A, 26*(4), 359-372. https://www.jstor.org/stable/25049340

Xu, X., & Douglas, J. (2006). Computerized adaptive testing under nonparametric IRT models. *Psychometrika, 71*(1), 121-137. https://doi.org/10.1007/s11336-003-1154-5

Appendices

Appendix A: Description of the task in PUMa

Ability	Instruction	Items	Task	Screen display
Approximate number system	Touch the side where there are more fireflies	32	Two sets of dots (fireflies) are shown in the task. The child must indicate which of the sets has the greatest number of elements. The points appear on the screen for a short time to avoid that the children have the possibility of counting; the ratio between the points on each side varies from 1.25 to 2.00	
Counting	You must load the number of stones that the order indicates	7	The task involves putting in the cart the amount of stones that is shown analogically on the paper containing the order (through points).The number of dots in the order varies between 2 and 23.	
Numerical ordering forward	Order the stones from smallest to largest	10	The task involves ordering from smallest to largest a set of stones that have a number symbol. The number of stones varies between 5 and 7 stones while the numbers that are presented are between 1 and 131.	
Numerical ordering backward	Order the stones from largest to smallest	10	The person being evaluated is asked to order, from highest to lowest, a set of stones that have a numerical symbol. 5 stones with non-consecutive numbers ranging from 1 to 112 are presented.	
Transcoding auditory-verbal stimuli to Arabic-visual symbols	Touch the sheep that has the number you heard	20	The task involves touching the sheep that has the same number written on it that they hear through the audio. The numbers vary from 2 to 150 and the child has the possibility to listen to the name of the number as many times as needed.	
Math fluency	Add as fast as you can the total number of animals	34	Two cards are displayed on the screen that have one digit each for the children to do the addition and indicate the result by touching one of the cards from 1 to 10 available at the bottom of the screen.	
Patterns		10	A pattern made up of a series of images is shown in which there is an empty space that must be completed from two options displayed at the bottom of the screen.	
Composition and decomposition of numbers		20	The child is asked to select the amount of money with which they will pay for the total purchase. The options are presented through unlimited coins of 1, 2, 5, 10 and 50 pesos. The totals of the purchases that the exercise proposes vary between 2 and 200 pesos.	

Appendix B: Item bank from the original PUMA test, and the brief parametric and non-parametric tests

Approximate number system Task				**Retained: Parametric Brief Test**	**Retained: Non-Parametric Brief Test**
Item	**Target 1**	**Target 2**	**Ratio**		
1	4	2	2	*	*
2	10	5	2	*	*
3	5	4	1,3		
4	10	6	1,7	*	*
5	7	5	1,4		
6	8	4	2	*	
7	10	8	1,3	*	
8	15	9	1,7	*	
9	20	16	1,3	*	*
10	14	10	1,4	*	
11	30	15	2	*	
12	20	12	1,7		
13	10	7	1,4		*
14	30	10	3	*	*
15	12	12	1		
16	40	20	2		*
17	30	21	1,4	*	*
18	50	25	2	*	*
19	40	24	1,7	*	*
20	40	28	1,4		*
21	30	24	1,3		*
22	48	40	1,2		*
23	33	20	1,7	*	*
24	28	20	1,4	*	*
25	20	14	1,4	*	
26	22	13	1,7	*	*
27	21	17	1,2		
28	20	10	2	*	
29	45	27	1,7	*	
30	6	5	1,2	*	*
31	50	35	1,4		
32	40	20	20	*	

Numerical ordering forward Task			**Retained: Parametric Brief Test**	**Retained: Non-Parametric Brief Test**
Item	**Set**	**Correct answer**		
1	5, 3, 6, 4, 2, 7, 1	1, 2, 3, 4, 5, 6, 7	*	*
2	11, 14, 16, 10, 15, 12, 13	10, 11, 12, 13, 14, 15, 16	*	*
3	31, 28, 30, 25, 26, 29, 27	25, 26, 27, 28, 29, 30, 31	*	*
4	39, 40, 43, 38, 41, 42	38, 39, 40, 41, 42, 43		
5	48, 52, 47, 49, 51, 50, 46	46, 47, 48, 49, 50, 51, 52		
6	67, 71, 70, 73, 68, 72, 69	67, 68, 69, 70, 71, 72, 73	*	
7	89, 91, 87, 88, 90	87, 88, 89, 90, 91	*	*
8	100, 99, 104, 101, 102, 103, 98	98, 99, 100, 101, 102, 103, 104		*
9	113, 112, 111, 109, 110	109, 110, 111, 112, 113	*	*
10	127, 131, 128, 129, 130	127, 128, 129, 130, 131	*	*

