

An early numeracy digital brief assessment: parametric and non-parametric IRT models

Cecilia Marconi^a, Dinorah de León^b, Mario Luzardo^a, Alejandro Maiche^b

^a*Instituto de Fundamentos y Métodos, Facultad de Psicología, Universidad de la República, Uruguay;*

^b*Centro Interdisciplinario en Cognición para la Enseñanza y el Aprendizaje, Universidad de la República, Uruguay*

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 test comprises both symbolic and non-symbolic dimensions, including tasks 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. 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, we generated preliminary evidence that the brief version retained both precision and validity. The resulting brief test, consisting of nearly half the original items, achieved a balanced representation of symbolic and non-symbolic tasks. These findings highlight the efficiency of digital tools for real-time data collection and their potential to enable scalable early interventions. Such advancements support personalized educational strategies, fostering cognitive and academic growth from an early age.

Keywords: Early Numeracy Assessment; Symbolic/Non-symbolic Mathematics Skills; Item Response Theory; Parametric/Non-parametric IRT models; Kernel Smoothing IRT.

Introduction

Early numeracy skills play a crucial role in shaping mathematical development and lay the foundation for understanding how early competence in this area can influence 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).

While many children begin formal education with the foundational skills needed to engage with early math concepts, others face specific learning difficulties, such as dyscalculia, which can significantly hinder their mathematical success. These challenges often persist throughout their educational trajectories, leading to enduring disparities in academic performance and broader life outcomes (Davis-Kean et al., 2022). Therefore, early identification of mathematical difficulties is crucial for developing effective detection and intervention strategies that prevent the widening of these gaps. By assessing cognitive and numerical abilities before formal schooling, educators and researchers can identify at-risk children and implement targeted support to mitigate long-term academic and cognitive disparities.

Although early identification of at-risk children is essential for implementing targeted support and mitigating long-term academic and cognitive disparities, the limited availability of culturally and linguistically adapted assessment tools remains a significant challenge, particularly in Spanish-speaking countries. This issue is especially pronounced in Latin America, where many children lack access to evaluations that meet their specific needs. To address this gap, robust tools are needed to provide reliable data for both research and educational practices. Digital tools have emerged as a scalable and accessible solution, with technology-based interventions, such as interactive applications, showing promise in supporting children with mathematical learning difficulties through personalized and adaptive approaches (Benavides-Varela et al., 2020).

However, many of these tests are developed within different cultural contexts, making their applicability less relevant for certain populations. Many of these tests also require a trained

administrator to ensure proper application, adding an additional layer of complexity to their use. Moreover, very few of these tools have demonstrated concurrent validity with other previously validated tests, which limits their reliability and comparative usefulness in diverse educational settings (Outhwaite et al., 2024).

One of the most widely used tests in research worldwide is the Basic Mathematical Competence Test of Early Mathematics Ability—Third Edition (TEMA-3), a standardized tool adapted for Spanish-speaking populations aged 3 years to 8 years and 11 months. Comprising 72 items, the test evaluates both informal and formal components of mathematical

knowledge. Informal knowledge reflects mathematical concepts acquired before schooling, such as basic number sense and pre-counting abilities, while formal components assess structured tasks like counting sequences and cardinality (Ginsburg & Baroody, 2003). While the TEMA-3 is comprehensive and validated, it requires one-on-one administration by a trained professional, making it time-intensive and limiting its scalability for assessing entire populations of young children. For instance, in smaller countries such as Uruguay, each cohort consists of approximately 30,000 children, posing logistical challenges for widespread implementation.

Similarly, the Woodcock-Johnson Muñoz Applied Problems subtest (Muñoz-Sandoval, et al., 2009) is a global standard for assessing mathematical skills, including calculation fluency, applied problem-solving, and mathematical reasoning. Its subtests are brief—each lasting approximately 3 minutes—and offer flexibility in administration, which makes it widely used in both clinical and research contexts. However, like the TEMA-3, this tool also requires individual administration, restricting its capacity to assess multiple children simultaneously and increasing the demand for specialized personnel.

