Cognition 263 (2025) 106212

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Full Length Article

More gist, better math: Fuzzy-trace theory-based investigation of the relationship between long-term memory and mathematical skills

Michał Obidziński [a,*], Nina Bażela [a,b], Mateusz Hohol [a,*]

[a] Mathematical Cognition and Learning Lab, Copernicus Center for Interdisciplinary Studies, Jagiellonian University in Krakow, Poland
[b] Doctoral School in the Social Sciences, Jagiellonian University in Krakow, Poland

ARTICLE INFO

Keywords:
Long-term memory
Mathematical cognition
Fuzzy-trace theory
Multinomial modeling

ABSTRACT

Despite extensive research on the cognitive basis for mathematical activity, the associations between long-term memory and math skills remain relatively understudied. In our fuzzy-trace theory-driven study, we addressed this issue by investigating the relationships between long-term memory for numbers and prominent math skills, namely approximate number processing, arithmetic fluency, and math reasoning, along with math self-concept. Individuals who performed better in the numerical memory task demonstrated better math reasoning, a higher math self-concept, and were more arithmetically fluent. We did not find an association between memory and approximate number processing. Crucially, our memory task, based on the conjoint recognition model, allowed us to go beyond merely measuring overall performance and, as a result, to test fine-grained memory processes related to two memory traces: verbatim (remembering exact numbers) and gist (remembering a general intuition about a number's magnitude). While both gist and verbatim processes correlated with math reasoning, the associations involving gist-based processes were more prominent, which is consistent with one of the main assumptions of fuzzy-trace theory. This pattern was further supported by the results of the cluster-based analysis. On the other hand, even though math self-concept was positively associated with overall numerical memory performance, it correlated significantly only with verbatim-based process. Overall, our study shows the nuanced role of long-term memory processes in mathematical skills and demonstrates the power of fuzzy-trace theory and multinomial processing tree modeling in the fine-grained investigation of mathematical cognition.

1. Introduction

Mathematical skills, along with reading and writing, are among the most crucial factors for individuals' well-being and professional growth, as well as for the overall progress of societies (Butterworth, 2011; OECD, 2016; Reyna et al., 2009). Consequently, significant research is dedicated to exploring the general architecture of mathematical cognition (Dehaene, 2011; Hohol, 2020; Lakoff & Núñez, 2000) and cognitive profiles associated with individual differences in numeracy (De Smedt, 2022; Dowker, 2019).

A vast body of research shows that mastering mathematics is scaffolded on both domain-specific cognitive functions, mainly the approximate number system (henceforth, ANS), sometimes identified as an innate number sense (Berch, 2005; Halberda et al., 2008), and domain-general ones, including attention, executive functions, and memory (Bull & Scerif, 2001; De Smedt, 2022; Hohol et al., 2017). Regarding the former, more precise – though still approximate – processing of non-symbolic numerosities has been shown to be positively associated with standardized math achievement tests (Halberda et al., 2008); however, this association is relatively small and frequently vanishes when domain-general cognition is controlled for (Amland et al., 2025; Gilmore et al., 2013; Schneider et al., 2017).

Regarding the relationships between domain-general functions and math skills, significant research is paid to memory (DeStefano & LeFevre, 2004). For instance, virtually all classic psychological accounts assume that performance in simple arithmetic depends on retrieving facts from long-term memory (Ashcraft, 1982, 1992; Campbell, 1987; Siegler, 1988). Also, working memory capacity has been shown to be a strong predictor of arithmetic fluency and word problem-solving capacities (Friso-van den Bos et al., 2013; Peng et al., 2016). Conversely, smaller working memory capacity has also been found in children with mathematics learning problems, as reported by Geary et al. (2004). Bearing in mind these credible lines of research, it is important to note that the associations between long-term memory and math skills remain

* Corresponding authors at: Main Square 34 (Spiski Palace, fl 3, rm 1), 31-010 Krakow, Poland.
E-mail addresses: michal.obidzinski@uj.edu.pl (M. Obidziński), mateusz.hohol@uj.edu.pl (M. Hohol).

https://doi.org/10.1016/j.cognition.2025.106212
Received 19 November 2024; Received in revised form 23 May 2025; Accepted 5 June 2025

relatively understudied. It is sufficient to mention that the term "long-term memory", or its abbreviation (LTM), appears only a few times in prominent state-of-the-art textbooks in the field (e.g., Campbell, 2005; Cohen Kadosh & Dowker, 2015; Gilmore et al., 2018; Skeide, 2022), while the topic of working memory receives much more attention.

This might seem like a peculiar oversight as, in addition to simple arithmetic, LTM plays a major role in acquiring mathematical concepts and learning how to solve mathematical problems (e.g., remembering the order of operations, geometric formulas, etc.). This apparent scarcity of research is further compounded by the fact that even research on arithmetic fact knowledge, a skill that is clearly associated with LTM capacity (Ashcraft, 1992), often examines working memory as a mechanism responsible for encoding and retrieving arithmetic facts from the LTM memory, not accounting for LTM memory itself (e.g., Kiefer et al., 2002; McLean & Hitch, 1999). However, the exact contribution of working memory to these cognitive operations depends on the theoretical model adopted (Fournier et al., 2004; Logie, 2016).

1.1. Long-term memory and mathematical skills

Available data consistently points to the important role of LTM in numeracy. As just stated, arithmetic calculation performance is one of the most frequently studied skills in relation to LTM (Ashcraft, 1992). Multiple studies have shown that as mathematical skills develop, children transition from counting or using other procedural strategies when performing calculations to retrieving increasingly complex arithmetic facts directly from LTM (Barrouillet & Fayol, 1998; Cooney et al., 1988; Geary et al., 2004). Consistently, Calderón-Tena and Caterino (2016) demonstrated that LTM retrieval skills become more critical with age, gradually becoming better predictors of proficiency in mathematical calculation and problem-solving. Moreover, deficits in numerical LTM, specifically a reduced ability to memorize and retrieve basic arithmetic facts, have been reported in several studies on math learning problems (Geary, 2011) and are among the diagnostic criteria for the specific learning disorder with impairment in mathematics (commonly referred to as dyscalculia), as outlined in the DSM-5-TR (American Psychiatric Association, 2022, F81.2).

However, some studies suggest that more precise LTM does not necessarily benefit all mathematical skills (Brainerd & Reyna, 1993; Reyna & Brainerd, 1991). While it may seem intuitive that remembering exact facts for specific mathematical problems would aid in solving them, Reyna and Brainerd (1993) highlight several quantitative reasoning tasks in which it is possible to solve problems correctly without remembering specific problem facts, or even where memory of specific facts can negatively impact performance. They identify, for example, frequency judgment, missing digit identification, transitive inference and class inclusion tasks as demonstrating such memory independence. This clearly shows that the association between LTM and mathematical skills is more complex than a simple positive correlation. One possible explanation for these puzzling findings is offered by fuzzy-trace theory.

1.2. Fuzzy-trace theory

Fuzzy-trace theory (henceforth, FTT) was proposed and is still being developed by Brainerd and Reyna (Brainerd & Reyna, 1990a, Brainerd and Reyna, 1990b, 2004; Reyna, 2012; Reyna & Brainerd, 1991; Reyna & Brust-Renck, 2020). Central to FTT is the notion that information is encoded in two distinct memory traces: gist and verbatim. The verbatim trace captures detailed, exact representations, such as the specific sequence of letters in a word (for verbal material) or precise numerical values (for numerical material). It reflects memory for the surface form, which, however, is not identical to the perceptual representation. In contrast, the gist trace encodes the underlying meaning of the information, such as a general intuition about the number's magnitude, providing a more general, less precise memory that facilitates intuitive and efficient problem-solving with higher retention and fewer errors (Brainerd & Reyna, 1990a, Brainerd and Reyna, 1990b).

A key principle of FTT is that gist and verbatim traces are formed independently and in parallel (Brainerd & Reyna, 1988; Chapman & Lindenberger, 1988). Research has shown, for example, that a child's ability to recall the exact numerical details (verbatim trace) does not affect their capacity to remember the gist trace of the same information. This principle contradicts the integration hypothesis, which suggests that gist memories are derived from verbatim memories and are actively constructed when gist information is required (Brainerd & Gordon, 1994).

FTT-driven studies have demonstrated that adults tend to rely mainly on gist traces when solving problems, often without even accessing verbatim traces (Brainerd & Reyna, 1988; Reyna & Brainerd, 1990). This explains why, in some tasks, memory of exact problem facts negatively impacts performance. Operating on precise verbatim information may lead to more errors not only because it requires a higher cognitive workload, but also because it can hinder one's ability to see the bigger picture and fully comprehend the task's overall meaning (Reyna & Brainerd, 2023). A much better strategy often is to remember and retrieve only approximate information or general patterns—the gist trace. This aligns with the finding that the preference for gist processing increases with age (Reyna & Brainerd, 1990) and the fact that enhancing verbatim memory for problem facts can interfere with reasoning, especially in the case of younger children (Reyna & Brainerd, 1991).

However, although relying on gist may appear beneficial in many cases, it is also a common source of literal errors and systematic biases. For instance, Reyna and Brust-Renck (2020) linked a personal tendency to rely on gist in reasoning with higher susceptibility to specific decision biases, such as the Allais paradox. It has also been shown that relying on gist may lead to memory errors, including false memory phenomena such as phantom recollection—a memory illusion where a presented stimulus is perceived as familiar when it is actually new but similar (sharing the same gist) to a previously presented target (Brainerd et al., 2001; Stahl & Klauer, 2009). Relying on the verbatim trace, although more difficult, very rarely leads to false memories and can prevent gist-based memory errors by supporting the rejection of distractors based on comparison to the exact memory of the target – a memory process called recollection rejection (Brainerd et al., 2003).

1.3. Conjoint recognition model

What is particularly noteworthy about the FTT framework are the original research methods grounded in its theoretical foundation, combined with methods of statistical analysis such as the multinomial processing tree modeling (MPT; Erdfelder et al., 2009; Schmidt et al., 2023). MPT is a statistical method assuming that, in a task with a limited set of stimuli and responses, sample frequencies of answers follow a multinomial distribution. This approach allows for the description of cognitive processes leading to specific responses within the chosen theoretical framework, both diagrammatically (as a decision tree) and formally (through underlying model equations). The simplicity of the model's mathematics facilitates the estimation of parameters that correspond to the cognitive processes assumed by the researcher.

One of the implementations of the MPT model on the ground of fuzzy-trace theory is the conjoint recognition model, allowing the investigation and analysis, with some probability, of which memory processes an individual relies on when solving a memory task rather than merely determining whether they remembered something correctly (Brainerd et al., 1999). This approach enables the simultaneous examination of both gist and verbatim memory using a single task, making it possible to compare the quality of memory traces in each type of memory, as well as identify specific deficits in particular processes and potential compensatory mechanisms.

The latest version of the conjoint recognition MPT model (Brainerd et al., 2022) comprises nine trees, representing all possible combinations of test stimuli and questions. In the task, a list of target stimuli is presented to participants during the learning phase, and they are asked to remember as many as possible. After a retention phase, they enter a test phase in which they are presented with three types of stimuli: 1) target items – those presented during the learning phase; 2) related new items – similar to those from the learning phase on either the gist or verbatim level; and 3) unrelated new items – completely different from those in the learning phase. Each stimulus is accompanied by one of three possible probe questions about the familiarity of the presented stimuli:

1. “Is the stimulus identical to the one from the study phase?” – the correct answer is “Yes” only for target items.
2. “Is the stimulus only similar to the one from the study phase?” – the correct answer is “Yes” only for related new items.
3. “Is the stimulus either identical to or similar to the one from the study phase?” – the correct answer is “Yes” for both target and related new items.

