

# Assessing Early Math Skills in Preschoolers by Using Digital Games

Jairo A. Navarrete-Ulloa<sup>1</sup> , David M. Gómez<sup>1,2</sup> , Llery Ponce<sup>3</sup> , Felipe Munoz-Rubke<sup>4</sup> , Pablo R. Dartnell<sup>3,5</sup> 

[1] *Institute of Education Sciences, Universidad de O'Higgins, Rancagua, Chile.* [2] *Millennium Nucleus for the Study of the Development of Early Mathematics Skills (MEMAT), Santiago, Chile.* [3] *Institute for Advanced Studies in Education, Universidad de Chile, Santiago, Chile.* [4] *Instituto de Psicología, Universidad Austral de Chile, Puerto Montt, Chile.* [5] *Center for Mathematical Modeling, Universidad de Chile, Santiago, Chile.*

Journal of Numerical Cognition, 2025, Vol. 11, Article e14249, <https://doi.org/10.5964/jnc.14249>

Received: 2024-04-10 • Accepted: 2025-01-17 • Published (VoR): 2025-04-08

Handling Editor: Geetha Ramani, University of Maryland, College Park, MD, United States

Corresponding Author: Jairo A. Navarrete-Ulloa, Av. Libertador Bernardo O'Higgins 611, of. 717, Rancagua, Chile. E-mail: [jairo.navarrete@uoh.cl](mailto:jairo.navarrete@uoh.cl)

Supplementary Materials: Code, Data, Materials [see [Index of Supplementary Materials](#)]



## Abstract

Improving early mathematical competence is a major priority worldwide; thus, assessing early math abilities is critical. Although various international standardized instruments serve this purpose, their usage in underdeveloped countries is prohibitive due to their resource-intensive requirements. In this report, we explore the development of the “Test de Pensamiento Matemático” (TPM, Test of Preschool Mathematics), which is an automated, game-based, digital instrument for assessing early math abilities in 4-to-6-year-old children in accordance with international curricular standards. A confirmatory factor analysis shows an optimal fit for two dimensions: numerical thinking and visuospatial reasoning. By drawing on technology, the TPM can be applied to large groups of children, so it becomes an efficient tool for assessing performance, monitoring learning improvements, and screening children who need additional support to develop their math abilities at the same pace with their peers.

## Keywords

educational assessment, mathematics education, mathematics tests, early childhood education, educational games

## Non-Technical Summary

### Background

Teaching young children math skills is important all over the world, but many countries, especially those with fewer resources, struggle to find good ways to test these skills. Although various international standardized instruments serve this purpose, their usage in underdeveloped countries is prohibitive due to their resource-intensive requirements.

### Why was this study done?

This study introduces and validates a new tool called the Test de Pensamiento Matemático (TPM), or Test of Preschool Mathematics. It's a digital, game-like test designed to measure math skills in kids aged 4 to 6.

### What did the researchers do and find?

The test focuses on two main areas: number skills and visual-spatial reasoning (like understanding space and patterns). The TPM is easy to use and can test many children at once, making it a practical option for schools and communities with limited resources.



**What do these findings mean?**

The assessment helps teachers see how well kids are learning math, track their progress, and identify children who might need extra help to keep up with their classmates. By using technology, the TPM makes math assessment more accessible and efficient for everyone.

**Highlights**

- The TPM (Test of Preschool Mathematics) is an innovative assessment tool designed to evaluate early math skills in children at pre-K and kindergarten levels.
- It focuses on two key dimensions of learning: numerical thinking and spatial reasoning, providing a comprehensive understanding of a child's foundational math abilities.
- As a game-based assessment, the TPM is engaging and user-friendly, allowing large groups of children to participate simultaneously using tablets or smartphones paired with headphones.
- The entire assessment process is efficient, taking approximately 90 minutes to evaluate large groups of children, making it a practical solution for educators and researchers working with young learners.

Developing early math skills is crucial for students' learning and development throughout their school and post-school trajectory. Recent studies reveal that students who begin elementary school with underdeveloped math skills demonstrate low academic performance throughout their school trajectory. More specifically, early math skill levels at age five predict the student's likelihood to pursue a college degree and directly relate to the curriculum the student develops during high school (Davis-Kean et al., 2022; Watts et al., 2018). Early math skills' progress between 4 and 6 years old is the strongest predictor of math performance afterward (Siegler et al., 2012; Watts et al., 2014, 2018). Poor development of math skills is associated with early dropout from educational opportunities, lack of productive skills, sporadic employment, long periods of unemployment, low pay, and few opportunities for career advancement (Bynner & Parsons, 1997; Parsons & Bynner, 2005). Consequently, the importance of monitoring the development of early math skills cannot be overstated.

The literature on early math development has proposed a variety of assessment instruments: For example, the Clinical Interview Method and the Birthday Party (CIM; BP; Ginsburg et al., 2016; Ginsburg & Pappas, 2016), the Early Grade Mathematics Assessment (EGMA; Platas, Ketterlin-Geller, & Sitabkhan, 2016), the Test of Early Mathematics Ability (TEMA-3; Hoffman & Grialou, 2005), and the Research-based Early Math Assessment (REMA; Clements et al., 2008; Dong et al., 2021). These instruments generally consider assessing numbers, their operations, and spatial topics such as shapes, space problems, and visual patterns (see Table 1). A recent study has systematically reviewed 59 tools for measuring math competence (Outhwaite et al., 2024). The authors identified 37 math assessments (1-14 years) and 22 screeners (3-14 years). Of these tools, 52 are child-direct measures that require individual presentation, whereas only seven of them can be applied to small groups of children. Moreover, 49 require a trained assessor, and 41 are paper-based. These features suggest that children's lack of autonomy makes these assessment tools costly in terms of the amount of labor needed to evaluate large groups of children. For example, the full version of the REMA assessment requires around sixty minutes to assess one student, meaning that evaluating a classroom with 30 children would take around thirty hours of professional work (see Table 1). Note that, from the 59 measurement tools reviewed by Outhwaite et al., 55 target number skills and 47 target arithmetic skills, whereas only 22 measure shape, space, and measure skills. More importantly, only nine assessments and two screeners were evaluated in countries, cultures, or language groups different from WEIRD societies and English-speakers (Outhwaite et al., 2024).

The aforementioned instruments are frequently used in research institutions and educational systems of resourceful educational communities. In contrast, they are barely used in less affluent backgrounds due to their high demands on time and resources. For example, among the instruments listed in Table 1, the shortest application requires 20 minutes per child, so evaluating a classroom with 30 participants would need more than ten hours of continuous professional labor. Under-resourced communities often lack resources and devote little time to performing math activities (e.g.,

Strasser et al., 2009). Consequently, applying these assessments in such communities becomes prohibitive due the high amount of labor that compensates for childrens' lack of autonomy that is associated with their early age.

**Table 1**

*Comparison of Different Instruments for Assessing Early Math Skills*

Features	CIM	BP	EGMA	TEMA-3	REMA-Full	REMA-SF	TPM
Number and Operations	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Shapes	No	Yes	No	No	Yes	Yes	No
Space	No	Yes	No	No	Yes	Yes	Yes
Patterns	No	Yes	No	No	Yes	Yes	Yes
Individual application time (one child)	40m	20m	20m	40m	60m	25m	50m
Group application time (30 children)	20h	10h	10h	20h	30h	12.5h	1.5h

For example, in Chile, prior studies have estimated that the average time devoted to performing math activities at kindergarten is around 30 minutes per week (Strasser et al., 2009), suggesting that time available specifically for math assessment is even lower. Consequently, in under-resourced backgrounds, the best alternative available is the application of systematic observation methods, documentation processes, and performance rubrics for assessment (Alsina, 2021). These assessment methods rely on observing children's behavior, taking notes, and registering children's performance in math activities during an extended period; they are also widely used because they are more affordable and flexible, and they provide practical assessment results. However, one of their drawbacks is that they tend to be subjective—they rely on the expert gaze of the teacher. In principle, this might not be an issue, but these results may not be appropriate for providing performance comparisons between classrooms or schools. Furthermore, the results of these observational methods are accessible only after the observation period, so the collected information might not be helpful in providing valuable immediate feedback.