Purpura and Lonigan (2015) have proposed the Early Numeracy Assessment (ENA), which offers a more concise approach by focusing on 12 key numeracy tasks, including verbal counting, one-to-one counting, cardinality, set comparison, subitizing, and number combinations. Designed to align with established frameworks such as the NCTM Preschool Standards and the Common Core State Standards, the ENA provides a quick and psychometrically robust measure of early numeracy skills. However, despite its simplicity and ease of administration, the ENA shares similar limitations, as it requires direct supervision and specific materials, making it unsuitable for autonomous use.

Although in Latin America these tests are often based on norms from other populations (such as Spanish norms), recent initiatives in some countries have focused on developing locally adapted tools, as demonstrated by the case of Chile (Fritz, Ehlert, Ricken, & Balzer, 2017). In Uruguay, there is a rich tradition in mathematical cognition, which has fostered the development of various locally-designed tools for assessing math (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. Due to its focus on formal mathematical skills, TECA is recommended for use starting from second grade, making it less suitable for assessing the foundational mathematical concepts typically developed during preschool and early primary years. Despite its strengths in identifying children at risk for learning difficulties in later stages of education, TECA does not address early numeracy skills, such as number sense, counting sequences, or basic numerical relationships. There are no instruments designed to evaluate the intuitive skills that underpin formal mathematical abilities, leaving a significant gap in the ability to measure and support young children's development.

To address this need, in 2020, we start to develop the initial version of the Prueba Uruguaya de Matemática (PUMa), a tool specifically designed for the Uruguayan context that focuses on early mathematical skills (Universidad de la República, 2022). PUMa provides a culturally relevant and scalable means of assessment for preschool-aged children, bridging a critical gap in early education. This tool equips educators with a valuable resource to better understand and support early mathematical learning in Uruguay.

In this paper, we present the theoretical framework underlying the initial version of the PUMa tool (Universidad de la República, 2022) and detail the process of developing a brief version of PUMa: a shorter and more practical alternative to the original one. Using a variety of statistical methods within the framework of Item Response Theory (IRT), we provide preliminary evidence of the brief version's effectiveness and reliability in the Uruguayan educational context. This new brief version of PUMa will help bridge the gap in early-stage assessment, offering educators, researchers, and policymakers a robust resource to better understand and support young learners in Uruguay.

The PUMa Test: construct and measurement

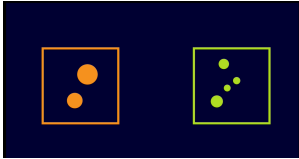



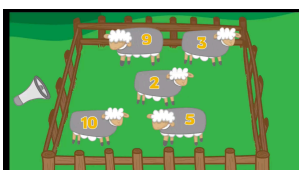
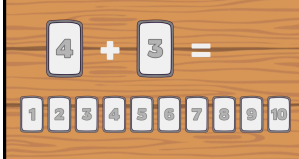
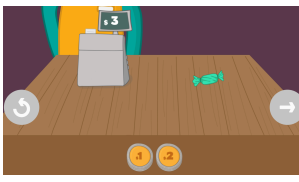
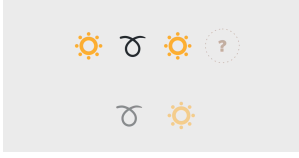
PUMa is an innovative, online, self-administered screening tool designed to evaluate early mathematical abilities in preschool and first-grade children. Children assessed by PUMa are guided through the assessment via audio instructions delivered through headphones. This approach allows the simultaneous 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. 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 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. These tasks are also 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 focus on the assessment goals.