Counting Task			Retained: Parametric Brief Test	Retained: Non-Parametric Brief Test
Item	Set	Correct answer		
1	22	15	*	*
2	8	2	*	*
3	8	4	*	*
4	10	3	*	*
5	30	18	*	*
6	32	23		*
7	20	9	*	*

Numerical ordering backward Task			Retained: Parametric Brief Test	Retained: Non-Parametric Brief Test
Item	Set	Correct answer		
1	2, 5, 1, 3, 4	5, 4, 3, 2, 1	*	*
2	10, 6, 7, 9, 8	10, 9, 8, 7, 6	*	*
3	16, 18, 20, 17, 19	20, 19, 18, 17, 16	*	*
4	27, 30, 31, 29, 28	31, 30, 29, 28, 27		
5	39, 37, 40, 38, 41	41, 40, 39, 38, 37		
6	49, 50, 53, 51, 52	53, 52, 51, 50, 49	*	
7	58, 59, 57, 56, 60	60, 59, 58, 57, 56	*	*
8	81, 77, 80, 79, 78	81, 80, 79, 78, 77		*
9	99, 98, 101, 102, 100	102, 101, 100, 99, 98	*	*
10	111, 112, 109, 110, 108	112, 111, 110, 109, 108	*	*

Transcoding auditive-verbal stimuli to Arabic-visual symbols Task			Retained: Parametric Brief Test	Retained: Non-Parametric Brief Test
Item	Set	Correct answer		
1	2, 5, 3, 20, 10	2	*	*
2	3, 10, 9, 16, 5	5	*	*
3	10, 9, 3, 13, 5	3		
4	16, 5, 10, 9, 20	9		
5	13, 10, 20, 3, 16	10	*	*
6	26, 16, 10, 13, 20	16	*	*
7	20, 31, 43, 10, 13	13		*
8	22, 20, 10, 16, 31	20		
9	43, 31, 20, 22, 56	22	*	
10	31, 43, 56, 13, 22	31		*
11	31, 65, 43, 13, 56	43	*	*
12	43, 89, 65, 56, 16	6	*	
13	89, 75, 56, 150, 65	65		
14	56, 89, 65, 75, 111	75	*	
15	89, 127, 103, 56, 43	89	*	
16	130, 150, 43, 103, 111	103		*
17	127, 111, 103, 130, 150	111		
18	111, 150, 130, 103, 127	130		
19	117, 103, 111, 127, 130	127	*	*
20	105, 130, 150, 103, 111	150		*

Math fluency Task			Retained: Parametric Brief Test	Retained: Non-Parametric Brief Test
Item	Target	Correct answer		
1	1 + 1	2	*	*
2	3 + 2	5	*	
3	2 + 2	4		*
4	3 + 1	4		
5	2 + 1	3	*	*
6	3 + 2	5		*
7	4 + 3	7		
8	1 + 9	10		
9	4 + 2	6		
10	6 + 1	7		*
11	2 + 5	7		*
12	3 + 3	6	*	
13	4 + 1	5		
14	2 + 7	9	*	
15	1 + 5	6		*
16	5 + 2	7	*	*
17	5 + 5	10		
18	2 + 4	6	*	*
19	4 + 5	9		*
20	3 + 6	9	*	
21	4 + 4	8		*
22	6 + 4	10		
23	5 + 3	8		
24	2 + 8	10		
25	6 + 3	9		
26	7 + 3	10		
27	6 + 2	8		
28	4 + 6	10		
29	8 + 1	9		
30	3 + 4	7		
31	5 + 1	6		
32	3 + 7	10		
33	1 + 7	8		
34	8 + 2	10		

Patterns Task				Retained: Parametric Brief Test	Retained: Non-Parametric Brief Test
Item	Pattern	Options	Target		
1	1-2-1-X	2, 1	2	*	
2	3-3-4-4-X	3, 4	3		*
3	X-7-8-7-8	8, 7	8		*
4	5-6-X-6-5	6, 5	5	*	
5	9-10-11-9-X	11, 10	10		*
6	12-X-14-12-13-14	14, 13, 12	13	*	*
7	2-2-X-2-2-1	1, 2, 2009	1	*	*
8	3-X-4-5-3-3	4, 3, 2005	3		*
9	6-7-8-9-X-7	9, 8, 7, 6	6		*
10	10-X-11-11-10-12	10, 11, 12	12	*	*

Note. For further information regarding the PUMa test and its applications, please refer to the official website: https://puma.ceibal.edu.uy.