Additionally, despite the importance of developing early math skills, there is little information regarding their developmental trajectory in early education and the potential differences between boys and girls that might appear at these levels. For example, a study in Chile used data from older students and growth models to predict the existence of a gap between boys and girls in early math performance with a size around 0.15 *SD* in classrooms of 4-6 year-old children (Perez Mejias et al., 2021). It has been argued that these differences in math performance may arise due to math-gender stereotypes (Zhu & Chiu, 2019) or sex differences in preferred strategies to approach mathematical problems (Spelke, 2005), among other factors. In consequence, the understanding of these sex gaps should be approached from a multicausal perspective (Casey & Ganley, 2021) and the development of an assessment instrument should be aware of these concerns. A relevant question is thus whether instruments' factor structure is similar for boys and girls, a psychometric property known as measurement invariance.

## A Technological Solution to Math Assessment in Early Education

The solution proposed here uses technology to automate some components of the assessment process. The previous section highlighted that the critical challenge is children's lack of autonomy, making assessment expensive due to the required mediation or subjective due to the teacher-dependent observation process. Hence, an assessment solution should be designed that can overcome children's lack of autonomy and remain affordable and objective. The core idea presented here is to automate the individual mediation processes through mobile devices and gamification to allow group applications of the assessment (all children in a classroom) in a simple, automated, and affordable way. In this regard, tactile mobile devices such as tablets and smartphones can facilitate children's autonomy (Holloway et al., 2013). These devices can provide verbal instructions, represent abstract ideas in interactive visual models, or allow manipulation of virtual objects through touch screens, thus automating specific evaluation processes.

Gamification could also increase the attention span of early education students (as early ages are associated with short attention spans). In fact, educational interest in gamification has grown exponentially due to its ability to engage

students innovatively and increase their motivation and creativity (Zainuddin et al., 2020). Although using electronic devices is not recommended at early ages due to the increase of emotional problems such as anxiety and social isolation, these adverse effects are associated with the intensive use of these devices during prolonged periods (Lin et al., 2020). The solution proposed here considers an adequate use of these devices by applying them only occasionally.

The present work aims to develop a proof of concept for an automated assessment tool that facilitates the quick and straightforward acquisition of high-quality information about children's early math abilities by allowing group assessments. This report presents a proof of concept of the "Test of Preschool Mathematics" (TPM), which enables group assessments by automating the mediation process through mobile devices and gamification. Its design aims to assess mathematical skills in prekindergarten and kindergarten levels in alignment with international curricular standards, and it uses digital games to present assessment tasks that evaluate numerical or visuospatial reasoning. We hypothesize that the data collected through this group assessment strategy would be highly informative about children's abilities related to early math skills. In this regard, we will present preliminary evidence about the validity of the TPM by collecting and analyzing data from more than 700 children in prekindergarten, kindergarten, and first grade levels.

The organization of this work is as follows. The next section provides an overview of international curricular standards and their relationship with Chile's curriculum—the background of the present work. Additionally, it briefly describes the expected mathematical learning for children in prekindergarten and kindergarten classrooms. The third section presents a brief description of the TPM development process. The methods section describes the data collection, the sample, and the participants, while the results section presents the data analysis supporting the TPM's criterion and construct validity. Finally, the last section discusses the results and argues that the TPM is an assessment tool that combines efficient resource use and good information quality regarding math performance in early education. The discussion highlights that this tool could present opportunities to assess early childhood math skills and develop timely and appropriate strategies to strengthen under-resourced educational communities.

## International Standards and the Chilean Curricular Framework

Developing early numeracy is a powerful predictor of future math performance (Watts et al., 2014, 2017, 2018), mainly because it lies at the foundation of mathematical structure, allowing us to describe generalizations of predictable sequences. Consequently, many countries have emphasized opportunities to strengthen students' mathematical learning in general and also in early childhood (Clements & Sarama, 2011; Melhuish & Petrogiannis, 2006). Indeed, progress has been made in consolidating robust curriculum standards for mathematics teaching at the international level. One of the curricular frameworks of reference is the Curriculum Focal Points (NCTM, 2006), which drove the emergence of collaborative environments promoting the creation and development of high-quality tools such as assessment instruments, curriculum frameworks, and instructional materials (e.g., books and software). For example, by drawing on the Focal Points curriculum, the Common Core State Standards (CCSS, 2010) have been massively adopted in different U.S. territories, creating a common framework that facilitates and promotes collaboration and curriculum development for a coherent educational system.

A comparative analysis between the Curriculum Focal Points and the Common Core State Standards concludes that, although the two documents differ in terms of their levels of specificity, both standards are highly comparable in their coherence, focus, and content (Achieve, 2010). According to the NCTM website, comparative analyses between the Canadian mathematics curriculum and the Focal Points are highly similar. Regarding math education in prekindergarten and kindergarten classrooms, these documents state that the majority of instructional time should focus on (1) representing, relating, and operating with whole numbers and (2) describing shapes and spatial relationships (CCSS, 2010).

In coherence with these standards, Chile's curriculum (MINEDUC, 2018) strongly emphasizes the development of number knowledge and aims to develop reasoning skills about spatial relationships. A detailed description of these learning goals is provided as supplementary material (see Navarrete-Ulloa et al., 2025S-b, Table S2). Chile's curriculum aims to teach students numerical skills, using numbers to represent quantities and solve quantitative problems. The learning objectives include comparison relationships; the notion of number, quantifier function, and numerical sequence; concrete, pictorial, and symbolic (COPISI) numerical representations of quantities; and solving problems

through addition and subtraction operations using concrete and pictorial representations (up to 10). Regarding spatial relationships, the learning objectives include the creation of two- and three-element patterns, identification/classification of shapes and objects by their attributes, seriation of objects by various attributes (such as height, length, and capacity), and spatial/temporal orientation and representation of objects from different perspectives (top, bottom, sides).

## Design of the Test of Preschool Mathematics (TPM)

We reviewed the literature for tasks to assess early mathematical skills. Experts analyzed each task to verify its alignment with the Chilean early education's learning objectives (MINEDUC, 2018). Not all target learning objectives could be covered, so we designed specific tasks to cover them by drawing on theoretical frameworks of analogical reasoning and analogical representations (Kalra & Richland, 2022; Navarrete & Dartnell, 2017; Navarrete et al., 2018; Navarrete-Ulloa & Munoz-Rubke, 2022; Ramani et al., 2012, 2020; Richland et al., 2004). Analogical reasoning consists in comparing the structure of two entities and is particularly useful for processing abstract domains through concrete representations (e.g., love is a journey) by identifying a structural correspondence between a known concrete domain and the abstract concept (Andrews et al., 2006; Halford et al., 2010; Holyoak, 2012; Lakoff, 2006). For example, a balanced scale can be a friendly representation of some first-degree equations (Araya et al., 2010). This property of analogical reasoning helps to design concrete representations of abstract and complex concepts (Gentner, 2010; Gentner & Colhoun, 2010; Navarrete & Dartnell, 2017; Navarrete-Ulloa & Munoz-Rubke, 2022; Richland & McDonough, 2010; Richland et al., 2012).