Each task starts automatically once the child finishes the previous one since the 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. 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)	<i>Touch the side where there are more fireflies</i>	
Counting (CON)	<i>You must load the number of stones that the order indicates</i>	
Numerical ordering forward (SNP)	<i>Order the stones from smallest to largest</i>	
Numerical ordering backward (SNR)	<i>Order the stones from largest to smallest</i>	
Transcoding auditory-verbal stimuli to Arabic-visual symbols (TRA)	<i>Touch the sheep that has the number you heard</i>	
Math fluency (CMV)	<i>Add as fast as you can the total number of animals</i>	
Composition and decomposition of numbers (CYD)	<i>Choose the amount of coins to pay the exact price of the snack</i>	
Patterns (PAT)	<i>Complete the pattern by touching the missing symbol</i>	

Approximate Number System

The approximate number system (ANS) is a component of number sense that implies 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 become 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; Szklarek & Brannon, 2017). For example the acuity in ANS is positively correlated with math in school years (Libertus & Feigenson, 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).

Numerical ordering backwards

Learning to count backwards correctly takes children longer, likely because they must first master the forward sequence (Walter-Lager et al., 2021). 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 & Samara, 2014). 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 (Aunola et al., 2004).

Transcoding Auditory-verbal stimuli to arabic-visual symbols

Number transcoding is the process in which a number is spoken aloud, and children are required to represent it using symbols. This ability is a well-documented milestone in mathematical knowledge but is considered one of the most challenging to acquire. One reason for this difficulty is that children must first fully understand the base-10 system (Geary, 2000). The mental process behind transcoding involves converting information from a verbal numerical expression into a graphic 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—both forwards and backwards—are considered essential for mathematical fluency (Koponen et al., 2007; Locuniak & Jordan, 2008). Like most math skills, the development of mathematical fluency is gradual. In this process, children must first master counting, as it serves as the foundation for future calculation strategies. For example, a typical calculation strategy involves using the largest number in a sum as the starting point and then adding the smallest number (Koponen, et al., 2012).

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 (Miller et al., 2016; Collins & Laski, 2015)

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)

Method

Participants

The study sample consisted of 443 children, ages five to six, enrolled in preschool and first grade across eight private schools in Uruguay. Due to pandemic-related access constraints, a convenience sampling approach was employed, leveraging established connections with schools for participant recruitment. The sample was collected in November 2020 and November 2021. Efforts were made to include schools representing different socioeconomic backgrounds to ensure a more diverse and representative sample. 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 -value = 0.6689).

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 result.

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 assessments were conducted in classroom settings by undergraduate psychology students who had received prior training to ensure standardized test administration. The test was delivered through tablets, with each child provided with their own tablet and headphones.

For the TEMA-3 administration, the starting point was determined by the child's age-appropriate entry item, assuming no developmental disorders. Testing continued until ceiling and floor criteria were established: a ceiling was 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., Rasch, 1960; Lord, 1980; Reise et al, 2023; Baker, 1985; Baker & Kim, 2004). 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 & Cohen, 2001; Ramsay, 1991; Douglas, 1997; 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, there is currently no definitive theorem 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 of 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

identified as the ability value at which the ICC equals 0.5, while pseudo-information is derived from the first derivative of the isotonic ICC to capture information levels across the theta scale. This method not only maintains the flexibility of non-parametric estimation but also aligns with the fundamental assumptions of IRT, enhancing the robustness and interpretability of the ICC estimates.

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 $p\text{-value} < .01$. 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 have weaker correlations with other tasks., with the ANS task notably exhibiting the lowest inter-task correlations, ranging between 0.26 and 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.

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 ($p\text{-value} < 0.05$) 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 ($p\text{-value} > 0.05$) 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 ($p\text{-value} < 0.05$), 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 T_i 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 (R Core Team, 2024), 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. 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 assigning relative weights to each task within symbolic or non-symbolic domains. Following this, the specified number of items was selected within each interval based on their information quality. For the parametric approach, items with the highest discrimination parameters within each grid interval were chosen, while for the non-parametric approach, items with the highest pseudo-information were prioritized.

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. Subsequently, 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 t-test was applied to assess the significance of any differences in their correlation coefficients.