Fig. 1 presents a modified version of the conjoint recognition model, in which false-new distractors (FD) are used instead of the unrelated distractors from the standard version. The figure presents all model trees with possible routes to “yes” and “no” responses in various test scenarios. MPT modeling enables researchers to investigate guessing strategies and the factors that influence guessing (e.g., experimental manipulations and group-specific characteristics). For instance, if participants are presented with a target stimulus and the probe “Is the stimulus the same as before?” the model posits that the correct response “yes” could be given based on identity (parameter *I*), representing the retrieval of a verbatim trace. If a verbatim trace for the given stimulus is not retrieved (probability 1 – *I*), a verbatim trace for another similar old stimulus might be retrieved, leading to erroneous recollection rejection (parameter *E*) and an incorrect “no” response. If no verbatim trace is retrieved (1 – *E*), the correct “yes” response can still occur through gist trace retrieval and gist familiarity (parameter S_t). If this route also fails (1 – S_t), the correct answer might still be given by guessing (parameter b_v), though there is also a probability of guessing incorrectly (1 – b_v). This interpretative approach applies similarly to other model tree graphs.

1.4. Associations between long-term gist and verbatim memory and mathematical skills

Since its development in the early 90s, FTT has proven to be a fruitful framework for memory and reasoning research, providing new insights that explain some seemingly counterintuitive developmental shifts in both non-verbal and verbal domains (Brainerd & Gordon, 1994; Brainerd & Reyna, 1990a, Brainerd and Reyna, 1990b; Reyna & Brainerd, 1991). FTT-driven research has also provided new insights into decision biases (Reyna & Brust-Renck, 2020) and the LTM-related characteristics of learning difficulties, such as developmental dyslexia (Obidziński,

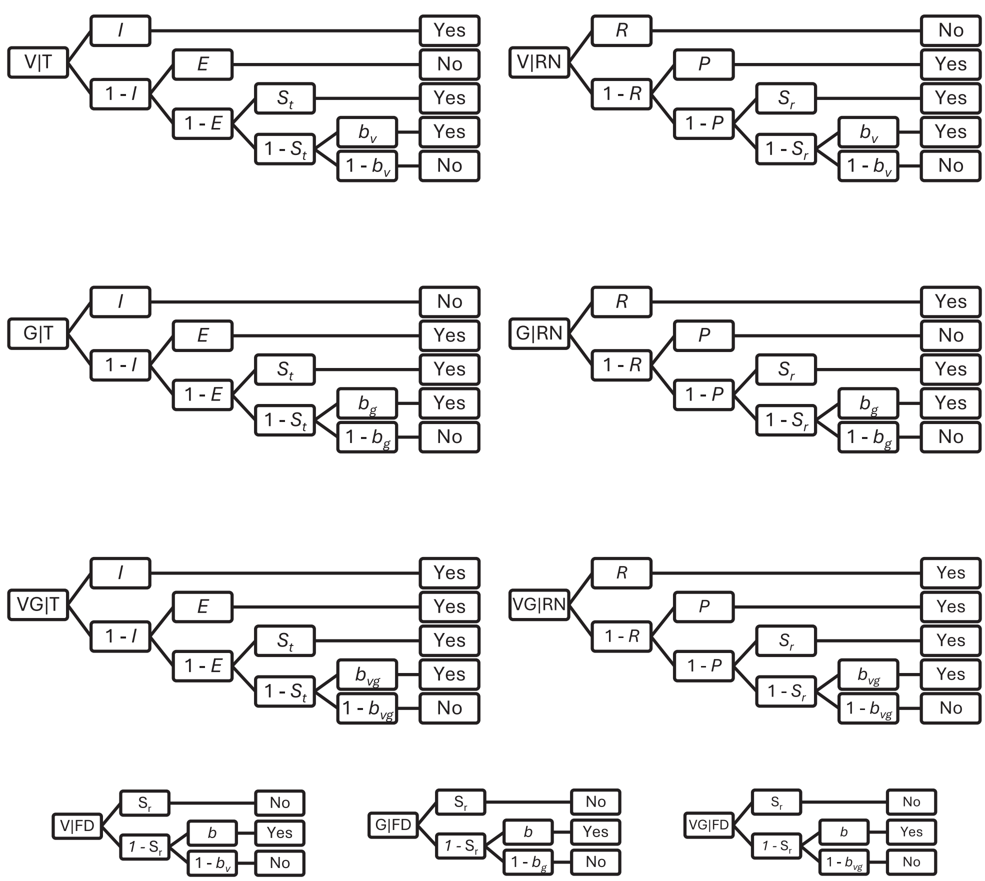

Fig. 1. Conjoint recognition model – possible routes to “yes” and “no” responses in various test scenarios.
Note. The nine parameters depicted in the figure are as follows: *I* - Identity (retrieving a verbatim trace of the target); *R* - Recollection rejection (retrieval of exact verbatim trace in situation of a gist-consistent probe, leading to rejecting incorrect answer); *E* - Erroneous recollection rejection (retrieval of verbatim trace of another stimulus, similar to the one presented with a probe); *P* - Phantom recollection (retrieval of a gist trace of the stimulus, similar to the one presented with a probe, leading to a strong belief that the presented stimulus is a target); S_t - Gist familiarity for target (retrieval of gist trace when a target is presented); S_r - Gist familiarity for related stimulus (retrieval of gist trace when a related new item is presented); b_v - Guessing a “yes” answer in the situation of “is it target” probe; b_g - Guessing a “yes” answer in the situation of “is it new but related” probe; b_{vg} - Guessing a “yes” answer in the situation of “either is it target or new related” probe.

2021; Obidziński & Nieznański, 2017, 2022).

Most of the FTT-driven research on mathematical skills has focused on exploring the relationship between mathematical skills and the tendency to rely on gist rather than verbatim information, as well as proficiency in extracting the gist from mathematical problems (Reyna & Brainerd, 1991, 2023), which is not the same as the ability to store and retrieve numerical information successfully from LTM. A notable exception is Liberali et al. (2012), which investigated the relationship between gist numerical memory and scores on numeracy scales. However, it is important to note that the scales used in this study primarily assessed probability judgment and the ability to deal with fractions and percentages, without accounting for other mathematical skills. Additionally, the task designed to measure gist memory did not allow for modeling memory processes in a way that could assess both gist and verbatim memory simultaneously. As a result, the study could not compare performance between these two types of memory for the same material and their relation to scores on numerical scales—a comparison that could yield valuable insights and highlight the novelty of FTT methods, which enable this kind of analysis.

As mentioned before, not all types of mathematical tasks benefit from LTM equally. While fuzzy-trace theorists include a broad range of tasks measuring mathematical and numerical skills in their studies—including objective numeracy scales (Reyna & Brust-Renck, 2020), relative numerosity judgments in class-inclusion tasks (Brainerd & Reyna, 1990a, Brainerd and Reyna, 1990b), transitive inference about relative quantities (Reyna & Brainerd, 1990), mental arithmetic (Brainerd & Reyna, 1988), and more—FTT has not yet been widely recognized within the mathematical cognition research community. Also, numeracy assessments used in FTT studies often reflect approaches more typical of probability judgment research than of those commonly used in the field of mathematical cognition. Relating verbatim and gist memory to a broader set of math skills, measured with tasks typically used in mathematical cognition research, can therefore help address important abilities that have not yet been studied within FTT. Also, this provides an excellent opportunity to corroborate the assumption of FTT—which distinguishes it from other theories—regarding the prominent role of gist in cognition.

Measuring both verbatim and gist LTM, along with their relationship to different mathematical skills could provide new insights into which types of memory abilities are associated with specific building blocks of overall numeracy. To achieve this, skills at different levels of complexity should be evaluated, ranging from basic processing of non-symbolic numerosities, presumably handled by the ANS, through performing arithmetic operations on symbolic magnitudes, up to mathematical reasoning in different domains. Furthermore, an interesting additional variable is an individual's perception of their own mathematical abilities, called math self-concept (Bong & Skaalvik, 2003; Pajares & Miller, 1994). According to Reyna and Brainerd (2023), when evaluating their mathematical skills, individuals often rely on their memory of past performance in mathematics. Thus, LTM may serve as an important yet implicit factor in self-assessing mathematical abilities. This is also supported by the moderate correlation observed between subjective ratings and objective measures of mathematical competence, as demonstrated in numerous studies (e.g., Fagerlin et al., 2007; Rossi et al., 2022), which suggests that people tend to recall and evaluate their mathematical performance fairly accurately. However, Peters and Bjalkebring (2015) emphasized the influence of context, emotions, and motivation on self-evaluation and found no association between subjective numeracy—a concept related to the self-evaluation of numeracy—and numerical memory performance. Investigating math self-concept in this study will help to better determine its relationship with numerical memory and explore whether self-perceptions of mathematical ability are associated with actual numerical recall.

1.5. The purpose of the study and hypotheses

In our FTT-driven study, we aimed to investigate associations between verbatim and gist LTM for numbers and mathematical skills. To this end, we developed a memory task based on the conjoint recognition model for testing a single sample of participants (Brainerd et al., 2010; Brainerd et al., 2022). This task allowed us to test verbatim and gist memory performance using the same material. To quantify levels of different mathematical skills, we used a dot comparison task tapping into the ANS performance, a speeded test of arithmetic fluency, and a set of problems testing mathematical reasoning. Additionally, we assessed the participant's math self-concept. To investigate the associations between verbatim and gist LTM for numbers and math skills, we performed both correlational and cluster-based analyses.

Research to date has shown that performing arithmetic calculations relies heavily on retrieving precise mathematical facts from memory (Ashcraft, 1992), particularly in the case of multiplication (Campbell & Xue, 2001). This suggests that arithmetic fluency is more strongly associated with exact, verbatim memory. In contrast, efficient mathematical reasoning has been linked to a preference for gist processing and proficiency in extracting gist from mathematical problems (Reyna & Brainerd, 2023; Reyna & Brust-Renck, 2020), indicating a more robust connection between performance in math reasoning tasks and gist memory.

Reyna and Brust-Renck (2020) suggested linking gist processing to the ANS. However, this claim has not yet been investigated in terms of a possible association between performance in a numerical LTM task and a task requiring the comparison of non-symbolic numerosities, which is a gold standard in ANS-related research (Gilmore et al., 2011; Halberda et al., 2008). Additionally, given that math self-concept has been shown to positively correlate with symbolic mathematical performance (Rossi et al., 2022), which in turn is linked to LTM performance (Calderón-Tena & Caterino, 2016), we anticipated a positive correlation between math self-concept and memory task performance, contrary to Peters and Bjalkebring's (2015) findings. However, we did not make any specific predictions about whether this association would be stronger for gist or verbatim memory.

We formulated the following hypotheses regarding the correlations between memory processes and math performance:

1. Overall performance in memory tasks is positively correlated with performance in all the tasks assessing mathematical skills and math self-concept.
2. Mathematical skills correlate:
   - positively with memory processes leading to correct memory performance, namely: Identity (I), Recollection rejection (R), Gist familiarity for target (S_t), and Gist familiarity for related stimuli (S_r);
   - negatively with faulty memory processes leading to incorrect responses: Phantom recollection (P), Erroneous recollection rejection (E) and guessing processes (b_v, b_g, b_{vg}).
3. There are significant differences in the strengths of the correlations between verbatim and gist processes and different mathematical skills. Based on theoretical considerations, the following specific hypotheses were proposed:
   - Performance in the mathematical reasoning task would correlate more strongly with gist memory processes than those associated with verbatim memory.
   - Performance in the dot comparison task would correlate more strongly with gist memory processes than those associated with verbatim memory.
   - Performance in the speeded arithmetic calculation would correlate more strongly with verbatim memory processes than those associated with gist memory.