The methods described above led to sixteen proposals for digital tasks to measure math abilities in children. These proposals consisted in a conceptual model of the task with focus on assessing the target learning goal along with a basic design of the user interface and its functionality. Each proposal was presented by a member of the team and evaluated by five experts who used a scoring rubric built on the following criteria: (a) factibility of development, (b) face validity based on literature support, and (c) intuitiveness of the user interface. From these proposals, twelve were selected for further development. The five experts conducted three cycles of review, evaluation, and feedback to develop these tasks further and improve them. Three of these experts had a doctoral degree and research experience, whereas the two other experts had a master's degree and preschool classroom experience. These experts then selected the best nine tasks based on the same criteria described above. Afterward, the development team implemented these nine tasks as digital games. To explore the usability and validity of the assessment, we carried out a piloting process where a group of children performed all nine digital games. One task was eliminated because its average duration was deemed too long for children to complete in actual testing. Another task was eliminated because its user interface was not intuitive enough and difficult to use, and thus, its validity was called into question.





The seven remaining tasks compose the TPM and cover nine out of the twelve learning goals of the Chilean curriculum in early education (see Navarrete-Ulloa et al., 2025S-b, Table S2). Furthermore, these seven tasks are strongly aligned with international curriculums such as the Focus Points (NCTM) and the Common Core State Standards (CCSS), as these tasks focus on (1) representing, relating, and operating with whole numbers and (2) describing shapes and spatial relationships (CCSS, 2010).



Table 2 briefly describes the seven TPM tasks, which, through narrative, illustrations, and animations, emulate digital games. The graphic and interaction design of each task had an edutainment focus implemented by an illustrator and a programmer, both experts in designing and implementing games for children. To lower children's anxiety, there was a progress indicator on the screen; the character's voice was calm and soft, and a ludic story surrounded each game, where the characters had to achieve a critical mission.



Table 2

Description of the Seven TPM's Tasks, Grouped Into Numerical Thinking and Visuospatial Reasoning

Game Name	Code	Description	Illustration	Scoring
<b>Numerical Thinking</b>				
Numerical Comparison (10 items)	NumComp	Evaluates the use of numbers to compare quantities up to 20 in an everyday situation. Two barrels are shown, each labeled with a numeral from 1 to 20. The instructions ask to choose the barrel with the larger (or smaller) quantity associated with it.		Score from 0 to 100 points. Two possible alternatives: correct or incorrect. Scores in {0, 100}.
Concrete Pictorial Symbolic Mapping (10 items)	CoPiSi	Evaluates skills to represent numbers and quantities up to 10 in a concrete, pictorial, and symbolic way. Three areas are shown on the main screen. The first presents quantities using the fingers of the hand (concrete), the second presents a collection of spaceships (pictorial), and the third presents numerals (symbolic). The instructions for this game indicate a specific quantity and ask to represent it in the three areas of the screen.		Score from 0 to 100 points. Proportional to the number of correct chosen representations. Scores in {0, 33, 50, 66, 100}
Additive Problem Solving (10 items)	AddProb	Evaluates the resolution of simple problems in a concrete and pictorial way by adding or removing up to 10 items. A character requiring an "a" amount of fruit is shown next to a basket with a "b" amount of fruit. The instructions ask to add or remove fruits from the basket to satisfy the requirement. For this, the player must choose an alternative (e.g., +2, -1, +3, etc.) from those shown in the bar presented on the screen.		Score from 0 to 100 points. Two possible alternatives: correct or incorrect. Scores in {0, 100}.
Number Line Estimation (10 items)	NumLine	Evaluates the ability to use numbers to indicate the position of elements in everyday situations. A horizontal line is shown crossing the screen joining two islands, the left one with the number 0 and the right one labeled with the number 10. At the center of the screen a numeral is shown (between 0 and 10) and the instructions ask to estimate the position of a ship at the position on the number line corresponding to that number.		$PAE = 10^* Estimate - Position $ Score = $100 - PAE$ Scores in [10, 100]
<b>Visuospatial Reasoning</b>				
Pattern Creation (10 items)	PattCrea	Evaluates the ability to copy, extend and create visual patterns of two or three elements. The left side of the screen shows a building constructed with floors whose colors follow a pattern, and the right side of the screen shows an incomplete building. The instructions ask you to complete the construction of the building on the right following the same pattern as the building on the left-although the colors may vary.		Score from 0 to 100 points. Proportional to the number of correctly repeated patterns. Scores in {0, 33, 50, 66, 100}

Game Name	Code	Description	Illustration	Scoring
Attribute Seriation (10 items)	AttribSer	Evaluates the ability to order different elements using attributes of height, width, length or capacity to contain. A train is shown that has six cars to place objects. The instructions ask to take objects from the bar presented on the screen and order them from left to right according to one of the attributes mentioned above.		Score from 0 to 100 points. Proportional to the number of correctly ordered items. Scores in {0, 33, 50, 66, 100}
Perspective Visualization (10 items)	PerspVis	It evaluates skills to represent objects from above, from the side or from below. The game shows a character in the center of the screen while taking a picture of an object from a specific perspective (front, side, top, bottom). The game asks the player to identify the photograph thus obtained from among other alternatives that show the same object, but seen from other perspectives (front, side, top, bottom).		Score from 0 to 100 points. Two possible alternatives: correct or incorrect. Scores in {0, 100}.

From now on, we will refer to each task of the TPM as a game. Each “game” (task) consists of “ten levels” (ten items) scored from 0 to 100. In other words, the TPM presents ten levels for each game, but each level differs due to slight modifications to modulate the difficulty level (see Navarrete-Ulloa et al., 2025S-b, Table S2). The TPM has seven games, each one with ten levels, meaning that the TPM presents 70 items to children. The TPM does not give feedback to children about the correctness of the answered item, but congratulates them each time they complete a game. The structure of every game has initial directions (given by a character), an example (accompanied by an animation showing the interaction with the screen), and the ten game levels (the task items). The Supplementary Materials contain the audio transcript (translated from Spanish) of the “Numerical Comparison” game to illustrate the types of questions posed to children (see Navarrete-Ulloa et al., 2025S-b, Table S3). Additionally, the ten game levels had a progression of difficulty: The first was the easiest, and the last was the most difficult. Nevertheless, every child had to answer all assessment items, since there was no algorithm for selecting the items presented to children according to their previous performance.

## The Present Study

This paper argues for the TPM’s validity based on the analysis of the data collected from a sample of Chilean children aged 4 to 6 years old in April 2022. This data collection was conducted just after the COVID-19 quarantine. It raised concerns that the data could reflect atypical learning progress, as participants had been isolated in their homes for almost two full years. For this reason, we decided to conduct a second round of data collection after roughly three months of regular school attendance. The data collected in this second round was thus expected to better reflect typical learning progress for children of these ages.

## Method

### Participants and Procedure

Measurements of children’s mathematical learning were conducted using the TPM in a sample of 824 participants in Prekindergarten (PK), Kindergarten (K), and Grade 1 (G1) classrooms, recruited from five educational centers located in O’Higgins Region, Chile. In the first and second rounds of data collection, 718 and 653 children participated, respectively. In what follows, we report the results of the second round of data collection and provide analog results for the first round in the Supplementary Materials (see Navarrete-Ulloa et al., 2025S-b). Participants’ mean ages and percentage of girls for each educational level are presented in Table S9 (first round) and Table 3 (second round). The

protocols of this research were approved by the Scientific Ethics Committee of Universidad de O'Higgins, certificate 05/2019, and the guardian of each participant signed an informed consent. Each participant also gave informed assent before the start of the activities.

**Table 3**

*Descriptive Statistics of the Sample*

Level	N	M (SD)	Number of children with missing age	% girls
PK	203	4.7 (0.3)	7	60%
K	237	5.7 (9.3)	5	57%
G1	213	6.7 (0.4)	2	48%
Total	653	5.8 (0.9)	14	55%

The design of the TPM's implementation in classrooms considered the acquisition of students' lists beforehand. Hence, typing a test identifier in the TPM showed the students' list, which facilitated a fast start of the assessment process and diminished the application time in classrooms. When children finished a game, their performance data was uploaded to an internet cloud for subsequent analysis. For the interested reader, a prior work details the TPM architecture (Navarrete-Ulloa et al., 2023).