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	35.00	133.00	189.00	63.00
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	0.00
RMSEA_5	0.041	0	0.041	0	0.000	0	0	0
RMSEA_95	0.053	0.059	0.077	0.1145	0.026	0.030	0.028	0.036
SRMSR	0.104	0.061	0.055	0.173	0.115	0.052	0.181	0.071
TLI	0.868	1.000	0.923	0.998	1.003	0.996	1.000	1.000
CFI	0.877	1.000	0.957	0.999	1.000	0.997	1.000	1.000

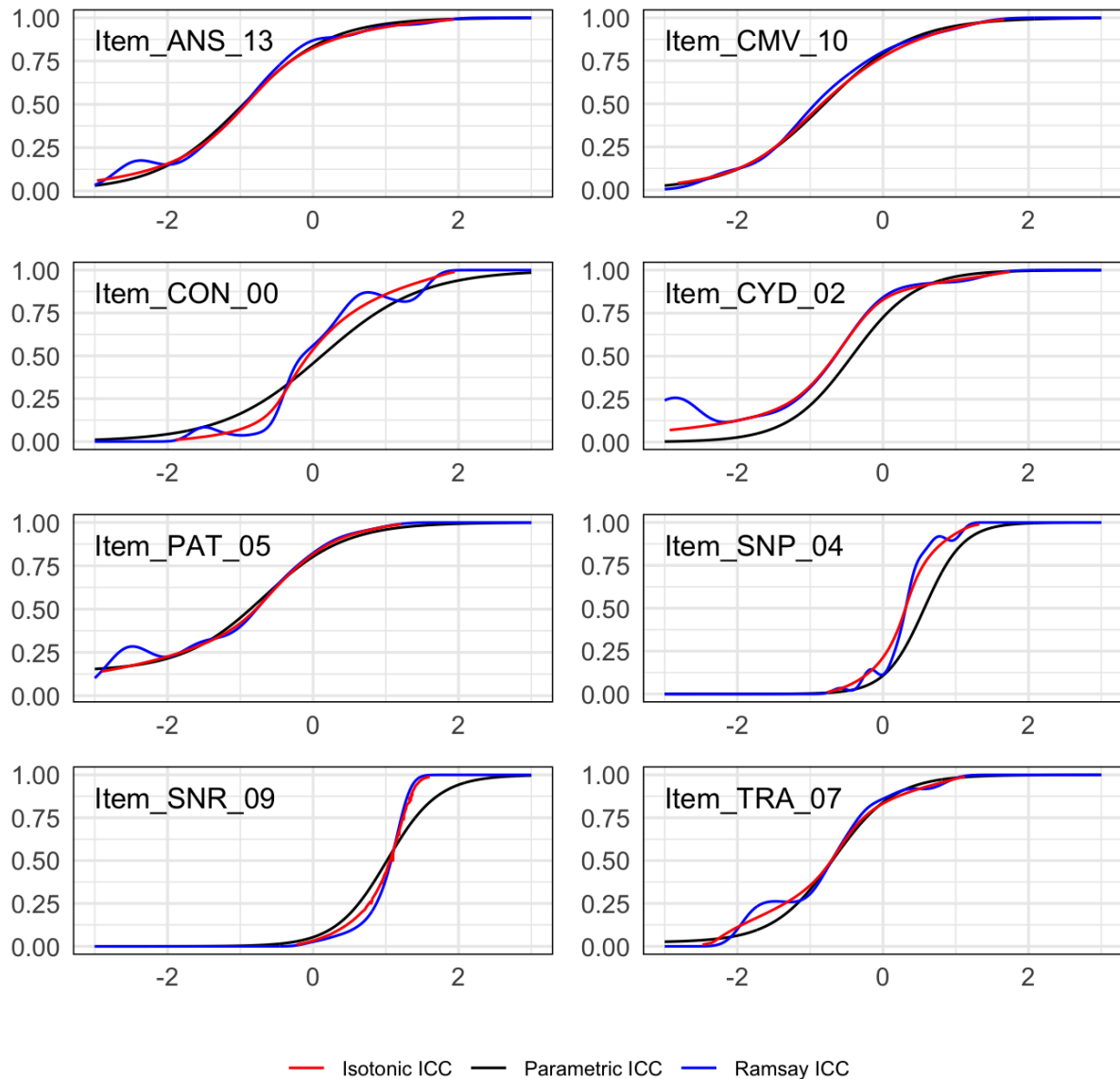


Figure 1. Comparison of Parametric and Non-parametric ICC estimation for a selection of sample items (one for each task).