2. Method

2.1. Participants

Based on an a priori power analysis conducted for a correlation measure (G*Power; v. 3.1.9.7; Faul et al., 2007), a sample size of 123 participants was determined to achieve a power level of 0.80, with $\alpha = 0.05$ (two-tailed), for detecting small correlations (ρ H1 $= 0.25$, ρ H0 $= 0$), and a critical $r = \pm\ 0.177$. This target sample size was reached during recruitment. However, due to a computer malfunction, data from nine participants was incomplete, and their results were excluded from the analysis. The final analyzed sample consisted of 114 participants (79 females, 32 males, 1 who reported another gender, and 2 who chose not to disclose their gender). Participants' age ranged from 18 to 34 years (mean age = 22.83 years; SD = 3.21). 102 (89.5 %) participants self-reported as right-handed, 11 (9.7 %) as left-handed, and 1 as ambidextrous (0.9 %). Regarding education levels, most of the sample consisted of master's students (52; 45.6 %) and bachelor's students (38; 33.3 %). The rest of the sample were graduates (22; 19.3 %) or had completed secondary/vocational education (2; 1.8 %). Participants represented various fields of study, including STEM (32; 28.5 %), social sciences (55; 48.2 %), humanities (15; 13.2 %), and medical fields (6; 5.3 %). Four participants (3.5 %) who declared themselves to be students/graduates did not provide information about their field of study. These fields differed in terms of students' exposure to math (see Supplementary Material 1). All participants had normal or corrected-to-normal vision, were legally adults, and were native Polish speakers. Each participant received a 50 PLN (approx. 12 USD) gift card to a media store for participating in the experiment. Written informed consent was obtained from all participants, who were also informed that they could withdraw from the study at any time without providing a reason.

2.2. Materials

Participants completed an FTT-driven task to assess numerical LTM, as well as a battery of tasks evaluating various aspects of mathematical performance: a dot comparison task tapping into ANS performance, a speeded test of arithmetic fluency, and a set of school math problems testing mathematical reasoning. Additionally, participants reported their math self-concept.

2.3. Numerical long-term memory test

The material used to test LTM for numbers was based on the conjoint recognition model procedure for a single sample of participants (Brainerd et al., 2010; Brainerd et al., 2022). This model allows for the separate analysis of verbatim and gist recognition accuracy for the memorized material. However, instead of standard unrelated new items, false new items were used. This change was made to prevent rejecting a test item based on verbal rather than numerical LTM. Therefore, an extended version of the conjoint recognition model, which includes equations for false but related items, was applied (e.g. Reyna et al., 2016; Singer & Remillard, 2008). In this version, the model equations were modified for the three trees representing unrelated new items by adding a rejection path based on the S_r parameter, resulting in trees for false new items. Thus, the equations for the acceptance of false new items are as follows: $P_V(\mathrm{FN}) = (1 - S_r) \times b_v$, $P_G(\mathrm{FN}) = (1 - S_r) \times b_g$, and $P_{VG}(\mathrm{FN}) = (1 - S_r) \times b_{vg}$.

Our procedure combines elements of the original verbal memory task, in which participants are presented with a list of sentences and are asked to remember as many as possible (Reyna & Kiernan, 1994), with later tasks that employ a conjoint recognition procedure using word lists (Brainerd et al., 2022; Obidziński & Nieznański, 2017). In the current study, after a short retention phase, during which participants perform a buffer task, they enter the test phase. In the latter, they are presented with three types of stimuli: 1) target items, which were shown during the learning phase; 2) related new items, which are similar to the learning-phase items at the gist level; and 3) false new items, which are related to one of the learning-phase items but differ in both verbatim and gist information. Each stimulus is accompanied by one of three possible probe questions regarding the familiarity of the presented stimuli.

To assess LTM for numbers, we created custom new material. The study list comprised 45 sentences, each presenting two groups of items with numbers expressed in Arabic digits (e.g., "The farmer has 7 dogs and 11 horses"). For each target item, two additional items were created: 1) related new items maintained the same numerical relationship but differed in the actual numbers (e.g., "The farmer has 10 dogs and 15 horses"—more horses than dogs, similar to the target sentence), thus differing on a verbatim level but remaining similar on a gist level to the target stimuli; 2) false new items differed both in the numbers and in the numerical relationship (e.g., "The farmer has 20 dogs and 9 horses"—more dogs than horses, reversing the numerical relationship compared to the target sentence). A full list of sentences is provided in Supplementary Material 2.

In the standard conjoint recognition task, all words that the participant has to memorize are presented during a single learning phase of the study (Brainerd et al., 2010; Brainerd et al., 2022; Obidziński & Nieznański, 2017). However, our initial pilot study with 5 participants demonstrated that the material for the numerical memory experiment was too difficult to remember. Presenting all 45 sentences in one study phase resulted in task performance near random chance. To address this, we modified the procedure by presenting the stimuli in three blocks, each consisting of a study phase, a short retention phase (in which participants performed a mental rotation task), and a test phase. The results of the mental rotation task were not analyzed. Between blocks, participants completed tasks assessing mathematical skills (see below). A short training block was added at the beginning of the procedure, with a study phase containing 3 items and a test phase featuring all possible probe types. During the training block, feedback on the given answer was provided along with explanations to ensure participants understood the task instructions. This modified procedure was tested on another group of 12 participants, confirming that the changes allowed for a reasonable level of memorization (0.69), significantly above random, without ceiling or floor effects. Pilot study results are not reported here.

In the final design, participants were asked to memorize 15 sentences with numbers in each of the three study blocks (45 sentences in total). Each sentence was displayed on the screen for 6 s, with a 500 ms fixation point between items.

During each test phase, participants were presented with 15 test stimuli (45 in total). There were three types of stimuli: 1) Target items – those presented during the study phase; 2) Related new items – similar to those from the study phase at the gist level; 3) False new items – related but different in both verbatim and gist of numerical information. For each type, five stimuli (15 in total) were presented, so for each stimulus presented during the study phase, only one type of corresponding test stimulus was shown during the test phase.

Each item was accompanied by one of three probe questions, and participants were instructed to respond with "yes" or "no" by pressing the T or N keys, respectively. The possible questions were:

1. "Is the stimulus identical to the one from the study phase?" – the correct answer was "Yes" only for target items;
2. "Is the stimulus only similar to the one from the study phase?" – the correct answer was "Yes" only for related new items;
3. "Is the stimulus either identical to or similar to the one from the study phase?" – the correct answer was "Yes" for both target and related new items.

In total, 15 stimuli were presented for each probe question—five were target items, five were related new items, and five were false new items.

During a short practice, which preceded the first memory block, a set of three training items was presented. After memorizing these three items, the presented instructions outlined the types of stimuli and the types of questions that would be presented during the test phase. An example item was presented to demonstrate to participants that different stimuli varied only in the specific numbers used, while the rest of the sentence wording remained identical to that in the study phase. During the practice test phase, participants were presented with each possible combination of stimuli and probe question types. To ensure that participants correctly understood the different test questions, they were informed of the correct answer, along with an explanation of why it was correct, for each combination of stimulus and question.

2.4. Dot comparison task

We used Clayton et al.'s (2015b) adaptation of Gebuis and Reynvoet's (2011) protocol to assess participants' performance in comparing non-symbolic numerosities, typically considered a measure of ANS performance. Participants were asked to select which of two dot arrays, yellow or blue, contained more dots by pressing Z for the left array and M for the right one. The numerosities of the dot arrays ranged from 22 to 36, with a numerosity ratio between 0.61 and 1.64, a cumulative surface area ratio from 0.10 to 11.06, and a convex hull ratio between 0.45 and 2.25. The practice session consisted of 8 trials, followed by 96 testing trials administered in a single block. Each trial began with a fixation point displayed for 600 ms, followed by a dot array shown for 600 ms. Afterward, a blank screen appeared while the participants' reactions were recorded. Participants had a maximum of 4400 ms to respond. A higher percentage of correct responses was considered the accuracy measure, indicating better performance in dealing with non-symbolic numerosities. The set of stimuli we used is freely available in (Clayton et al., 2015a), with further details discussed in (Clayton et al., 2015b).

2.5. Math4Speed

Arithmetic fluency was measured using a computerized version of Math4Speed (M4S) – a speeded task involving the four basic operations: addition, subtraction, multiplication, and division (Loenneker et al., 2024). Participants were instructed to solve as many arithmetic problems as possible for each operation within a two-minute time limit per operation (eight minutes total). Each operation was presented in a block consisting of 50 items displayed one after another in a fixed pseudorandomized order, for a total of 200 items. The blocks were presented in the following order: addition, subtraction, multiplication, and division. All problems were presented centrally on the screen, and participants entered their answers using the numeric keypad, confirming each answer by pressing the "Enter" key. To mirror the original paper-and-pencil M4S task, where participants could skip items, we allowed participants to omit problems by pressing "Enter" without providing a solution. However, they were encouraged to avoid skipping items whenever possible. The trials varied in complexity, with 25 simple and 25 complex problems per operation. All addition and subtraction problems involved two-digit numbers. Complexity in addition stemmed from whether carrying was required (e.g., $16 + 37$ vs. $23 + 15$), and for subtraction, from whether borrowing was needed (e.g., 92–79 vs. 67–52). Complex multiplication problems involved one single-digit operand (range: 2–9) and one two-digit operand (range: 12–19), e.g., 7×19, while simple multiplication problems involved only single-digit operands (e.g., 4×5). Division problems included up to three-digit operands (up to 153), with complexity inversely related to multiplication (e.g., $102 \div 6$ vs. $24 \div 6$). The difficulty, stemming from the order of simple and complex problems and the fact that participants were asked not to omit problems, was distributed randomly across the participants. Thus, and according to M4S authors' guidelines (Loenneker et al., 2024), we did not analyze the performance in solving simple and complex problems separately. The total score was the sum of correctly solved arithmetic problems. A higher score in the M4S indicated greater arithmetic fluency. In addition to the total score, we calculated scores for all arithmetic operation types separately. Individual response times were not analyzed.

2.6. Math reasoning problems

Participants completed a set of 20 problems designed to assess mathematical reasoning in the domains of arithmetic (e.g., "600 g of cake costs £8.40. How much does 1 kg of this cake cost?"), algebra (e.g., "Which number belongs to the solution set of the inequality $(x-2)(x+3) < 0$?"), and geometry (e.g., "The total surface area of a cuboid with dimensions $5 \times 3 \times 4$ is equal to…"). For each problem, participants were tasked with selecting the correct answer from five options. The problems were sourced from educational materials used for elementary and secondary school students. This set of problems had been previously used in Szczygieł and Hohol's (2024) study, with some items also appearing in Szczygieł and Sari's (2024) investigation. Each problem was displayed centrally on the screen, with possible answers and corresponding response key mappings listed below. Participants were allowed to use paper and pen to solve the problems, and these supplies were provided. Higher accuracy was interpreted as better math reasoning performance. Response times were not considered, and there was no time pressure.

2.7. Math self-concept

Participants' math self-concept was measured using the Mathematical Ability subscale of the Self-Description Questionnaire III (SDQ III; Marsh & O'Neill, 1984). The subscale consists of four items: "I am quite good at mathematics," "I have trouble understanding anything that is based upon mathematics," "I have always done well in mathematics classes," and "I never do well on tests that require mathematical reasoning." Participants rated their level of agreement with each statement using an 8-point Likert scale (1 = "I don't agree at all" to 8 = "I completely agree"), using the numerical keypad. The second and fourth items were reverse-coded. Responses to all items were summed, with a higher total score indicating a stronger math self-concept.

2.8. Procedure

The study was conducted in a group setting, with up to 5 participants in the lab simultaneously. The entire procedure was computerized and conducted on PCs with 27″ monitors, using OpenSesame (Mathôt et al., 2012). Our procedure and dataset are available on the Open Science Framework (Obidziński et al., 2024). Participants sat approximately 50 cm from the monitor. The tasks were displayed in white Calibri font (40 px) on a dark grayish-purple background (#3d3846) with a resolution of 1920×1080 px. Responses were collected using standard QWERTY keyboards. At the beginning of each session, participants provided written informed consent. Subsequently, they completed a brief demographic survey covering age, gender, handedness, educational level, and, if relevant, their field of study. Following this, participants performed the numerical memory task, the dot comparison task, the M4S task, and solved a set of math reasoning problems. The order of the tasks was counterbalanced across participants, with blocks of the memory task interspersed between tasks measuring mathematical performance. Lastly, participants reported their math self-concept. The entire procedure lasted between 45 and 60 min, was conducted in Polish, and used Arabic digits for numerical representations. The study adhered to the Declaration of Helsinki and received approval from the Rector's Committee on Research Ethics of the Jagiellonian University in Krakow (decision number 1027.0041.5.2024).