Data was collected in the classrooms by assessing complete groups during each visit. For data collection, two research assistants visited each classroom carrying the necessary technological equipment (tablets, mobile internet, etc.). When entering the classroom, the assistants connected a router to provide wireless internet to the classroom. After the assistants introduced themselves and gave group instructions, they provided each child a tablet with the app ready for use and headphones. Afterward, they approached each child to type the test identifier and selected the child's name to start the first game. The assistants followed a protocol including the following guidelines: a) Provide instructions where the TPM is presented as a fun game that they brought for them to play; b) In case they notice a child does not understand the instructions, they should approach the child and explain the task at hand personally. However, they should try their best for these explanations not to influence the child's performance; c) For each child, there should be a five-minute break in the middle of the assessment process (at the end of one of the games).

Children carried out the TPM games autonomously using a tablet, with few exceptions where an adult was required to mediate the test application. In PK classrooms, the assessment process lasted less than 80 minutes. In K and G1 classrooms, the evaluation process lasted less than 60 minutes. The application of the games in each classroom was divided into two sessions, separated by a five-minute break and monitored by the leading researcher. Ninety-four percent of the participants completed both sessions on the same day, while 6% completed the sessions on different days ( $M = 14$  days;  $SD = 6.4$ ; Range = 1-21 days). Regarding class size, classrooms of the PK, K, and G1 levels had an average of 17.6 (range 6-29), 24.4 (range 15-33), and 26.8 (range 22-33) children, respectively.

## Data Analysis

To verify the TPM's criterion validity, the analyses presented below seek to confirm that the TPM can detect significant differences between the three educational levels considered in the present study (PK, K, G1). Concerning construct validity, confirmatory factor analyses were performed to show that the TPM captures two dimensions of children's mathematical thinking: numerical thinking and spatial reasoning (according to international curricular standards). The test has been designed to facilitate diagnostic and formative assessment procedures two years before school; therefore, it is valuable to interpret the data collected in terms of this objective. The three cohorts in our sample represent different assessment points in the learning trajectory of the educational period of interest. The first assessment point is seen as a diagnosis for entry to early education (Pre-Kinder cohort; PK), the second assessment point is a measurement of progress at the end of the first year of early education (Kinder cohort; K), and the third assessment point evaluates the



competencies acquired at the end of early education (first grade cohort; G1). Thus, our selection of cohorts allows us to conceptualize three different levels of student progress regarding their early math skills.

## Results

We included only children with complete data in the seven TPM tasks in the analysis. Here, we present results for the second round ( $n = 653$ ), which was less affected by the COVID-19 quarantine. For completeness, we provide the results of the first round ( $n = 718$ ) as Supplementary Materials (see Navarrete-Ulloa et al., 2025S-b).

### Descriptive Statistics

Table 3 shows descriptive statistics of the sample of children in the second data collection round. Table S4 (Navarrete-Ulloa et al., 2025S-b), provides Pearson correlations between each game's scores and age.

### Reliability Analysis

We first looked at item discrimination levels per task. We computed discrimination indices for each item and task by subtracting each item's average scores for the 27% of participants with top and bottom overall task scores (Ebel & Frisbie, 1991). In what follows, we considered an item's discrimination as good or not based on a threshold of 30% (Ebel & Frisbie, 1991). Number Line Estimation was the task with the least number of good items, while all the items in Numerical Comparison and Perspective Visualization were above the threshold. Table 4 presents the number of items with above-threshold discrimination and Cronbach's alpha coefficients for each task. The Numerical Comparison and the Perspective Visualization tasks obtained the lowest alphas (about .60). In contrast, the Pattern Creation task obtained the highest value (about .80). From this point onwards, we discarded the nine items that did not reach a good level of discrimination.

**Table 4**

*Discrimination and Reliability Data for Each Task*

Game	# good items	Cronbach's alpha	Cronbach's alpha after item removal
Numerical Comparison	10	0.66	0.66
Concrete Pictorial Symbolic Mapping	7	0.71	0.76
Additive Problem Solving	10	0.83	0.83
Number Line Estimation	6	0.77	0.84
Pattern Creation	10	0.86	0.86
Attribute Seriation	8	0.75	0.76
Perspective Visualization	10	0.75	0.75

### Correlations Between First and Second Data Collection

As already mentioned, the first collection of data (T1) was performed in April 2022, and the second one after roughly three months (T2). Table 5 presents the correlations between scores of T1 and T2 for each of the tasks, considering only the subset of children who provided complete data in both times of testing. Table 5 shows strong correlations between the scores of each task at both times (ranging from .37 to .68). However, these correlations should not be taken as estimates of test-retest reliability, due to the long time elapsed between the two measurements: Nunnally and Bernstein (1994) have suggested a 2-week interval for estimating test-retest reliability regarding achievement-type tests. Hence, correlations in Table 5 more likely suggest a change in the sample. For example, a comparison between Figure 1 and Figure S1 (Navarrete-Ulloa et al., 2025S-b) shows that children improved their knowledge during the 3-month period. Recent research on analysis of pretest-posttest data indicates that the acquisition of knowledge between T1 and T2

is associated with lower correlation values between the scores obtained in T1 and T2 (Navarrete-Ulloa, 2024). Hence, these relatively low correlation values more likely reflect a change in the sample due to learning, a better adaptation to school environment, children's developmental changes, among other factors that are confounded with the time elapsed. Nevertheless, Table 5 show significant correlations between all the TPM tasks, and suggest that all of them measure somewhat stable constructs.

**Table 5**

*Pearsons' Correlations Between Children's Scores for Each TPM Game Between the First Data Collection (T1) and the Second Data Collection (T2) (n = 547)*

Game	T1-T2 Correlation
Numerical Comparison	.60***
Concrete Pictorial Symbolic	.45***
Additive Problem Solving	.53***
Number Line Estimation	.37***
Pattern Creation	.68***
Attribute Seriation	.65***
Perspective Visualization	.61***

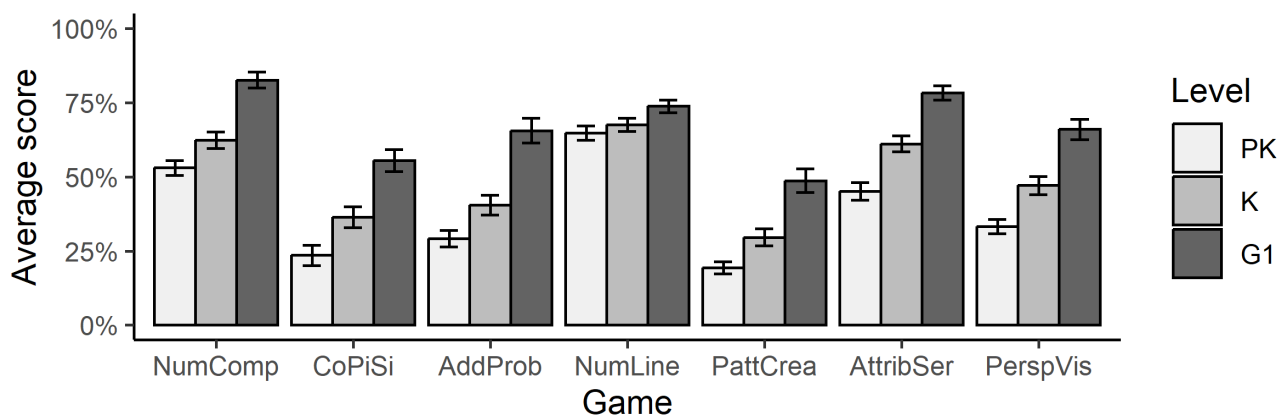
\*\*\* $p < .0001$ .