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.

Applying this parametric selection process resulted in a reduction of items from 144 to 69, representing a 52% reduction in test length. It is important to note that for the Counting (CON) and Progressive Numeric Series (SNP) tasks, the selection procedure was not applied due to the limited initial number of items (six and four, respectively).

The initial version of the PUMa test exhibited a greater representation of symbolic skill items, while non-symbolic skills were less represented. The development of the brief parametric version, however, resulted in a more balanced distribution, achieving a closer alignment between non-symbolic and symbolic skills in the final item set.

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.

It is noteworthy that all correlations are positive and strong, with most exceeding 0.9, underscoring the robustness of the brief test. For tasks CON and SNP, the correlation is 1.0, as the brief test version comprises the same items as the initial version test. 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 development of the parametric brief test, the non-parametric version was created after removing items that failed to meet CTT psychometric standards or exhibited a high number of missing responses. However, unlike the parametric approach, items identified through the ICC estimation process were not eliminated beforehand. The analytical procedure followed the same structure (for details, refer to the Analytic Approach section); item selection was based on the estimated pseudo-difficulty parameter, with items containing higher pseudo-information selected within predetermined grid intervals. Table 6 provides an overview of items removed and retained through this process. Ultimately, the non-parametric approach, using isotonic ICC estimates, resulted in a brief test of 73 items—a reduction of 49%, closely aligning with the reduction achieved by the parametric method. Regarding the distribution of non-symbolic and symbolic items, the final proportions reached 44% and 56%, respectively, reflecting the balanced distribution achieved with the parametric approach

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	Removed: high level Nas	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%

Following the same analytical procedure used in the parametric approach, ability estimates for the non-parametric brief test were calculated using both IRT and CTT, and correlations were subsequently computed. Table 7 presents these correlations, showing that, consistent with the parametric approach, the correlations are predominantly positive and strong (exceeding 0.9) in nearly all cases, with the exception of CYD, which yielded a correlation of 0.884. As previously noted, the CON and SNP tasks exhibit a correlation of 1.0, as a brief version was not developed for these tasks due to the limited number of items

Table 7. Correlation Coefficients Using the Non-Parametric Approach

Correlations: Abilities estimated from the initial and the brief test versions under IRT NonParametric 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

Considering these results, the non-parametric approach proves as effective as the parametric method in developing a psychometrically sound brief test. Both approaches achieved a

similar reduction in test length, approximately 50%, while maintaining high correlations with ability estimates derived 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, they both evaluate closely related constructs and 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	Number of items	
		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). 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 for the psychometric validity of the initial digital version of PUMa test, while addressing the limitations posed by its extensive 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.edu.uy), 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, fostering 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 arrived at 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 the two brief test versions showed moderate to high positive correlations with TEMA-3 scores, a benchmark tool in early numeracy assessment. Consistent with previous studies, symbolic skills showed stronger correlations with TEMA-3 scores, while non-symbolic skills exhibited weaker associations (Purpura & Lonigan, 2015). These findings indicate that it is possible to develop more efficient assessment tools without compromising accuracy, addressing frequent concerns about the reliability and predictive validity of existing screening tools.

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. Furthermore, the brief format enables educators to support repeated test administrations over time, facilitating continuous progress monitoring without inducing fatigue or disengagement, thus enhancing the tool's practicality for both classroom and research settings.

However, several limitations must be acknowledged. The sample was limited to Uruguayan children, which may restrict the generalizability of the findings to other populations with different cultural or educational contexts. Furthermore, the reliance on a single dataset constrains the validation of findings across diverse contexts or age groups. Expanding this research to include varied populations and additional datasets would enhance the robustness and applicability of the test, ensuring it can be effectively used in broader educational settings.