2.9. Software and analysis

Data from the numerical memory task was analyzed using the *TreeBUGS* package (Heck et al., 2018) for R language for statistical computing (Posit Team, 2025; R Core Team, 2023), based on the conjoint recognition model (Brainerd et al., 1999), adapted for a single participant group (Brainerd et al., 2010). Then we calculated correlations between variables of interest—namely, the results for gist and verbatim recall in the memory task, and scores related to mathematical performance and math self-concept using *psych* (Revelle, 2024) and *correlation* (Makowski et al., 2020) R-packages. In further analyses, we separated the sample into groups with different levels of math skills using the random forest clustering algorithm provided by the Machine Learning module of JASP (JASP Team, 2024) and compared the clusters' mean levels of each memory parameter using ANOVA.

3. Results

3.1. Memory correctness and measures of mathematical abilities

Descriptive statistics for the numerical LTM test, ANS (dot comparison task), M4S (arithmetic fluency), scores obtained in a set of math reasoning problems, and math self-concept are presented in Table 1. To examine the relationships between numerical LTM and various dimensions of mathematical performance, we used Spearman's ρ correlation and applied a False Discovery Rate (FDR) correction to adjust for multiple comparisons (Benjamini & Hochberg, 1995). Table 2 presents the results of the correlation analysis. Overall memory performance was significantly correlated (two-tailed test) with scores obtained in a set of math reasoning problems (small positive correlation). Given that mathematics generally requires memorization of facts and rules, we tested for the significance of the one-tailed test to identify positive correlations. The test revealed very small, significant positive correlations between memory performance and math self-concept, and multiplication and division scales of M4S. ANS performance (measured as the percentage of correct responses) correlated significantly (very small positive correlation) only with performance in solving math reasoning problems. Math self-concept was significantly correlated with both math reasoning performance and M4S scores, including the overall score and four subscales – all of these correlations were moderate and positive. Finally, we found significant correlations between math reasoning performance and the overall M4S score, as well as with each of the M4S subscales—all of these correlations were moderate and positive.

Table 1
Descriptive statistics for memory and mathematical skills.

	M	Median	*SD*	Shapiro-Wilk *p*
LTM	0.70 [31.36]	0.71	0.12 [5.18]	**.031**
ANS	0.70 [66.88]	0.70	0.12 [11.23]	**< .001**
M4S Addition	0.44 [22.15]	22	0.12 [6.16]	.686
M4S Subtraction	0.32 [15.96]	16	0.12 [5.96]	.317
M4S Multiplication	0.32 [16.18]	14.5	0.17 [8.37]	**< .001**
M4S Division	0.30 [14.78]	12	0.19 [9.31]	**< .001**
M4S All	0.35 [69.05]	66	0.13 [26.32]	**< .001**
Math reasoning	0.73 [14.67]	15	0.19 [3.79]	**< .001**
Math self-concept	0.63 [21.62]	22	0.24 [6.60]	**< .001**

Note: Significant results are bolded. In columns *M* and *SD*, the proportion of correct answers, ranging from 0 to 1 is presented outside the brackets. Inside the brackets, a raw score for the given variable is presented. Raw scores range from 0 to 45 for LTM, 0 to 96 for ANS, 0 to 50 for the individual M4S subscales, 0 to 200 for the M4S overall scale, 0 to 20 for math reasoning, and 4 to 32 for math self-concept. For the median, the only score presented is the one relevant to the conducted analysis (proportion of correct answers for LTM and ANS and the raw score for the rest of the variables). For SD, once again, values for both the proportion of correct answers and the raw score are presented.

3.2. Multinomial modeling results

Latent-trait model analysis was conducted using the *treeBUGS* package (Heck et al., 2018). The model was fitted using 100,000 iterations of an MCMC sampler and a 10,000 *burn-in period*. Goodness of fit was inspected visually (Fig. 2), and with the use of *T1* (observed = 0.07, predicted = 0.07, $p = .462$) and *T2* (observed = 3.85, predicted = 3.56, $p = .423$) statistics. All these approaches indicate that the model fits the data well and allows for an informative interpretation of model parameters. Estimated parameter means, standard errors (SE), and confidence limits (CL) are presented in Table 3.

Individual parameter estimates were extracted using the *getParam* function in treeBUGS to examine within-subject differences and the relationship between memory parameters for numerical material and various measures of mathematical performance.

3.3. Within-subject analysis of differences between model parameters

To test the difference between the *verbatim* and *gist* parameters, a within-subjects difference test was conducted using the transformedParameters function of TreeBUGS. The results are presented in Table 4. The results of the test should be interpreted as follows: a substantial difference between the tested variables exists if and only if both ends of the confidence interval (2.5 % and 97.5 % in the table) have positive or negative values (Schmidt et al., 2023).

The results indicate the existence of substantial differences between most of the analyzed variable pairs. Specifically: The *recollection rejection* parameter is larger than *identity*; *Gist familiarity* for target items is larger than *gist familiarity* for related items; *Gist familiarity* for target items is also larger than *identity*; Both *recollection rejection* and *gist familiarity* for related items are larger than *phantom recollection*; *Erroneous recollection rejection* is smaller than *gist familiarity* for target items but larger than *identity*. There is no substantial difference between *recollection rejection* and *gist familiarity* for related items, *identity* and *erroneous recollection rejection*, or between *erroneous recollection rejection* and *phantom recollection*.

3.4. Correlation of model parameters and mathematical skills

Spearman's ρ, with FDR correction applied to account for multiple comparisons, was used to investigate the relationships between memory parameters obtained through multinomial modeling and measures of mathematical skills. The results are presented in Table 5. All memory parameters, except for two (*Identity* and *bias* for *vg* probes), correlate significantly with performance in math reasoning problems. The strength of these significant correlations is small. *Bias* for *v* probe and *phantom recollection* correlate negatively, while *recollection rejection*, *erroneous recollection rejection*, and *familiarity* for both target and related items and *bias* for *g* probe correlate positively—this pattern of correlation direction holds for all other significant correlations. Math self-concept correlates significantly with one verbatim memory parameter, the *erroneous recollection rejection*. The strength of this correlation is small. No other significant correlations were found.

3.5. Differences in memory parameters between clusters based on math skills

To investigate the robustness of the findings on the relationships between numerical LTM for verbatim and gist information and math skills reported in the previous section, we conducted additional cluster analyses. More specifically, using random forest cluster analysis algorithms, we separated our sample into groups based on participants' mathematical skills, and then compared the clusters' mean levels for each memory parameter using ANOVA. Random forest cluster analysis is an unsupervised algorithm which works by creating multiple decision tree models, each using a random subset of features and bootstrapped

Table 2
Spearman's ρ correlation (FDR corrected) results for memory and mathematical skills.

	LTM	ANS	M4S Addition	M4S Subtraction	M4S Multiplication	M4S Division	M4S All	Math reasoning	Math self-concept
LTM	–	0.10	0.11	0.07	0.18†	0.17†	0.15	0.28**	0.18†
ANS	0.10	–	0.10	0.06	0.13	−0.07	0.07	0.19†	< 0.01
Math reasoning	0.28**	0.19†	0.50***	0.53***	0.52***	0.52***	0.60***	–	0.62***
Math self-concept	0.18†	< 0.01	0.48***	0.57***	0.47***	0.58***	0.61***	0.62***	–

Note: *$p < .05$, two-tailed. **$p < .01$, two-tailed. ***$p < .001$, two-tailed. †$p < .05$, one-tailed

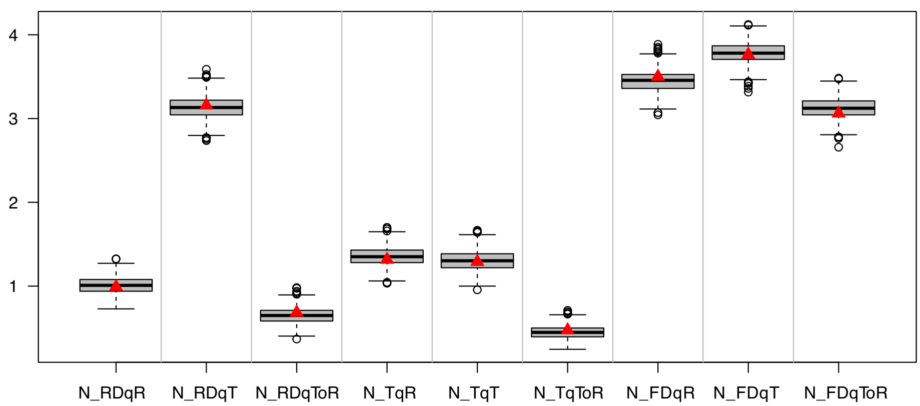

Fig. 2. Observed (red triangles) and predicted (boxplot) mean frequencies of responses of the conjoint recognition model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Note. N_RDqR – Answer "No" to the related distractor when the "is it similar?" probe was presented; N_RDqT – Answer "No" to the related distractor when the "is it identical?" probe was presented; N_RDqToR – Answer "No" to the related distractor when the "is it either identical or similar?" probe was presented; N_TqR – Answer "No" to the target item when the "is it similar?" probe was presented; N_TqT – Answer "No" to the target item when the "is it identical?" probe was presented; N_TqToR – Answer "No" to the target item when the "is it either identical or similar?" probe was presented; N_FDqR – Answer "No" to the false distractor when the "is it similar?" probe was presented; N_FDqT – Answer "No" to the false distractor when the "is it identical?" probe was presented; N_FDqToR – Answer "No" to the false distractor when the "is it either identical or similar?" probe was presented.

Table 3
Estimated parameters of conjoint recognition model for numerical material.

Parameter	Mean	SE	95 % CL
b_g	0.52	0.04	0.45–0.60
b_v	0.38	0.04	0.31–0.45
b_{vg}	0.62	0.03	0.55–0.68
E	0.11	0.04	0.02–0.19
S_r	0.35	0.04	0.26–0.43
S_t	0.76	0.06	0.64–0.87
P	0.03	0.03	0.00–0.10
R	0.47	0.04	0.39–0.53
I	0.04	0.02	0.004–0.10

Note. Guessing a "yes" answer in the situation of "is it target" probe (b_v); Guessing a "yes" answer in the situation of "is it new but related" probe (b_g); Guessing a "yes" answer in the situation of "either is it target or new related" probe (b_{vg}); Erroneous recollection rejection (E); Gist familiarity for related stimuli (S_r); Gist familiarity for target (S_t); Phantom recollection (P); Recollection rejection (R); Identity (I).

Table 4
Measures of within-subject differences between model parameters.

Difference between	M	SD	2.5 %	97.5 %
$S_t - S_r$	0.41	0.07	0.28	0.53
I – R	−0.39	0.05	−0.47	−0.29
S_t – I	0.72	0.06	0.61	0.84
S_r – R	−0.07	0.06	−0.19	0.05
S_r – P	0.32	0.05	0.21	0.41
R – P	0.39	0.04	0.30	0.48
S_t – E	0.65	0.06	0.53	0.77
I – E	−0.07	0.04	−0.16	0.02
P – E	−0.08	0.05	−0.17	0.02

Note. Erroneous recollection rejection (E); Gist familiarity for related stimuli (S_r); Gist familiarity for target (S_t); Phantom recollection (P); Recollection rejection (R); Identity (I).

data. Each tree creates a proximity matrix that is used in a further step as a basis for a singular clustering model (e.g., Bicego, 2019; Breiman, 2001). Taking advantage of this method, we built two models that were used in the subsequent analysis.