## Criterion Validity

Figure 1 presents the average scores obtained by children in each educational level and game. As expected, average scores on all the TPM games increased along with educational level. We used  $t$ -tests to separately compare children's scores across consecutive educational levels (PK vs. K, K vs. G1) for each game. Since this amounts to 14 contrasts, we corrected for multiple comparisons using the Holm-Bonferroni procedure (Holm, 1979). This process showed statistically significant differences between all three educational levels in all but one case: only the Number Line Estimation game failed to show a significant difference (between PK and K,  $p = .08$ ; all other  $ps < .0003$ ).

**Figure 1**

*Children's Scores for All Games and Levels*



Note. Vertical bars depict 95% confidence intervals.

## Construct Validity

We then looked at the seven-game set and asked about its factorial structure. According to our theoretical framework, we hypothesized that these tasks would be grouped into two subsets: numerical tasks (Numerical Comparison, Concrete

Pictorial Symbolic Mapping, Additive Problem Solving, and Number Line Estimation) and visuospatial tasks (Pattern Creation, Attribute Seriation, and Perspective Visualization). A confirmatory factor analysis considering these two factors in predicting total scores per task showed a good degree of fit: RMSEA = .015, SRMR = .017, CFI = .999, TLI = .998,  $\chi^2(13) = 14.90$ ,  $p = .31$ . Table 6 shows the factor loadings for this model, revealing that Number Line Estimation was the only task with a loading smaller than .60 in its respective factor.

**Table 6**

*Factor Loadings From the CFA*

Game	Numerical Thinking factor	Visuospatial Reasoning factor
Numerical Comparison	.63	
Concrete Pictorial Symbolic Mapping	.71	
Additive Problem Solving	.81	
Number Line Estimation	.35	
Pattern Creation		.74
Attribute Seriation		.74
Perspective Visualization		.64

To understand how the TPM behaves at the different individual tasks and grade levels, we computed Cronbach's alpha values separately for each task, dimension, and level. Table 7 shows that tasks improve their internal consistency with grade level, that most tasks show acceptable values at the target level (K), and that two tasks show very low alpha values in the youngest age group (PK): Numerical Comparison and Perspective Visualization. It should be noted that both dimensions (Numerical Thinking and Visuospatial Reasoning) have good internal consistency for all levels.

**Table 7**

*Cronbach's Alpha for Each Task, Grade Level, and Dimension*

Task / Dimension	PK	K	G1
Numerical Comparison	.23	.56	.72
Concrete Pictorial Symbolic Mapping	.71	.72	.70
Additive Problem Solving	.62	.74	.86
Number Line Estimation	.79	.84	.87
Pattern Creation	.71	.81	.87
Attribute Seriation	.65	.68	.64
Perspective Visualization	.42	.71	.76
<b>Dimension</b>			
Numerical Thinking	.65	.82	.87
Visuospatial Reasoning	.69	.82	.87

## Measurement Invariance

Finally, we evaluated whether the two factors underlying the TPM show measurement invariance between boys and girls, that is to say, if the factor structure is comparable between these two groups (Kline, 2023). There are different measurement invariance levels: configural, metric, scalar, and strict (Bialosiewicz et al., 2013). Each of these introduces additional requirements to the previous ones. In configural invariance, both groups exhibit the same factor structure. In metric invariance, factor loadings are equal between groups. In scalar invariance, factor loadings and intercepts are equal between groups. Finally, in strict invariance, factor loadings, intercepts, and residuals are equal between groups. The data should display at least scalar invariance to meaningfully compare factor means for boys and girls.

We computed four multigroup (boys/girls) CFAs, one per each invariance level. A chi-square difference test revealed no significant difference between the configural invariance model and the metric invariance model ( $\chi^2_{\text{diff}}(5) = 6.3, p = .27$ ), and a significant difference between the metric invariance model and the scalar invariance one ( $\chi^2_{\text{diff}}(5) = 20.6, p = .001$ ). In addition, the metric invariance model presented a good degree of fit: RMSEA = .033, SRMR = .033, CFI = .993, TLI = .990,  $\chi^2(31) = 41.8, p = .09$ . Altogether, this suggested that the two-factor model for the TPM exhibits metric invariance.

We analyzed each factor's measurement invariance separately to shed further light on the TPM's properties. For the Numerical Thinking factor, the chi-square difference test revealed no significant differences between models ( $\chi^2_{\text{diff}}(3) = 3.3, p = .34$ ;  $\chi^2_{\text{diff}}(3) = 3.5, p = .32$ ; and  $\chi^2_{\text{diff}}(4) = 7.8, p = .10$ , respectively). Since the strict invariance model displayed a good fit (RMSEA = .029, SRMR = .040, CFI = .993, TLI = .994,  $\chi^2(14) = 17.9, p = .21$ ), we concluded that this factor exhibits strict invariance. Instead, for the Visuospatial Reasoning factor, the chi-square difference test indicated no significant difference between the configural invariance and the metric invariance models ( $\chi^2_{\text{diff}}(2) = 2.2, p = .34$ ), and a significant difference between the metric invariance and the scalar invariance models ( $\chi^2_{\text{diff}}(2) = 17.8, p = .0001$ ). Together with the good fit indices of the metric invariance model (RMSEA = .016, SRMR = .021, CFI = .999, TLI = .999,  $\chi^2(2) = 2.2, p = .34$ ), we concluded that this factor shows metric invariance.

In summary, the measurement invariance analysis suggests that the TPM's Numerical Thinking factor is measured with the same degree of precision for both boys and girls (Kline, 2023). In contrast, the Visuospatial Reasoning factor only exhibited metric invariance, indicating that means between boys and girls should not be directly compared.

## Percentile Scales

In the Supplementary Materials, we provided percentile scales for the Numerical Thinking and Visuospatial Reasoning dimension scores and total TPM scores for the entire sample (Navarrete-Ulloa et al., 2025S-b, Table S5). We also provide percentile scales for the PK (Table S6), K (Table S7), and G1 (Table S8) subsamples.

## Discussion

The assessment of mathematical skills in childhood is challenging. Most affordable strategies rely on recording observations during long periods of the learning process. Although there are standardized assessments, they tend to be slow and costly due to the need for individual mediation of early-aged learners. These barriers limit the ability of different educational stakeholders to make effective decisions and implement corrective measures promptly. This work introduced the TPM as an instrument to facilitate assessment in the mathematical area of early childhood education. International standards suggest that math learning at this stage should focus on representing, relating, and operating with whole numbers and describing shapes and spatial relationships (CCSS, 2010). Consistent with these standards, our confirmatory factor analysis shows that the TPM has two dimensions: The first one associated with numerical thinking and the second one with visuospatial reasoning. Primary consistency analyses indicate that TPM scores reflect at least three degrees of mathematical ability and that the difficulty of each game is well-calibrated for the age and ability of the target population. Consequently, the TPM has criterion and construct validity. Furthermore, the reliability analysis (see Table 7) indicates that each TPM dimension can be used for individual assessment (at least) for K and G1 levels—though it is less clear whether they can be used in PK for individual evaluations. Regarding the subgroups of boys and girls, the measurement invariance analysis suggests that the TPM measures Numerical Thinking robustly. Nevertheless, the Visuospatial Reasoning measurement may not be comparable between boys and girls. Finally, although the TPM's application requires mobile devices and headphones, its application in large groups is simple and fast (around 80 minutes), and its results are immediate. Consequently, this first study provides evidence supporting the fact that a technologically-based assessment instrument such as the TPM may be an efficient way to offer high-quality inputs for the quick assessment of math learning in early childhood.