To address these limitations, the brief PUMa test developed under the parametric approach is being administered to representative samples of Uruguayan schools. Although the study demonstrated the effectiveness of both brief versions, the parametric approach was chosen for implementation due to their mathematical simplicity and interpretability. This effort aims to support the test's standardization and broaden its potential for widespread use. Additionally, a longitudinal study has been initiated to follow children throughout their academic journeys. This will provide valuable insights into how early mathematical skills predict future academic

achievement and validate the effectiveness of the test in forecasting long-term mathematical performance.

While efforts are underway to address the limitations of the current study, such as expanding the test's applicability through broader samples and longitudinal studies, additional considerations arise due to PUMA's digital format. For instance, although the tool offers significant advantages in efficiency and scalability, it presents challenges in assessing certain mathematical skills. Specifically, the digital format does not support direct evaluation of counting through verbal responses, prompting us to focus on a subskill: one-to-one correspondence. Furthermore, the inability to manipulate concrete objects limits the exploration of geometric and spatial concepts, emphasizing the need for complementary approaches to achieve a more comprehensive assessment of mathematical abilities.

In summary, this study provides evidence for the development and application of an efficient, reliable, and scalable digital tool for assessing early numeracy skills in young learners. The brief versions of the PUMa test demonstrate that it is possible to balance precision and practicality without compromising psychometric validity. These advancements contribute not only to the field of educational research but also to the development of personalized interventions that can better support children's learning trajectories.

References

- Aubrey, C., Godfrey, R., & Dahl, S. (2006). Early mathematics development and later achievement: Further evidence. *Mathematics Education Research Journal*, *18*(1), 27–46. <https://doi.org/10.1007/BF03217516>
- 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>
- Aunio, P., Korhonen, J., Ragpot, L., Törmänen, M., & Henning, E. (2021). An early numeracy intervention for first-graders at risk for mathematical learning difficulties. *Early Childhood Research Quarterly*, *55*, 252–262. <https://doi.org/10.1016/j.ecresq.2020.07.003>
- Archbald, D., & Farley-Ripple, E. N. (2012). Predictors of placement in lower level versus higher level high school mathematics. *The High School Journal*, 33–51.
- Baker, F. B. (1985). *The basics of item response theory*. Portsmouth, NH: Heinemann.

- Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). CRC Press. <https://doi.org/10.1201/9781482276725>
- 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 & Education, 157*, 103953. <https://doi.org/10.1016/j.compedu.2020.103953>
- Clements, D. H., & Sarama, J. (2014). The importance of the early years. In R. E. Slavin (Ed.), *Science, technology & mathematics (STEM)* (pp. 5–9). Corwin Press. <https://doi.org/10.4135/9781483380957.n2>
- 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.2015.04.003>
- 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), 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 & Adolescent Psychiatry, 9*(S2), S102–S110. <https://doi.org/10.1007/s007870070021>
- Dette, H., Neumeyer, N., & Pilz, K. (2006). A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli, 12*(3), 469–490. <https://doi.org/10.3150/bj/1151525131>
- Douglas, J. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement, 25*(3), 234–243. <https://doi.org/10.1177/01466210122032046>
- Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika, 62*, 7–28. <https://doi.org/10.1007/BF02294778>
- Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., & 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 (MARKO-DI+)*. Göttingen: Hogrefe.
- Geary, D. C. (2000). From infancy to adulthood: The development of numerical abilities. *European Child & Adolescent Psychiatry, 9*(S2), S11–S16. <https://doi.org/10.1007/s007870070019>