In the Model 1, the random forest algorithm was provided with all math skills measures, namely, ANS performance, overall M4S score, separate scores for the four arithmetic operations, math reasoning, and math self-concept. Cluster determination was optimized according to the Bayesian Information Criterion (BIC) and was not fixed. Based on this measure, the algorithm identified three clusters within the sample. The first cluster (C1), consisting of 24 participants (21 % of the sample), was characterized by a low mean level of all variables used by the algorithm. The second cluster (C2), consisting of 66 participants (58 %), was characterized by a medium mean level of all variables used by the algorithm. The final cluster (C3), consisting of 24 participants (21 %), was characterized by a medium level of mathematical reasoning and a high mean level of all other variables used by the algorithm.

Table 6 presents the results of the ANOVA (with Welch's correction applied, as the assumption of homoscedasticity was violated) and the cluster mean levels for each memory parameter. The Games–Howell post hoc test was conducted when a significant difference in a given memory parameter was observed. We found significant differences between clusters: C1-C3 and C2-C3 for S_r parameter ($p = .025$ and $p = .044$, respectively); C2-C3 for S_t parameter ($p = .017$); C1-C3 for I parameter ($p = .023$).

A second cluster analysis (Model 2) was conducted to identify and compare groups in that show a larger difference in math reasoning levels. All variables, except ANS performance, were used in the random forest algorithm. As in the previous analysis, cluster determination was

Table 5
Spearman's correlations (FDR corrected) between model parameters and mathematical skills.

	ANS	M4S Addition	M4S Subtraction	M4S Multiplication	M4S Division	M4S All	Math reasoning	Math self-concept
b_g	0.08	0.08	0.04	0.06	0.06	0.07	0.22*	0.07
b_v	−0.08	−0.07	−0.07	−0.15	−0.16	−0.13	−0.20*	−0.12
b_{vg}	< −0.01	−0.06	−0.11	−0.07	−0.15	−0.11	< −0.01	−0.11
E	0.16	0.15	0.19	0.16	0.19	0.18	0.33**	0.28*
S_r	0.11	0.11	0.10	0.21	0.14	0.16	0.32**	0.18
S_t	0.14	0.07	< −0.01	0.13	0.05	0.08	0.27*	0.13
P	−0.11	−0.05	−0.02	−0.10	−0.14	−0.11	−0.22*	−0.17
R	0.14	0.07	0.04	0.13	0.14	0.11	0.26*	0.19
I	−0.02	0.07	0.04	0.09	0.15	0.09	0.14	0.06

Note. *p < .05, two-tailed. **p < .01, two-tailed. ***p < .001, two-tailed. †p <.05, one-tailed. Guessing a "yes" answer in response to the "is it target" probe (b_v); Guessing a "yes" answer in response to the "is it new but related" probe (b_g); Guessing a "yes" answer in response to the "either is it target or new related" probe (b_{vg}); Erroneous recollection rejection (E); Gist familiarity for related stimuli (S_r); Gist familiarity for target (S_t); Phantom recollection (P); Recollection rejection (R); Identity (I).

Table 6
The results of ANOVA (with Welch's correction) and cluster mean levels for each memory parameter (Model 1).

Parameters	M_{C1}	M_{C2}	M_{C3}	F	df	p	η_p^2
b_g	0.52	0.51	0.54	1.29	2; 111	.280	0.02
b_v	0.41	0.39	0.37	1.88	2; 111	.158	0.03
b_{vg}	0.61	0.61	0.61	0.35	2; 111	.703	0.01
E	0.15	0.15	0.17	1.19	2; 48.47	.314	0.01
S_r	0.33	0.37	0.48	3.73	2; 111	**.027**	0.06
S_t	0.65	0.67	0.77	5.13	2; 53.49	**.001**	0.04
P	0.07	0.08	0.05	2.04	2; 111	.135	0.04
R	0.42	0.42	0.47	1.43	2; 111	.245	0.03
I	0.08	0.15	0.25	4.43	2; 50.13	**.017**	0.07

Significant results are bolded.

optimized according to the BIC and was not fixed. Based on this measure, the algorithm identified three clusters within the sample. The first cluster (C1'), consisting of 29 participants (~25 % of the sample), was characterized by a low mean level of all variables used by the algorithm. The second cluster (C2'), consisting of 57 participants (50 %), was characterized by a medium mean level of all variables used by the algorithm. The final cluster (C3'), consisting of 28 participants (~25 %), was characterized by a high mean level of all variables used by the algorithm.

Table 7 presents the results of ANOVA (again, with Welch's correction applied, as the assumption of homoscedasticity was violated) and cluster mean levels for each memory parameter. The Games–Howell post hoc test was conducted when a significant difference in a given memory parameter was observed. We found significant differences between clusters: C1'-C3' and C2'-C3' for S_r parameter (p = .008 and p = .049, respectively); C1'-C3' and C2'-C3' for S_t parameter (p = .027 and p = .023, respectively).

Table 7
The results of ANOVA (with Welch's correction) and cluster mean levels for each memory parameter (Model 2).

Parameters	$M_{C1'}$	$M_{C2'}$	$M_{C3'}$	F	df	p	η_p^2
b_g	0.51	0.52	0.53	0.91	2; 111	.407	0.02
b_v	0.41	0.39	0.37	2.53	2; 111	.085	0.04
b_{vg}	0.62	0.61	0.61	0.78	2; 111	.462	0.01
E	0.15	0.15	0.16	0.34	2; 111	.712	0.01
S_r	0.33	0.37	0.47	4.33	2; 111	**.016**	0.07
S_t	0.64	0.67	0.77	5.60	2; 63.47	**.006**	0.06
P	0.08	0.07	0.06	2.09	2; 111	.130	0.04
R	0.42	0.41	0.48	2.35	2; 111	.100	0.04
I	0.10	0.15	0.22	2.76	2; 60.10	.071	0.04

Note. Guessing a "yes" answer in the situation of "is it target" probe (b_v); Guessing a "yes" answer in the situation of "is it new but related" probe (b_g); Guessing a "yes" answer in the situation of "either is it target or new related" probe (b_{vg}); Erroneous recollection rejection (E); Gist familiarity for related stimuli (S_r); Gist familiarity for target (S_t); Phantom recollection (P); Recollection rejection (R); Identity (I). Significant results are bolded.

In both models, clusters characterized by higher math skills showed a significantly higher probability of retrieving gist for two gist-related parameters: *gist familiarity for related stimuli* (S_r) and for *targets* (S_t). The *identity* (I), which is one of the processes involved in verbatim memory was also significantly higher in this group, but only in Model 1. On the other hand, we did not find significant differences between the low and medium math skill clusters.

4. Discussion

The study investigated the relationship between LTM of numerical material and mathematical skills. In particular, using the fuzzy-trace theory as our study's framework, we were interested in how two types of numerical memory traces—verbatim and gist—relate to dealing with non-symbolic numerosities, speeded symbolic arithmetic, and school math reasoning problems of various domains. Additionally, we investigated an association between numerical memory and self-perceiving of mathematical competence, called "math self-concept".

4.1. Consistency of mathematical performance and math self-concept

We decided to include a broad range of tasks that assess distinct math skills, as each of them may vary in their association on gist and verbatim LTM for numbers. Against the results of previous studies, one could expect that the performance in tasks tapping into distinct mathematical skills will be related positively (Amland et al., 2025; Schneider et al., 2017). The same holds true regarding math skills measured objectively and math self-concept (Rossi et al., 2022). In this line, we found that better mathematical reasoning is positively associated with performance in comparing non-symbolic numerosities (very small correlation) and symbolic arithmetic fluency (moderate correlation), as well as with math self-concept (moderate correlation). This last result supports the use of self-evaluation measures of math skills as a useful proxy for objectively measured math skills. Specifically, we found a correlation above ρ = 0.60 between subjective (math self-concept) and objective (math reasoning, overall arithmetic fluency) measures. This finding aligns with the results of pioneering work in this field (Fagerlin et al., 2007) but contrasts with the results of Peters (2020), who reported smaller correlations across different studies.

4.2. Overview of verbatim and gist traces in numerical memory

Examining the specificity of numerical LTM multinomial model parameter analysis shows that verbatim and gist traces play important, although different, roles in memory for numerical material. For target items, there is a significantly lower probability of retrieving the verbatim trace (*identity* and *erroneous recollection rejection*), while the probability of gist retrieval (*familiarity for target*) is relatively high. On the other hand, there is no significant difference in the probability of retrieving a verbatim trace (*recollection rejection*) when a related new item is present, compared to the probability of retrieving a gist trace (*familiarity for related new items*). However, the probability of retrieving a verbatim trace is significantly higher than that of *phantom recollection*, indicating a difference in this comparison. Moreover, the probability of retrieving verbatim is higher for related than target items. Note that this pattern is different from what is typically observed in studies using the conjoint recognition model with verbal material (e.g. Brainerd et al., 2022). In our study, we used a new, complex material including numerical stimuli represented with Arabic digits. We speculate that the unexpected pattern we observed could be characteristic of the numerical domain. For example, it might be the case that even a relatively small change in verbatim information in related new items leads to greater incentives to use verbatim memory compared to changes in word stimuli. Future studies should take this possibility into consideration.

4.3. Relationship between numerical LTM and mathematical skills

Analysis of both overall memory performance and separate memory parameters of the conjoint recollection model shows a significant correlation between numerical memory and some (but not all) mathematical skills measured. On the level of overall memory performance, higher scores in the numerical memory task relate to better performance solving math reasoning problems, better performance in multiplication and division subscales of the M4S task, and a higher math self-concept. We found no significant correlation between memory and performance in comparing non-symbolic numerosity nor performance in addition and subtraction subscales of the M4S. These results indicate that the relationship between LTM and mathematical skills is complex, even if tasks tapping into symbolic math skills are considered.

The analysis using the memory parameters of the conjoint recognition model shows that both verbatim and gist traces are related to math skills, although not in the same way. The performance in solving math reasoning problems correlates significantly with most memory parameters: it increases with the higher levels of *identity*, *recollection rejection*, *erroneous recollection rejection*, *familiarity* for both *targets* and *related new* items, and *bias* for *g* probe; it decreases with the higher levels of *bias* for *v* probes, and *phantom recollection* parameters. Thus, all gist, but not all verbatim, parameters are associated with math reasoning performance. A higher math self-concept is associated with higher values of *erroneous recollection rejection*, which is one of the verbatim parameters. Despite a correlation with overall LTM performance, arithmetic fluency was not significantly related to any memory parameter.

4.4. Hypothesis validation

Overall, the just summarized findings partially support two of the three proposed hypotheses. As predicted in Hypothesis 1, overall numerical LTM performance is positively associated with mathematical skills, however, these correlations are not significant for all of them. In full accordance with the hypothesis, math reasoning performance correlates significantly with numerical memory performance. For arithmetic fluency, multiplication and division performance are correlated significantly, while performance of addition and subtraction are not. Our positive finding may be linked to the fact that multiplication and division tables are memorized during elementary education (Campbell & Xue, 2001), however, the associations between LTM and these operations that we found appeared to be small. The performance in the dot comparison task did not correlate significantly with LTM in any capacity. This null finding can be accounted for by the fact that instant non-symbolic comparisons, supposed to be handled by the ANS, are the least related to the memory of all the tasks we used—given its elementary nature. Alternatively, lack of correlation between ANS and LTM performance might be due to non-symbolic vs. symbolic modality of used tasks—a non-symbolic memory test might show a different pattern of results. Consistent with the hypothesis, LTM performance correlated with math self-concept, though the correlation was small. Although this result differs from the one obtained by Peters and Bjalkebring (2015), it does not necessarily contradict their view on the self-evaluation of math skills. Peters emphasized the strong dependence of mathematical self-evaluation on context, confidence, and emotions. In this study, the memory task constituted a large part of the experimental procedure, and the math self-concept questionnaire was completed at the end. As a result, performance on the memory task could have strongly influenced participants' confidence in their mathematical abilities. While this limitation does not rule out an association between numerical memory and math self-concept, it highlights the need to counterbalance conditions between participants and to ask different participants to assess their math self-concept at different points in the procedure. This approach would better account for the multi-faceted nature of math self-concept, which may be influenced by specific task conditions.