A key challenge for young children's classroom assessment is that these children's abilities are underdeveloped and insufficient to answer traditional tests. Most solutions to this problem require individual mediation to apply the assessment, which implies a significant consumption of resources. In Chile, for example, the evaluation process in early

childhood education establishments is based on systematic observation and recording observations. Because this process is idiosyncratic to the classroom, its results cannot be used for comparative analysis or the monitoring of learning and remedial interventions. Although there are standardized instruments in Chile such as the Precalculus Test (Milicic & Schmidt, 2011), the adaptation of the Utrecht Early Mathematics Assessment Test (Cerdeza Etchepare et al., 2012), and the Kinder Test (Educa UC), the application of these instruments is traditional, with an adult mediating the interaction; thus, their application is expensive for large groups.

Similarly, at the international level, there are standardized instruments such as the CIM, BP, EGMA, REMA, and TEMA (Dong et al., 2021; Ginsburg et al., 2016; Ginsburg & Pappas, 2016; Hoffman & Grialou, 2005; Platas et al., 2016), whose expenses are out of reach for under-resourced communities. The assessment instrument presented in this paper provides a solution that combines efficiency in the assessment process with adequate information delivery. Hence, the TPM could help to empower these communities by providing an assessment alternative that aligns with international standards and requires fewer resources in terms of time and labor. For example, in Chile, only affluent schools can access standardized methods to assess early math skills, as they can afford them. On the other hand, less affluent schools commonly have classrooms with thirty children under the care of a teacher and a teaching assistant, where assessment relies on systematic observation methods. This feature makes it difficult to ensure similar quality standards across public schools for early math education. Although the TPM requires access to the enabling technologies (internet, tablets, etc.), it requires small amounts of teacher labor. Consequently, the TPM would be a valuable tool for educational environments with access to technology where extra professional labor is difficult to obtain.

Additionally, we verified that the play experience guides the assessment, as mandated by pedagogical principles (Hirsh-Pasek, 2009) and the Chilean early education's curricular guidelines. As mentioned earlier, the research assistants introduced the TPM as a fun game to play, along with the graphic design that uses animated characters in everyday fantasy contexts and missions to create a play environment during the evaluation process. Throughout the data collection process, several children expressed that they wanted to "play more games" after finishing all the tasks, while several others asked research assistants, "When do you come back to play again?" These memories constitute anecdotal evidence consistent with the TPM having a game-based design that provides an enjoyable play experience during assessment.

## Limitations and Future Directions

The participants in this study have demographic characteristics associated with the geographic region where it was carried out, so the results should not be generalized to the population of Chile in general. This point is relevant because although the scores obtained through this assessment instrument can be used to make specific comparisons between students (or schools), the lack of further information at the national level that identifies the levels of competence obtained makes it challenging to interpret the test results objectively. More details about this would enable the interpretation of the results to facilitate educational decision-making relevant to the student's (school's) reality and appropriate to their regional or national environment.

Considering that the children took the first TPM during the first days of the 2022 school year, which coincided with the return to face-to-face classes interrupted by the COVID-19 quarantines, the scores of this first data collection could have reflected atypical mathematical learning trajectories in contrast to an everyday schooling context. We thus decided to present the analysis of the second data collection, expecting it to reflect a more typical mathematical learning trajectory. While overall scores were indeed higher in the second data collection (compare Figure 1 and Figure S1, Navarrete-Ulloa et al., 2025S-b), this difference might not be only due to math learning and development but also due to children being more familiar with the teaching methodology and school environment, reflected in an improved relation with the teacher and better behavior in the classroom. Children feeling more at ease in the classroom might have provided better conditions to conduct the second TPM assessment and might partly explain the improved performance. Future research should establish whether there is any significant difference in the score distributions relative to cohorts whose educational process is typical. Researchers aiming to generalize the results presented here would need to evaluate a population at this academic level with a stratified sample according to key educational variables such as geographic



region, administrative unit, gender, and educational level of participants. Better yet, in the future, a more rigorous validation study could consider better standards of criterion validity and add discriminant validity criteria.

This work presented evidence of reliability and construct validity resulting from data analysis associated with two different data collections. This redundancy is because the first data collection was conducted at the end of a lockdown period of almost two years (due to the COVID-19 health contingency). Consequently, participants lacked school experience, their levels of autonomy were compromised, and their learning trajectories were likely atypical. These factors suggest that the evidence of construct validity and reliability from the first data collection might be underestimated (see Navarrete-Ulloa et al., 2025S-b). There is a similar effect for the second data intake, but in the opposite direction, as most students had prior experience with the instrument. This suggests that the evidence associated with the second data collection might be overestimated. By analyzing both data collections, we can ensure that the evidence of construct validity and reliability for the TPM is bounded by the levels of evidence associated with these two data intakes.

The TPM's design strongly aligned with worldwide curricular guidelines, which point out that prekindergarten and kindergarten levels should spend the majority of time learning numbers, their relations and representations (numerical reasoning), along with spatial concepts, their relations and logic (spatial reasoning) (Achieve, 2010; MINEDUC, 2018; NCTM, 2006). Nevertheless, those researchers and practitioners focused on assessing only numerical skills might want to drop the dimension of spatial reasoning to obtain a shorter version of the TPM which might be more efficient to this aim. Using different frameworks would have suggested measuring other important precursors for arithmetic learning, such as verbal counting and ordinality (e.g., Merkley & Ansari, 2016). Still, our theoretical framework and the technology constraints prioritized measuring the seven tasks reported in this study. In this context, estimating the number line tasks has been widely used in the infant mathematical learning literature (Ramani & Siegler, 2008; Siegler & Ramani, 2009; Siegler et al., 2012). Our results indicate that the associated task (NumLine) has standardized factor loading  $\lambda = 0.35$ , which is much lower than the loadings associated with the other three number sets. Hence, there is room to improve the TPM by modifying this estimation task. Along the same lines, the invariance measurement analysis indicates that the Visuospatial Reasoning factor shows only metric invariance, implying that at least a subset of these items may be approached differently by boys and girls, so the comparisons of group means may be contaminated (Gregorich, 2006). Further research is required to understand the origin of such invariance results for this TPM dimension.

Although the TPM technology assesses large groups of children more efficiently, each child experiences an activity of around sixty minutes. This long period of activity might cause exhaustion, and such a factor might have impacted the results presented here. Still, in our experience, the tablet and game-based design succeeded in maintaining children's attention, and we did not see signs of exhaustion in kindergarten. However, some signs of fatigue were observed in prekindergarten. In this context, practitioners might want to apply the two assessment sessions on two separate days, especially for younger children. Additionally, since the TPM does not require a teacher to mediate between a child and their assessment, it cannot capture the learning dimensions that are more readily observable through social interaction with the adult or between children. Concerning criterion validity—that is, the degree of effectiveness with which the TPM detects levels of early mathematical ability—convergent, retrospective, and predictive validity criteria should be studied. For convergent validity, the correlation between the TPM scores and a test of mathematical skills should be estimated, for example, the "TEMA-III battery" (Hoffman & Grialou, 2005), a widely used international standard of early mathematical skills which has been adapted to Spanish. For retrospective and predictive validity, we suggest a comparison of participants' TPM scores with their overall school performance in the academic semester immediately prior to (retrospective) and at the end of (predictive) the academic semester in which the TPM assessment is performed. Concerning discriminant validity, it is essential to distinguish whether the TPM measures early mathematical skills' development or the general cognitive development of boys and girls. For this purpose, a picture vocabulary test such as the PPVT-III may be used, expecting higher correlations between the TPM and TEMA-III and lower correlations between the TPM and PPVT-III.

## Conclusions

This paper highlights the difficulties in conducting mathematical learning assessments in early childhood, especially for under-resourced educational communities with large student groups. Considering international curricular standards for early childhood mathematics, this paper presented the Test of Preschool Mathematics (TPM), which uses automation and gamification technologies to deliver a pleasant, effective, and efficient assessment experience. Although there is room for improvement of the TPM, the current version meets the minimum desirable requirements that could provide valuable information as one of the inputs considered for the assessment of 4- and 5-year-old children's development of mathematical abilities.