- Ginsburg, P., & Baroody, J. (2007). *TEMA 3: Test de competencia matemática básica: Manual*. ISBN: 9788471748645.
- 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/a0012617>
- Hayfield, T., & Racine, J. S. (2008). The np package: Nonparametric and semiparametric kernel methods in R. ETH Zürich; McMaster University. Recuperado de <http://CRAN.R-project.org/package=np>
- 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), 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). Numerical cognition in Uruguay: From clinics and laboratories to the classroom (Cognición numérica en Uruguay: De la clínica y los laboratorios al aula). *Studies in Psychology*, *41*(2), 294–318. <https://doi.org/10.1080/02109395.2020.1727345>
- Koponen, T., Aunola, K., Ahonen, T., & Nurmi, J. E. (2007). Cognitive predictors of single-digit and procedural calculation skills and their covariation with reading skill. *Journal of Experimental Child Psychology*, *97*(3), 220–241. <https://doi.org/10.1016/j.jecp.2007.06.003>
- Koponen, T., Salmi, P., Eklund, K., & Aro, T. (2013). Counting and RAN: Predictors of arithmetic calculation and reading fluency. *Journal of Educational Psychology*, *105*(1), 162–174. <https://doi.org/10.1037/a0030396>
- Libertus, M. E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental Science*, *14*(6), 1292–1304. <https://doi.org/10.1111/j.1467-7687.2011.01062.x>
- Lord, F. M. (1980a). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lyons, I. M., & Beilock, S. L. (2011). Numerical ordering ability mediates the relation between number-sense and arithmetic competence. *Cognition*, *121*(2), 256–271. <https://doi.org/10.1016/j.cognition.2011.06.007>
- Maiche, A., De León, D., Díaz-Simón, N., Puyol, L., San Román, N., Dutra, M., & González, M. (2022). *Prueba Uruguaya de matemática: Manual de aplicación*.

- Maiche, A., de León, D., Puyol, L., Díaz-Simón, N., López, F., & San Román, N. (2022). Prueba Uruguaya de Matemáticas: PUMa (Versión 1.0.10) [Software] Universidad de la República. <https://puma.cicea.uy/es/puma/demo/>
- 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>
- Mulligan, J., Oslington, G., & English, L. (2020). Supporting early mathematical development through a ‘pattern and structure’ intervention program. *ZDM*, *52*(4), 663–676. <https://doi.org/10.1007/s11858-020-01111-3>
- Mulligan, J., Papic, M., Prescott, A. E., & Mitchelmore, M. (2006). Improving early numeracy through a pattern and structure mathematics awareness program (PASMMap). In *Conference of the Mathematics Education Research Group of Australasia (MERGA)*.
- 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.12272>
- Purpura, D. J., & Lonigan, C. J. (2015). Early numeracy assessment: The development of the preschool early numeracy scales. *Early Education and Development*, *26*(2), 286–313. <https://doi.org/10.1080/10409289.2015.991084>
- Purpura, D. J. (2010). *Informal number-related mathematics skills: An examination of the structure of and relations between these skills in preschool* (Doctoral dissertation, The Florida State University).
- Rajlic, G. (2020). Visualizing items and measures: An overview and demonstration of the Kernel Smoothing item response theory technique. *Quantitative Psychology and Measurement*, *5*(1), 28–46. <https://doi.org/10.22140/qpm.2020.0013>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.
- Räsänen, P., Salminen, J., & Niemi, P. (2009). Development of early numeracy skills. *Learning and Instruction*, *19*(2), 149–159. <https://doi.org/10.1016/j.learninstruc.2008.01.002>
- 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>

- Siegler, R. S., & Ramani, G. B. (2009). Playing linear number board games—But not circular ones—Improves low-income preschoolers' numerical understanding. *Journal of Educational Psychology, 101*(3), 545–560. <https://doi.org/10.1037/a0014239>
- Siegler, R. S., & Lortie-Forgues, H. (2014). Early arithmetic development: New questions and insights. *Current Directions in Psychological Science, 23*(2), 88–93. <https://doi.org/10.1177/0963721414522813>
- Sian, M., Paredes, M., & Vezina, J. (2019). Contributions of number line estimation and counting to young children's mathematics achievement. *Journal of Educational Psychology, 111*(4), 779–792. <https://doi.org/10.1037/edu0000345>
- Trivett, S., & Thomas, M. O. J. (2017). Early development of arithmetic skills. *Journal of Educational Psychology, 109*(2), 299–312. <https://doi.org/10.1037/edu0000154>
- 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://doi.org/10.1007/BF02868620>
- 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>