For Hypothesis 2, correlations of almost all parameters have expected directions, for both significant and insignificant correlations. The only two exceptions are *erroneous recollection rejection* and *bias* for *g* probe, both of which correlate positively, not negatively, with all mathematical skills. The unexpected results for *erroneous recollection rejection* might relate to the fact that when the test probe asks *if an item is identical or similar* (VG condition), it leads to the correct answer. Thus, participants with better math skills may utilize this process to their advantage. Alternatively, considering that nearly all correlations with this parameter are insignificant, it is likely that *erroneous recollection rejection* indicates an overall stronger verbatim trace—which is connected with better math skills—while it does not play a critical role in the investigated math skills. The unexpected result regarding *bias* for the *g* probe calls for investigation in future studies. It is possible that, due to strong recollection rejection, it is beneficial to guess "yes" for items for which one does not recall verbatim or gist information. Thus, it could represent a guessing strategy indicative of good memory performance and, therefore, correlate with better mathematical performance.

As we did not find significant correlations between arithmetic fluency or ANS performance and any memory parameters, no conclusions can be drawn regarding the second and third points of Hypothesis 3, which predicted that these math skills would correlate more strongly with either gist or verbatim processes. Regarding the first point of this hypothesis, which predicted a stronger association between mathematical reasoning and gist processes than with verbatim ones, the paired *r*-test revealed no significant differences in the strength of correlations between math reasoning and any of the multinomial model parameters. Therefore, based solely on the comparison of the correlation strengths, no conclusions can be drawn about the first point of Hypothesis 3.

However, a closer inspection of the significant correlations reveals a meaningful pattern: all three gist parameters significantly correlate with mathematical reasoning, whereas only two verbatim parameters do. This pattern lends some support to Hypothesis 3 in the context of math reasoning, even in the absence of statistically significant differences in the correlation strengths.

4.5. Associations between gist and verbatim memory traces and mathematical skills

Although the research conducted so far within the framework of FTT has not explicitly investigated associations between gist and verbatim LTM and various mathematical skills, it has characterized the memory representations of numbers encoded in memory, namely: categorical gist (some vs. none), ordinal gist (more vs. less), and verbatim (exact numbers) (Reyna, 2012; Reyna & Brust-Renck, 2020). It has also explored multiple decision-making and reasoning effects, showing that adults prefer to make various decisions—including those based on probabilities or frequencies—using the simplest level of representation that allows them to discriminate between options and solve the task, resorting to verbatim representations only when necessary (Reyna, 2012; Reyna et al., 2023). Similar to decision problems, where people tend to rely on different levels of precision in numerical representations depending on the task, it can be assumed that the ability to memorize and operate on gist-based numerical information may be more crucial for several types of mathematical tasks.

In our study, we investigated only ordinal gist and verbatim memory and hypothesized differences in their associations with various mathematical skills. We expected stronger associations between gist memory and mathematical reasoning (since some problems in this domain may be solved without relying on exact calculations), as well as ANS. Conversely, we anticipated a stronger association between verbatim memory and arithmetic calculations. We found significant positive associations between various gist and verbatim memory processes and math reasoning. Thus, our results suggest greater involvement of the gist memory trace, with still significant involvement of the verbatim trace. Nevertheless, we did not observe a clear difference in the strength of the relationship between the two memory traces and other math skills. Although null results are challenging to interpret, this lack of difference would suggest that both gist and verbatim might play considerable, although different, roles in mathematical performance (at least in more complex mathematical tasks). Aligned with this interpretation, a broader analysis of differences in parameters and correlations shows a considerable pattern of differences between gist and verbatim involvement in numerical LTM. For the accurate memory of target items, gist plays a significantly more important role than verbatim, supporting the claim that developed mathematical strategies are based on gist. At the same time, accurate memory for related new items relies equally on *recollection rejection* (a verbatim process) and *familiarity* (a gist process). This, along with the finding that the probability of *recollection rejection* is significantly higher than *identity*, highlights the important role of numerical verbatim memory as a control mechanism against errors.

There is also the possibility that the discrepancy in the significance of gist and verbatim processes observed in previous studies stems from the experiment's methodology. In studies concerning gist and mathematical cognition (Liberali et al., 2012; Reyna & Brust-Renck, 2020), numeracy scales were used. Objective numeracy scales are task-based measurement methods; however, they address only specific types of mathematical reasoning, namely those related to problem-solving and decision-making (cf. Reyna & Brust-Renck, 2020). Thus, as shown in multiple studies within FTT, gist should dominate in the performance of adults. Our study used a battery of tasks that tapped into more diverse mathematical skills, which may have resulted in greater involvement of verbatim trace.

We did not observe significant correlations between any memory parameter and the arithmetic fluency task (M4S), despite a significant correlation between overall numerical LTM performance and two of the task's scales (namely, multiplication and division). This might be due to the cumulative influence of both traces being strong enough to be observable at the level of the correlation between arithmetic fluency and overall LTM performance, but not when the influence of separate parameters is considered. However, this relationship requires further investigation.

Despite our predictions driven by Reyna and Brust-Renck's (2020) statements on the ANS, we found no significant association between verbatim or gist memory and performance in the dot comparison task. However, note that Reyna and Brust-Renck (2020) applied only symbolic numerical tasks and our study is the first attempt to investigate the relationship between the processing of non-symbolic numerosities and memory traces as distinguished in the FTT. This is crucial, as there is no consensus in numerical cognition literature regarding the processing scope of the ANS. While numerous researchers indeed tend to assume that the ANS handles comparisons of both non-symbolic and symbolic numerical magnitudes, as Weber's law drives them both (Hohol et al., 2017; see footnote 1, Hohol et al., 2020), others argue that the term ANS is restricted to the processing of non-symbolic numerosities, with the comparison of dot arrays being a prototypical experimental task (Gilmore et al., 2011; Halberda et al., 2008). In recent literature, the prevailing view seems to be that even most elementary instances of symbolic numerical processing (including Arabic symbols and numerals) fall outside the scope of the ANS (see, e.g., the Discrete Semantic System account; Krajcsi et al., 2022; Krajcsi & Kojouharova, 2023). For this reason, we employed a non-symbolic task that is widely recognized as a robust measure of ANS processing.

The fact that we did not find a correlation between performance on the dot comparison task and gist memory processes should be interpreted with caution. On one hand, rapid non-symbolic comparisons are likely the least memory-dependent of all the tasks we used. On the other hand, it is possible that the ANS (understood strictly as non-symbolic processing) is not closely related to gist processing, at least in adults. As we tested only adults in our study, it is possible that the association between ANS and verbatim and gist processing exists in childhood and disappears in further ontogeny. This speculation should be investigated in future cross-sectional or, better yet, longitudinal studies. To facilitate more direct comparisons between gist processing and ANS performance in the same non-symbolic modality, future research should also incorporate numerical non-symbolic memory tasks based on FTT.

The result that deserves to be also highlighted is that the math self-concept correlates significantly with overall numerical LTM performance and one of the conjoint recognition model's parameters, related to verbatim memory. The correlation has the same direction as for mathematical skills measures, further supporting self-evaluation of math skills as a useful measurement in studies on mathematical cognition, in line with the results of other studies in the field (e.g., Fagerlin et al., 2007). Moreover, as math self-concept correlates only with a memory process related to verbatim, it suggests that only the quality of verbatim numerical memory may affect one's math self-concept. To investigate this in a causal way, future studies should implement stronger, i.e., longitudinal or experimental, designs.

Last but not least, to investigate whether our findings of more prominent associations between gist-based processes and math skills did not stem from a methodological artifact, we tested differences in memory parameters between groups characterized by different levels of math skills, identified using a random forest cluster analysis algorithm. The observed differences are generally in line with the results of the classic correlational analysis. Clusters characterized by higher math skills showed a significantly higher probability of retrieving gist for both target and related new items. The identity process was also higher in this group, but only in Model 1. However, there was no significant difference between the low and medium math skill clusters. These findings once again suggest that the involvement of gist processes in adults' math skills is greater than that of verbatim processes (as two out of three gist-related parameters, and only one out of three verbatim-related parameters differed significantly). However, this does not change the fact that verbatim information remains important.

4.6. The study limits and future research directions

Drawing strong conclusions from our results is limited for several reasons. Firstly, as mentioned before, although all the behavioral tasks were counterbalanced among the participants, the mathematical self-concept was always reported at the end of the procedure. Thus, we cannot exclude the possibility that performance in tasks measuring objectively math performance and numerical memory influenced participants' self-assessments. This scenario is probable, as Peters (2020) states that self-reported confidence in math skills varies more between participants when they complete math performance tests before evaluating their own skills. Although this suggests that assessing self-reported math competence at the end of the procedure may be a better approach than doing so at the beginning, future studies should counterbalance the timing of math self-concept assessment to eliminate the influence of task order at the sample level.

Secondly, to quantify mathematical reasoning performance, we did not use a standardized test (though the math problems we used in our study had been used previously; Szczygieł & Hohol, 2024; Szczygieł & Sari, 2024), as none are available for adults in Poland. Standardized math reasoning tests should be included in future studies.

Thirdly, while we decided to use Gebuis and Reynvoet's (2011) dot comparison protocol to test performance in comparing non-symbolic numerosities, there are other protocols to list only Panamath (Halberda et al., 2008) and the accuracy scores in them did not show a significant correlation (Clayton et al., 2015b; see also Gilmore et al., 2011; Krajcsi et al., 2024). The choice of protocol is not merely a methodological nuance. Although both aim to investigate the ANS, they likely measure distinct cognitive constructs. Although, as already stated, the dot comparison task is the least memory-dependent of all tasks we used, future FTT-driven studies using Panamath may reveal another pattern of the association with gist/verbatim processes.

Fourthly, although the math skills we measured in our study—namely, comparing non-symbolic numerosities, arithmetic fluency, and math reasoning—have been considered prototypical in several studies (see Amland et al., 2025; De Smedt et al., 2013; Halberda et al., 2008; Schneider et al., 2017), we chose to exclude another fundamental skill, symbolic number comparison (Moyer & Landauer, 1967), to ensure a reasonable testing time. Testing it would be particularly important, as it would deliver evidence for the association between elementary numerical cognition and gist/verbatim processes more comparable to previous studies (Reyna & Brust-Renck, 2020). Also, our finding that a better memory of exact numbers is positively associated with a more positive math self-concept would encourage researchers to investigate the relationships between numerical memory and other mathematics-related dimensions of individual differences, such as math self-efficacy (Bong & Skaalvik, 2003) and trait math anxiety (Cipora et al., 2022; Richardson & Suinn, 1972).

Finally, we want to highlight that we recruited only adults and did not use purposive sampling based on mathematical skills. In the FTT-related literature, it is well-established that while adults mainly rely on gist memory, children primarily depend on verbatim memory (Reyna, 2012). Considering the developmental shift that occurs during adolescence, further studies should explore other age groups, as associations between math skills and LTM processes may differ significantly from the patterns we observed. The associations we found also suggest testing numerical memory performance in specific groups with either extremely low or extremely high math skills, particularly individuals with developmental dyscalculia and, conversely, those gifted mathematically. These groups may reveal effects that are blurred in the typical level group we tested (e.g., due to limited variance). FTT-driven investigation of numerical memory in individuals with mathematics learning problems is further justified by a longstanding line of research indicating that they manifest persistent problems with memorizing and retrieving elementary arithmetic facts (Geary, 2011). Also, the FTT has already delivered insights into LTM characteristics in developmental dyslexia (Obidziński & Nieznański, 2017, 2022), which is worth noting in the context of dyscalculia, as a discussion on shared mechanisms in these two conditions is ongoing (Banfi et al., 2022; Marks et al., 2023). To sum up, we generally recommend including a broader range of tasks, scales, and populations in future FTT-driven studies on the relationship between mathematical performance and the structure of LTM.