---

**Funding:** This work was funded by the Chilean Agency of Research and Development (ANID), grant ANID/FONDEF/IT23I0012 (JN). Additionally, JN was supported by internal grant PI2402 from Universidad de O'Higgins. In addition, DMG, LP, and PD were supported by the grants PIA/Basal FB210005 and PIA/Support 2024 AFB240004; DMG was also supported by the grant Milenio/NCS2021\_014.

---

**Acknowledgments:** The authors have no additional (i.e., non-financial) support to report.

---

**Competing Interests:** The authors have declared that no competing interests exist.

---

**Related Versions:** The present work is related to a prior proceedings paper presented at the 11<sup>th</sup> International Conference on Information and Education Technology (Navarrete-Ulloa et al., 2023). The proceedings paper was focused on the software architecture of the TPM and an exploratory analysis of the data associated to the numerical reasoning dimension.

---

**Data Availability:** The data and analysis script of this study are publicly available (see Navarrete-Ulloa et al., 2025S-a).

---

## Supplementary Materials

The Supplementary Materials contain the following items:

- The data and analysis script (Navarrete-Ulloa et al., 2025S-a)
- Additional materials (Navarrete-Ulloa et al., 2025S-b): These materials provide additional details and analyses related to the TPM (Test of Preschool Mathematics) assessment, which is aligned with the Chilean curriculum and others (see the paper's main text). The supplementary materials enhance the transparency and comprehensiveness of the TPM assessment by providing detailed mappings, transcripts, and additional analyses. The inclusion of the T1 data analysis complements the T2 findings, offering a more complete picture of the research outcomes. The percentile scales and correlation data help educators and researchers interpret the assessment results in a meaningful way.

### Index of Supplementary Materials

Navarrete-Ulloa, J. A., Gómez, D. M., Ponce, L., Munoz-Rubke, F., & Dartnell, P. R. (2025S-a). *Supplementary materials to "Assessing early math skills in preschoolers by using digital games"* [Research data and code]. OSF. <https://doi.org/10.17605/OSF.IO/87XVQ>

Navarrete-Ulloa, J. A., Gómez, D. M., Ponce, L., Munoz-Rubke, F., & Dartnell, P. R. (2025S-b). *Supplementary materials to "Assessing early math skills in preschoolers by using digital games"* [Additional materials]. PsychOpen GOLD. <https://doi.org/10.23668/psycharchives.16191>

## References

- Achieve. (2010). *Comparing the Common Core State Standards in Mathematics and NCTM's "Curriculum Focal Points"*. Achieving the Common Core. <https://eric.ed.gov/?id=ED512110>
- Alsina, Á. (2021). Estableciendo niveles de adquisición de conocimientos matemáticos importantes de 3 a 6 años: Rúbrica ACMI 3-6. *Edma 0-6: Educación Matemática en la Infancia*, 8(2), 17–43. <https://doi.org/10.24197/edmain.2.2019.17-43>
- Andrews, G., Birney, D., & Halford, G. S. (2006). Relational processing and working memory capacity in comprehension of relative clause sentences. *Memory & Cognition*, 34(6), 1325–1340. <https://doi.org/10.3758/BF03193275>
- Araya, R., Calfucura, P., Jiménez, A., Aguirre, C., Palavicino, M. A., Lacourly, N., Soto-Andrade, J., & Dartnell, P. (2010). The effect of analogies on learning to solve algebraic equations. *Pedagogies*, 5(3), 216–232. <https://doi.org/10.1080/1554480X.2010.486160>
- Bialosiewicz, S., Murphy, K., & Berry, T. (2013). *Do our measures measure up? The critical role of measurement invariance*. Claremont Evaluation Center.
- Bynner, J. M., & Parsons, S. (1997). *Does numeracy matter? Evidence from the National Child Development Study on the impact of poor numeracy on adult life*. Basic Skills Agency.
- Casey, B. M., & Ganley, C. M. (2021). An examination of gender differences in spatial skills and math attitudes in relation to mathematics success: A bio-psycho-social model. *Developmental Review*, 60, Article 100963. <https://doi.org/10.1016/j.dr.2021.100963>
- CCSS. (2010). *Common Core State Standards for Mathematics*. CCSSO. [https://www.nctm.org/uploadedFiles/Standards\\_and\\_Positions/Common\\_Core\\_State\\_Standards/Math\\_Standards.pdf](https://www.nctm.org/uploadedFiles/Standards_and_Positions/Common_Core_State_Standards/Math_Standards.pdf)
- Cerda Etchepare, G., Pérez Wilson, C., Moreno Araya, C., Núñez Risco, K., Quezada Herrera, E., Rebolledo Rojas, J., & Sáez Tisnao, S. (2012). Adaptación de la versión española del Test de Evaluación Matemática Temprana de Utrecht en Chile. *Estudios Pedagógicos (Valdivia)*, 38(1), 235–253. <https://doi.org/10.4067/S0718-07052012000100014>
- Clements, D. H., & Sarama, J. (2011). Early childhood mathematics intervention. *Science*, 333(6045), 968–970. <https://doi.org/10.1126/science.1204537>
- Clements, D. H., Sarama, J., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The Research-Based Early Maths Assessment. *Educational Psychology*, 28(4), 457–482. <https://doi.org/10.1080/01443410701777272>
- 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>
- Dong, Y., Clements, D. H., Day-Hess, C. A., Sarama, J., & Dumas, D. (2021). Measuring early childhood mathematical cognition: Validating and equating two forms of the Research-based Early Mathematics Assessment. *Journal of Psychoeducational Assessment*, 39(8), 983–998. <https://doi.org/10.1177/07342829211037195>
- Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. Prentice Hall.
- Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, 34(5), 752–775. <https://doi.org/10.1111/j.1551-6709.2010.01114.x>
- Gentner, D., & Colhoun, J. (2010). Analogical processes in human thinking and learning. In B. Glatzeder, V. Goel, & A. Müller (Eds.), *Towards a theory of thinking* (pp. 35–48). Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-642-03129-8\\_3](https://doi.org/10.1007/978-3-642-03129-8_3)
- Ginsburg, H. P., Lee, Y.-S., & Pappas, S. (2016). A research-inspired and computer-guided clinical interview for mathematics assessment: Introduction, reliability and validity. *ZDM*, 48(7), 1003–1018. <https://doi.org/10.1007/s11858-016-0794-8>
- Ginsburg, H. P., & Pappas, S. (2016). Invitation to the birthday party: Rationale and description. *ZDM*, 48(7), 947–960. <https://doi.org/10.1007/s11858-016-0818-4>
- Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the Confirmatory Factor Analysis Framework. *Medical Care*, 44(11), S78–S94. <https://doi.org/10.1097/01.mlr.0000245454.12228.8f>
- Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences*, 14(11), 497–505. <https://doi.org/10.1016/j.tics.2010.08.005>
- Hirsh-Pasek, K. (Ed.). (2009). *A mandate for playful learning in preschool: Presenting the evidence*. Oxford University Press.
- Hoffman, H., & Grialou, T. (2005). Test of Early Mathematics Ability (3rd ed.) by Ginsburg, H. P., & Baroody, A. J. (2003). Austin, TX: PRO-ED. *Assessment for Effective Intervention*, 30(4), 57–60. <https://doi.org/10.1177/073724770503000409>