5. Conclusions

Our finding that better LTM performance is associated with better mathematical reasoning and greater arithmetic fluency in multiplication and division—but not non-symbolic numerosity processing—adds to the knowledge base on the relationship between memory and mathematical performance. Simultaneously, our FTT-driven investigation of more elementary memory processes reveals more complex patterns, as there are significant associations between memory parameters and math reasoning, but not arithmetic fluency or approximate number processing. Although through both correlational and cluster-based analyses we found overall evidence that both verbatim and gist memory traces are associated with math skills, the gist trace appears to play a more prominent role, which is consistent with one of the main assumptions of FTT. Our study highlights the explanatory power of FTT in the fine-grained investigation of mathematical cognition and emphasizes the importance of multinomial modeling in advancing our understanding of the connection between mathematical skills and basic cognitive processes.

Open practices statement

We report all data exclusions, all manipulations, and all measures in the study. The dataset and experimental procedure are available on the Open Science Framework (doi:10.17605/OSF.IO/ZTKQA). This study's design and its analysis were not pre-registered.

Availability of data and materials

The dataset and experimental procedure are available on the Open Science Framework (doi:10.17605/OSF.IO/ZTKQA).

CRediT authorship contribution statement

Michał Obidziński: Writing – review & editing, Writing – original draft, Visualization, Resources, Methodology, Formal analysis, Conceptualization. **Nina Bażela:** Writing – original draft, Resources, Methodology, Investigation, Conceptualization. **Mateusz Hohol:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization.

Consent to participate

Written informed consent was obtained from all participants included in the study.

Ethics approval

This study was performed in line with the principles of the Declaration of Helsinki and approved by the Rector's Committee on Research Ethics of the Jagiellonian University in Krakow (decision number 1027.0041.5.2024).

Funding

This scientific study was funded by the state budget through the Ministry of Science and Higher Education of Poland under the program "Science for Society II", project ID: NdS-II/SN/0332/2024/01; the total value of the project 1 000 000 PLN. Open access for this paper was funded by the Priority Research Area 'Society of the Future' under the Strategic Programme Excellence Initiative at the Jagiellonian University.

Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial, institutional, financial, or personal relationships that could be construed as a potential conflict of interest.

Acknowledgements

The authors thank the two anonymous reviewers for their insightful remarks, and Charles Brainerd, Valerie Reyna, Marek Nieznański, and Daniel W. Heck for their consultation on data analysis.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2025.106212.

Data availability

The dataset and experimental procedure are available on the Open Science Framework (https://doi.org/10.17605/OSF.IO/ZTKQA).

References

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders (5th, text rev. Ed.)*.

Amland, T., Grande, G., Scherer, R., Lervåg, A., & Melby-Lervåg, M. (2025). Cognitive factors underlying mathematical skills: A systematic review and meta-analysis. *Psychological Bulletin, 151*(1), 88–129. https://doi.org/10.1037/bul0000457

Ashcraft, M. H. (1982). The development of mental arithmetic: A chronometric approach. *Developmental Review, 2*(3), 213–236. https://doi.org/10.1016/0273-2297(82)90012-0

Ashcraft, M. H. (1992). Cognitive arithmetic: A review of data and theory. *Cognition, 44*(1), 75–106. https://doi.org/10.1016/0010-0277(92)90051-I

Banfi, C., Landerl, K., & Moll, K. (2022). Cognitive profiles and co-occurrence of dyslexia and dyscalculia. In M. A. Skeide (Ed.), *The Cambridge handbook of dyslexia and dyscalculia* (pp. 65–82). Cambridge University Press. https://doi.org/10.1017/9781108973595.006.

Barrouillet, P., & Fayol, M. (1998). From algorithmic computing to direct retrieval: Evidence from number and alphabetic arithmetic in children and adults. *Memory & Cognition, 26*(2), 355–368. https://doi.org/10.3758/BF03201146

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berch, D. B. (2005). Making sense of number sense: Implications for children with mathematical disabilities. *Journal of Learning Disabilities, 38*(4), 333–339. https://doi.org/10.1177/00222194050380040901

Bicego, M. (2019). K-random forests: A K-means style algorithm for random forest clustering. In *2019 international joint conference on neural networks (IJCNN)* (pp. 1–8). https://doi.org/10.1109/IJCNN.2019.8851820

Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review, 15*(1), 1–40. https://doi.org/10.1023/A:1021302408382

Brainerd, C., Reyna, V. F., Wright, R., & Mojardin, A. (2003). Recollection rejection: False-memory editing in children and adults. *Psychological Review, 110*(4), 762. https://doi.org/10.1037/0033-295x.110.4.762

Brainerd, C. J., Bialer, D. M., & Chang, M. (2022). Fuzzy-trace theory and false memory: Meta-analysis of conjoint recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(11), 1680–1697. https://doi.org/10.1037/xlm0001040

Brainerd, C. J., & Gordon, L. (1994). Development of verbatim and gist memory for numbers. *Developmental Psychology, 30*(2), 163. https://doi.org/10.1037//0012-1649.30.2.163

Brainerd, C. J., & Reyna, V. F. (1988). Generic resources, reconstructive processing, and children's mental arithmetic. *Developmental Psychology, 24*(3), 324. https://doi.org/10.1037//0012-1649.24.3.324

Brainerd, C. J., & Reyna, V. F. (1990a). Inclusion illusions: Fuzzy-trace theory and perceptual salience effects in cognitive development. *Developmental Review, 10*(4), 365–403. https://doi.org/10.1016/0273-2297(90)90020-5

Brainerd, & Reyna, V. F. (1990b). Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review, 10*(1), 3–47. https://doi.org/10.1016/0273-2297(90)90003-m

Brainerd, C. J., & Reyna, V. F. (1993). Memory independence and memory interference in cognitive development. *Psychological Review, 100*(1), 42. https://doi.org/10.1037/0033-295X.100.1.42

Brainerd, C. J., & Reyna, V. F. (2004). Fuzzy-trace theory and memory development. *Developmental Review, 24*(4), 396–439. https://doi.org/10.1016/j.dr.2004.08.005

Brainerd, C. J., Reyna, V. F., & Aydin, C. (2010). Remembering in contradictory minds: Disjunction fallacies in episodic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 711–735. https://doi.org/10.1037/a0018995

Brainerd, C. J., Reyna, V. F., & Mojardin, A. H. (1999). Conjoint recognition. *Psychological Review, 106*(1), 160–179. https://doi.org/10.1037/0033-295X.106.1.160

Brainerd, C. J., Wright, R., Reyna, V. F., & Mojardin, A. H. (2001). Conjoint recognition and phantom recollection. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(2), 307. https://doi.org/10.1037//0278-7393.27.2.307

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bull, R., & Scerif, G. (2001). Executive functioning as a predictor of children's mathematics ability: Inhibition, switching, and working memory. *Developmental Neuropsychology, 19*(3), 273–293. https://doi.org/10.1207/s15326942dn1903_3

Butterworth, B. (2011). Dyscalculia: From brain to education. *Science, 334*(6057), 1049–1053. https://doi.org/10.1126/science.334.6057.761-b

Calderón-Tena, C. O., & Caterino, L. C. (2016). Mathematics learning development: The role of long-term retrieval. *International Journal of Science and Mathematics Education, 14*(7), 1377–1385. https://doi.org/10.1007/s10763-015-9655-0

Campbell, J. I. D. (1987). Network interference and mental multiplication. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(1), 109–123. https://doi.org/10.1037/0278-7393.13.1.109

Campbell, J. I. D. (Ed.). (2005). *The handbook of mathematical cognition*. Psychology Press.

Campbell, J. I. D., & Xue, Q. (2001). Cognitive arithmetic across cultures. *Journal of Experimental Psychology: General, 130*(2), 299–315. https://doi.org/10.1037/0096-3445.130.2.299

Chapman, M., & Lindenberger, U. (1988). Functions, operations, and decalage in the development of transitivity. *Developmental Psychology, 24*(4), 542. https://doi.org/10.1037/0012-1649.24.4.542

Cipora, K., Santos, F. H., Kucian, K., & Dowker, A. (2022). Mathematics anxiety—Where are we and where shall we go? *Annals of the New York Academy of Sciences, 1513*(10), 10–20. https://doi.org/10.1111/nyas.14770

Clayton, S., Gilmore, C., & Inglis, M. (2015a). *Data and stimuli for "dot comparison stimuli are not all alike: The effect of different visual controls on ANS measurement" (version 2) [Dataset]*. Figshare. https://doi.org/10.6084/m9.figshare.1546747.v2

Clayton, S., Gilmore, C., & Inglis, M. (2015b). Dot comparison stimuli are not all alike: The effect of different visual controls on ANS measurement. *Acta Psychologica, 161*, 177–184. https://doi.org/10.1016/j.actpsy.2015.09.007

Cohen Kadosh, R., & Dowker, A. (Eds.). (2015). *The Oxford handbook of numerical cognition*. Oxford University Press.

Cooney, J. B., Swanson, H. L., & Ladd, S. F. (1988). Acquisition of mental multiplication skill: Evidence for the transition between counting and retrieval strategies. *Cognition and Instruction, 5*(4), 323–345. https://doi.org/10.1207/s1532690xci0504_5

De Smedt, B. (2022). Individual differences in mathematical cognition: A Bert's eye view. *Current Opinion in Behavioral Sciences, 46*, Article 101175. https://doi.org/10.1016/j.cobeha.2022.101175

De Smedt, B., Noël, M.-P., Gilmore, C. K., & Ansari, D. (2013). How do symbolic and non-symbolic numerical magnitude processing skills relate to individual differences in children's mathematical skills? A review of evidence from brain and behavior. *Trends in Neuroscience and Education, 2*(2), 48–55. https://doi.org/10.1016/j.tine.2013.06.001

Dehaene, S. (2011). *The number sense*. Oxford University Press.

DeStefano, D., & LeFevre, J. (2004). The role of working memory in mental arithmetic. *European Journal of Cognitive Psychology, 16*(3), 353–386. https://doi.org/10.1080/09541440244000328

Dowker, A. (2019). *Individual differences in arithmetic: Implications for psychology, neuroscience and education*. Routledge. https://doi.org/10.4324/9781315755526-2

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models. *Zeitschrift Für Psychologie / Journal of Psychology, 217*(3), 108–124. https://doi.org/10.1027/0044-3409.217.3.108

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the subjective numeracy scale. *Medical Decision Making, 27*(5), 672–680. https://doi.org/10.1177/0272989X07304449

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fournier, S., Larigauderie, P., & Gaonac'h, D. (2004). Exploring how the central executive works: A search for independent components. *Psychologica Belgica, 44*(3). https://doi.org/10.5334/pb.1022

Friso-van den Bos, I., van der Ven, S. H. G., Kroesbergen, E. H., & van Luit, J. E. H. (2013). Working memory and mathematics in primary school children: A meta-analysis. *Educational Research Review, 10*, 29–44. https://doi.org/10.1016/j.edurev.2013.05.003

Geary, D. C. (2011). Consequences, characteristics, and causes of mathematical learning disabilities and persistent low achievement in mathematics. *Journal of Developmental & Behavioral Pediatrics, 32*(3), 250. https://doi.org/10.1097/DBP.0b013e318209edef

Geary, D. C., Hoard, M. K., Byrd-Craven, J., & DeSoto, M. C. (2004). Strategy choices in simple and complex addition: Contributions of working memory and counting knowledge for children with mathematical disability. *Journal of Experimental Child Psychology, 88*(2), 121–151. https://doi.org/10.1016/j.jecp.2004.03.002

Gebuis, T., & Reynvoet, B. (2011). Generating nonsymbolic number stimuli. *Behavior Research Methods, 43*(4), 981–986. https://doi.org/10.3758/s13428-011-0097-5

Gilmore, C., Attridge, N., Clayton, S., Cragg, L., Johnson, S., Marlow, N., … Inglis, M. (2013). Individual differences in inhibitory control, not non-verbal number acuity, correlate with mathematics achievement. *PLoS One, 8*(6), Article e67374. https://doi.org/10.1371/journal.pone.0067374

Gilmore, C., Attridge, N., & Inglis, M. (2011). Measuring the approximate number system. *The Quarterly Journal of Experimental Psychology, 64*(11), 2099–2109. https://doi.org/10.1080/17470218.2011.574710

Gilmore, C., Göbel, S. M., & Inglis, M. (2018). *An introduction to mathematical cognition*. Routledge.

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature, 455*(7213), 665–668. https://doi.org/10.1038/nature07246

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods, 50*, 264–284. https://doi.org/10.3758/s13428-017-0869-7

Hohol, M. (2020). *Foundations of geometric cognition*. Routledge. https://doi.org/10.4324/9780429056291

Hohol, M., Cipora, K., Willmes, K., & Nuerk, H.-C. (2017). Bringing back the balance: Domain-general processes are also important in numerical cognition. *Frontiers in Psychology, 8*, 499. https://doi.org/10.3389/fpsyg.2017.00499

Hohol, M., Willmes, K., Nęcka, E., Brożek, B., Nuerk, H.-C., & Cipora, K. (2020). Professional mathematicians do not differ from others in the symbolic numerical distance and size effects. *Scientific Reports, 10*(11531). https://doi.org/10.1038/s41598-020-68202-z

JASP Team. (2024). *JASP* (version 0.19.3) [computer software]. https://jasp-stats.org/.

Kiefer, M., Apel, A., & Weisbrod, M. (2002). Arithmetic fact retrieval and working memory in schizophrenia. *Schizophrenia Research, 53*(3), 219–227. https://doi.org/10.1016/S0920-9964(01)00152-9

Krajcsi, A., Chesney, D., Cipora, K., Coolen, I., Gilmore, C., Inglis, M., Libertus, M., Nuerk, H.-C., Simms, V., & Reynvoet, B. (2024). Measuring the acuity of the approximate number system in young children. *Developmental Review, 72*, Article 101131. https://doi.org/10.1016/j.dr.2024.101131

Krajcsi, A., & Kojouharova, P. (2023). Stimulus frequency alone can account for the size effect in number comparison. *Acta Psychologica, 232*, Article 103817. https://doi.org/10.1016/j.actpsy.2022.103817

Krajcsi, A., Kojouharova, P., & Lengyel, G. (2022). Processing symbolic numbers: The example of distance and size effects. In J. Gervain, G. Csibra, & K. Kovács (Eds.), *A life in cognition: Studies in cognitive science in honor of Csaba Pléh* (pp. 379–394). Springer International Publishing. https://doi.org/10.1007/978-3-030-66175-5_27.

Lakoff, G., & Núñez, R. (2000). *Where mathematics comes from*. Basic Books.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making, 25*(4), 361–381. https://doi.org/10.1002/bdm.752

Loenneker, H. D., Cipora, K., Artemenko, C., Soltanlou, M., Bellon, E., De Smedt, B., … Huber, J. F. (2024). Math4Speed: A freely available measure of arithmetic fluency. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale.*. https://doi.org/10.1037/cep0000347

Logie, R. H. (2016). Retiring the central executive. *Quarterly Journal of Experimental Psychology, 69*(10), 2093–2109. https://doi.org/10.1080/17470218.2015.1136657

Makowski, D., Ben-Shachar, M., Patil, I., & Lüdecke, D. (2020). Methods and algorithms for correlation analysis in R. *Journal of Open Source Software, 5*(51), 2306. https://doi.org/10.21105/joss.02306

Marks, R. A., Pollack, C., Meisler, S. L., D'Mello, A. M., Centanni, T. M., Romeo, R. R., … Christodoulou, J. A. (2023). Neurocognitive mechanisms of co-occurring math difficulties in dyslexia: Differences in executive function and visuospatial processing. *Developmental Science, e13443*. https://doi.org/10.1111/desc.13443

Marsh, H. W., & O'Neill, R. (1984). Self description questionnaire III: The construct validity of multidimensional self-concept ratings by late adolescents. *Journal of Educational Measurement, 21*(2), 153–174. https://doi.org/10.1111/j.1745-3984.1984.tb00227.x

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

McLean, J. F., & Hitch, G. J. (1999). Working memory impairments in children with specific arithmetic learning difficulties. *Journal of Experimental Child Psychology, 74*(3), 240–260. https://doi.org/10.1006/jecp.1999.2516

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature, 215*(5109), 1519–1520. https://doi.org/10.1038/2151519a0

Obidziński, M. (2021). Response frequencies in the conjoint recognition memory task as predictors of developmental dyslexia diagnosis: A decision-trees approach. *Dyslexia, 27*(1), 50–61. https://doi.org/10.1002/dys.1655

Obidziński, M., Bażela, N., & Hohol, M. (2024). *Mathematical performance and long-term memory for numbers [Dataset]*. Open Science Framework. https://doi.org/10.17605/OSF.IO/ZTKQA

Obidziński, M., & Nieznański, M. (2017). False memory for orthographically versus semantically similar words in adolescents with dyslexia: A fuzzy-trace theory perspective. *Annals of Dyslexia, 67*(3), 318–332. https://doi.org/10.1007/s11881-017-0146-6

Obidziński, M., & Nieznański, M. (2022). Context and target recollection for words and pictures in young adults with developmental dyslexia. *Frontiers in Psychology, 13*, Article 993384. https://doi.org/10.3389/fpsyg.2022.993384

OECD. (2016). *Skills matter: Further results from the survey of adult skills*. OECD. https://doi.org/10.1787/9789264258051-en

Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology, 86*(2), 193–203. https://doi.org/10.1037/0022-0663.86.2.193

Peng, P., Namkung, J., Barnes, M., & Sun, C. (2016). A meta-analysis of mathematics and working memory: Moderating effects of working memory domain, type of mathematics skill, and sample characteristics. *Journal of Educational Psychology, 108*(4), 455. https://doi.org/10.1037/edu0000079

Peters, E. (2020). *Innumeracy in the wild: Misunderstanding and misusing numbers*. Oxford University Press.

Peters, E., & Bjalkebring, P. (2015). Multiple numeric competencies: When a number is not just a number. *Journal of Personality and Social Psychology, 108*(5), 802–822. https://doi.org/10.1037/pspp0000019

Posit Team. (2025). *RStudio: Integrated development environment for R [computer software]*. Posit Software, PBC. http://www.posit.co/.

R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing https://www.R-project.org/.

Revelle, W. (2024). *Psych: Procedures for psychological, psychometric, and personality research (version 2.4.6.26) [Computer software]*. Northwestern University. https://cran.r-project.org/package=psych.

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision making, 7*(3), 332–359. https://doi.org/10.1017/S1930297500002291

Reyna, V. F., & Brainerd, C. J. (1990). Fuzzy processing in transitivity development. *Annals of Operations Research, 23*, 37–63. https://doi.org/10.1007/bf02204838

Reyna, V. F., & Brainerd, C. J. (1991). Fuzzy-trace theory and children's acquisition of mathematical and scientific concepts. *Learning and Individual Differences, 3*(1), 27–59. https://doi.org/10.1016/1041-6080(91)90003-J

Reyna, V. F., & Brainerd, C. J. (1993). Chapter 3 fuzzy memory and mathematics in the classroom. In *, Vol. 100. Advances in psychology* (pp. 91–119). Elsevier. https://doi.org/10.1016/S0166-4115(08)61096-1.

Reyna, V. F., & Brainerd, C. J. (2023). Numeracy, gist, literal thinking and the value of nothing in decision making. *Nature Reviews Psychology, 2*(7), 421–439. https://doi.org/10.1038/s44159-023-00188-7

Reyna, V. F., & Brust-Renck, P. G. (2020). How representations of number and numeracy predict decision paradoxes: A fuzzy-trace theory approach. *Journal of Behavioral Decision Making, 33*(5), 606–628. https://doi.org/10.1002/bdm.2179

Reyna, V. F., Corbin, J. C., Weldon, R. B., & Brainerd, C. J. (2016). How fuzzy-trace theory predicts true and false memories for words, sentences, and narratives. *Journal of Applied Research in Memory and Cognition, 5*(1), 1–9. https://doi.org/10.1016/j.jarmac.2015.12.003

Reyna, V. F., & Kiernan, B. (1994). Development of gist versus verbatim memory in sentence recognition: Effects of lexical familiarity, semantic content, encoding instructions, and retention interval. *Developmental Psychology, 30*(2), 178–191. https://doi.org/10.1037/0012-1649.30.2.178

Reyna, V. F., Müller, S. M., & Edelson, S. M. (2023). Critical tests of fuzzy trace theory in brain and behavior: Uncertainty across time, probability, and development. *Cognitive, Affective, & Behavioral Neuroscience, 23*(3), 746–772. https://doi.org/10.3758/s13415-022-01058-0

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135*(6), 943–973. https://doi.org/10.1037/a0017327

Richardson, F. C., & Suinn, R. M. (1972). The mathematics anxiety rating scale: Psychometric data. *Journal of Counseling Psychology, 19*(6), 551–554. https://doi.org/10.1037/h0033456

Rossi, S., Xenidou-Dervou, I., Simsek, E., Artemenko, C., Daroczy, G., Nuerk, H., & Cipora, K. (2022). Mathematics–gender stereotype endorsement influences mathematics anxiety, self-concept, and performance differently in men and women. *Annals of the New York Academy of Sciences, 1513*(1), 121–139. https://doi.org/10.1111/nyas.14779

Schmidt, O., Erdfelder, E., & Heck, D. W. (2023). How to develop, test, and extend multinomial processing tree models: A tutorial. *Psychological Methods*. https://doi.org/10.1037/met0000561

Schneider, M., Beeres, K., Coban, L., Merz, S., Susan Schmidt, S., Stricker, J., & De Smedt, B. (2017). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: A meta-analysis. *Developmental Science, 20*(3), 1–16. https://doi.org/10.1111/desc.12372

Siegler, R. S. (1988). Strategy choice procedures and the development of multiplication skill. *Journal of Experimental Psychology: General, 117*(3), 258–275. https://doi.org/10.1037/0096-3445.117.3.258

Singer, M., & Remillard, G. (2008). Veridical and false memory for text: A multiprocess analysis. *Journal of Memory and Language, 59*(1), 18–35. https://doi.org/10.1016/j.jml.2008.01.005

Skeide, M. A. (Ed.). (2022). *The Cambridge handbook of dyslexia and dyscalculia*. Cambridge University Press.

Stahl, C., & Klauer, K. C. (2009). Measuring phantom recollection in the simplified conjoint recognition paradigm. *Journal of Memory and Language, 60*(1), 180–193. https://doi.org/10.1016/j.jml.2008.08.001

Szczygieł, M., & Hohol, M. (2024). *The gender gap in math anxiety (and in the link between math anxiety and math performance) is not so salient when other anxieties are controlled for*. PsyArxiv. https://doi.org/10.31234/osf.io/5trew. preprint.

Szczygieł, M., & Sari, M. H. (2024). The relationship between numerical magnitude processing and math anxiety, and their joint effect on adult math performance, varied by indicators of numerical tasks. *Cognitive Processing, 25*(3), 421–442. https://doi.org/10.1007/s10339-024-01186-0