- Holloway, D., Green, L., & Livingstone, S. (2013). *Zero to eight: Young children and their internet use*. LSE, London: EU Kids Online. [https://eprints.lse.ac.uk/52630/1/Zero\\_to\\_eight.pdf](https://eprints.lse.ac.uk/52630/1/Zero_to_eight.pdf)
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
- Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 234–259). Oxford University Press.
- Kalra, P. B., & Richland, L. E. (2022). Relational reasoning: A foundation for higher cognition based on abstraction. *Mind, Brain and Education*, 16(2), 149–152. <https://doi.org/10.1111/mbe.12323>
- Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford Press. <https://books.google.cl/books?id=t2CvEAAAQBAJ>
- Lakoff, G. (2006). Conceptual metaphor. In D. Geeraerts (Ed.), *Cognitive linguistics: Basic readings* (pp. 185–238). De Gruyter Mouton. <https://doi.org/10.1515/9783110199901.185>
- Lin, H.-P., Chen, K.-L., Chou, W., Yuan, K.-S., Yen, S.-Y., Chen, Y.-S., & Chow, J. C. (2020). Prolonged touch screen device usage is associated with emotional and behavioral problems, but not language delay, in toddlers. *Infant Behavior and Development*, 58, Article 101424. <https://doi.org/10.1016/j.infbeh.2020.101424>
- Melhuish, E., & Petrogiannis, K. (2006). *Early childhood care & education: International perspectives*. Routledge.
- Merkley, R., & Ansari, D. (2016). Why numerical symbols count in the development of mathematical skills: Evidence from brain and behavior. *Current Opinion in Behavioral Sciences*, 10, 14–20. <https://doi.org/10.1016/j.cobeha.2016.04.006>
- Milicic, N., & Schmidt, S. (2011). *Manual De La Prueba De Precalculo*. Editorial Universitaria.
- MINEDUC. (2018). *Bases Curriculares Educación Parvularia*. Ministerio de Educación de Chile. [https://parvularia.mineduc.cl/wp-content/uploads/2019/09/Bases\\_Curriculares\\_Ed\\_Parvularia\\_2018-1.pdf](https://parvularia.mineduc.cl/wp-content/uploads/2019/09/Bases_Curriculares_Ed_Parvularia_2018-1.pdf)
- Navarrete-Ulloa, J. A. (2024). Learning rates: A correction of gain scores to assess math learning interventions. *Journal of Experimental Education*. Advance online publication. <https://doi.org/10.1080/00220973.2024.2352768>
- Navarrete, J. A., & Dartnell, P. (2017). Towards a category theory approach to analogy: Analyzing re-representation and acquisition of numerical knowledge. *PLoS Computational Biology*, 13(8), Article e1005683. <https://doi.org/10.1371/journal.pcbi.1005683>
- Navarrete, J. A., Gómez, D. M., & Dartnell, P. (2018). Promoting preschoolers' numerical knowledge through spatial analogies: Numbers' spatial alignment influences its learning. *Contemporary Educational Psychology*, 54, 112–124. <https://doi.org/10.1016/j.cedpsych.2018.06.006>
- Navarrete-Ulloa, J. A., & Munoz-Rubke, F. (2022). Playing board games to learn rational numbers: A proof-of-concept. *Mind, Brain and Education*, 16(4), 293–299. <https://doi.org/10.1111/mbe.12335>
- Navarrete-Ulloa, J. A., Ponce Pradenas, L., Flores, C. A., & Verschae, R. (2023). An automated assessment of early math abilities based on digital games. In *ICIET 2023: 2023 11th International Conference on Information and Education Technology. Proceedings* (pp. 172–176). <https://doi.org/10.1109/ICIET56899.2023.10111319>
- NCTM. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence* (2nd print). National Council of Teachers of Mathematics.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed). McGraw-Hill.
- 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(4), Article 110. <https://doi.org/10.1007/s10648-024-09950-6>
- Parsons, S., & Bynner, J. (2005). *Does numeracy matter more?* National Research and Development Centre for Adult Literacy and Numeracy.
- Perez Mejias, P., McAllister, D. E., Diaz, K. G., & Ravest, J. (2021). A longitudinal study of the gender gap in mathematics achievement: Evidence from Chile. *Educational Studies in Mathematics*, 107(3), 583–605. <https://doi.org/10.1007/s10649-021-10052-1>
- Platas, L. M., Ketterlin-Geller, L. R., & Sitabkhan, Y. (2016). Using an assessment of early mathematical knowledge and skills to inform policy and practice: Examples from the Early Grade Mathematics Assessment. *International Journal of Education in Mathematics, Science and Technology*, 4(3), 163–173. <https://doi.org/10.18404/ijemst.20881>
- Ramani, G., Daubert, E., Lin, G., Kamarsu, S., Wodzinski, A., & Jaeggi, S. M. (2020). Racing dragons and remembering aliens: Benefits of playing number and working memory games on kindergartners' numerical knowledge. *Developmental Science*, 23(4), Article e12908. <https://doi.org/10.1111/desc.12908>

- Ramani, G., & Siegler, R. S. (2008). Promoting broad and stable improvements in low-income children's numerical knowledge through playing number board games. *Child Development*, 79(2), 375–394. <https://doi.org/10.1111/j.1467-8624.2007.01131.x>
- Ramani, G., Siegler, R. S., & Hitti, A. (2012). Taking it to the classroom: Number board games as a small group learning activity. *Journal of Educational Psychology*, 104(3), 661–672. <https://doi.org/10.1037/a0028995>
- Richland, L. E., Holyoak, K. J., & Stigler, J. W. (2004). Analogy use in eighth-grade mathematics classrooms. *Cognition and Instruction*, 22(1), 37–60. [https://doi.org/10.1207/s1532690Xci2201\\_2](https://doi.org/10.1207/s1532690Xci2201_2)
- Richland, L. E., & McDonough, I. M. (2010). Learning by analogy: Discriminating between potential analogs. *Contemporary Educational Psychology*, 35(1), 28–43. <https://doi.org/10.1016/j.cedpsych.2009.09.001>
- Richland, L. E., Stigler, J. W., & Holyoak, K. J. (2012). Teaching the conceptual structure of mathematics. *Educational Psychologist*, 47(3), 189–203. <https://doi.org/10.1080/00461520.2012.667065>
- Siegler, R. S., Duncan, G., Davis-Kean, P., Duckworth, K., Claessens, A., Engel, M., Susperreguy, M. I., & Chen, M. (2012). Early predictors of high school mathematics achievement. *Psychological Science*, 23(7), 691–697. <https://doi.org/10.1177/0956797612440101>
- Siegler, R. S., & Ramani, G. (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>
- Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science? A critical review. *The American Psychologist*, 60(9), 950–958. <https://doi.org/10.1037/0003-066X.60.9.950>
- Strasser, K., Lissi, M. R., & Silva, M. (2009). Gestión del Tiempo en 12 Salas Chilenas de Kindergarten: Recreo, Colación y Algo de Instrucción. *Psyke (Santiago)*, 18(1), 85–96. <https://doi.org/10.4067/S0718-22282009000100008>
- Watts, T. W., Clements, D. H., Sarama, J., Wolfe, C. B., Spitler, M. E., & Bailey, D. H. (2017). Does early mathematics intervention change the processes underlying children's learning? *Journal of Research on Educational Effectiveness*, 10(1), 96–115. <https://doi.org/10.1080/19345747.2016.1204640>
- Watts, T. W., Duncan, G. J., Clements, D. H., & Sarama, J. (2018). What is the long-run impact of learning mathematics during preschool? *Child Development*, 89(2), 539–555. <https://doi.org/10.1111/cdev.12713>
- Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher*, 43(7), 352–360. <https://doi.org/10.3102/0013189X14553660>
- Zainuddin, Z., Chu, S. K. W., Shujahat, M., & Perera, C. J. (2020). The impact of gamification on learning and instruction: A systematic review of empirical evidence. *Educational Research Review*, 30, Article 100326. <https://doi.org/10.1016/j.edurev.2020.100326>
- Zhu, J., & Chiu, M. M. (2019). Early home numeracy activities and later mathematics achievement: Early numeracy, interest, and self-efficacy as mediators. *Educational Studies in Mathematics*, 102(2), 173–191. <https://doi.org/10.1007/s10649-019-09906-6>



*Journal of Numerical Cognition* (JNC) is an official journal of the Mathematical Cognition and Learning Society (MCLS).



leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.