The role of metacognition in monitoring performance and regulating learning in early readers

Ioanna Taouki[1], Marie Lallier[1,2], David Soto[1,2]

[1]Basque Center on Cognition, Brain and Language, San Sebastian, Spain

[2] Ikerbasque, Basque Foundation for Science, Bilbao, Spain

🖂 Ioanna Taouki

i.taouki@bcbl.eu

ORCID: 0000-0003-4464-4708

Accepted in *Metacognition and Learning*: 8 February 2022

Abstract

Metacognition refers to the capacity to reflect upon our own cognitive processes. Its contribution to reading development, when children start building their orthographic lexicon, still remains unknown. Here, we evaluate the metacognitive efficiency of children aged between 6 and 7 years old (N=60) in 5 experimental tasks; four linguistic tasks assessing orthographic lexical processing and a non-linguistic task unrelated to reading skills. First, we investigated how metacognition on the experimental tasks related to standardised on-paper reading performance, hence participants' general reading level. Second, we assessed whether these developing readers recruited common metacognitive mechanisms across the different experimental tasks. Third, we explored whether metacognition in this early stage was related to the longitudinal improvement in performance on a linguistic vs a non-linguistic task. No association was found between students' metacognition in the reading-related tasks and performance on the standardised reading tests, notwithstanding first-order performance correlated across these tasks. Remarkably, some negative correlations were noted between students' metacognitive ability in one task and task performance in another task. Moreover, we found some evidence consistent with shared metacognitive mechanisms for monitoring performance across tasks. Finally, metacognitive ability significantly predicted children's performance improvement across domains 10 months later. These results suggest that the development of metacognitive processing may be dissociated to some extent from reading-related linguistic abilities during the early stages of formal education. Nevertheless, it may play a fundamental role in guiding students' learning across domains. These data highlight the importance of creating educational programs fostering students' metacognition as a long-term learning tool.

Keywords - confidence, development, metacognition, reading, learning

The role of metacognition in monitoring performance and regulating learning in early readers

Metacognition refers to the ability of an individual to reflect on their own cognition and behaviour, e.g., to track the correctness of one's thoughts, actions, and behavioural responses (Metcalfe & Shimamura, 1994) across multiple task contexts (Narens, 1990). Metacognition was first introduced as a term by Flavell (1979), defined as our ability to "think about thinking" (Flavell, 1979). A widely used metacognitive framework in educational settings has been proposed by Efklides (2008, 2011), who stratified metacognition in: a) *metacognitive knowledge* or *metacognitive awareness*, referring to knowledge about our own and other people's cognitive processes, b) *metacognitive experiences*, which include procedural knowledge, feelings, and judgments generated in on-line task performance (analogue of Flavell's procedural metacognition, see: Flavell, 1979), and c) *metacognitive skills*, referring to the intentional employment of cognitive strategies in order to regulate cognition and guide behaviour (Efklides, 2008, 2011).

In the present study, we will focus on the concept of *procedural metacognition/metacognitive experiences* component, which relates directly to online processes associated with task performance. Influential models of procedural metacognition propose that it is mediated by an interaction between an object-level process (e.g., a reading or perceptual task, namely, the object-level or *type-1* performance) and a second-order process (e.g., the meta-level or *type-2* performance). The meta-level component monitors the first-order process and, when cognition fails (i.e., following an error), exerts control processes in order to promote adaptive behaviour (Koriat & Goldsmith, 1996; Nelson, 1990). These two processes have been referred to as "*metacognitive monitoring*" and "*metacognitive control*", respectively.

The current study aims to understand the role of metacognitive monitoring in the early stages of reading acquisition and specifically in the development of students' orthographic lexical processing skills, which refers to the orthographic knowledge and processing of whole word forms (visual word recognition), and is essential for individuals to develop fast and fluent reading (Ehri, 2014; Frith, 1985). Orthographic lexical processing requires that individuals: a) successfully process visually letter sequences at a glance (operationalized as the “Visual Attention (VA) Span”, see: Bosse et al., 2007; Bosse & Valdois, 2009), b) detect orthographic statistical regularities in words, which strengthens orthographic traces in memory for speeding up their later access (measured in statistical learning tasks, see: Arciuli & Simpson, 2012; Mano & Kloos, 2018; Boukadi et al., 2016), and c) activate whole-word orthographic representations and their respective quality from memory (measured by the orthographic lexical decision task, see: Chetail, 2017; Ginestet et al., 2019). The role of metacognition in students’ orthographic lexical processing will be studied here through the measurement of these three skills, which have been considered essential in the process of visual word recognition and for allowing the development of automatic or sight word reading. Increase of an individuals’ automaticity in visual word recognition has been suggested to free up cognitive resources, which can in turn be employed in reading comprehension (Verhoeven et al., 2019). For this reason, automaticity in word recognition has been considered to be an important predictor of academic achievement in reading (Cunningham & Stanovich, 1997).

Lexical orthographic processing has been suggested to develop from the beginning of reading acquisition (Martinet et al., 2004). However, it is not known how and whether metacognition contributes to the development of lexical orthographic knowledge and its cognitive prerequisites during the early stages of formal reading tuition. We also do not know the extent to which metacognitive processes predict or explain individual variability in reading skills or how and if metacognitive ability has a long-term effect on learning during the first year of formal reading instruction, and if it has, whether this is specific to linguistic tasks. These questions will be explored in the current study in the context of linguistic tasks assessing orthographic lexical processing and, by extension, in the context of a perceptual, non-linguistic task involving emotion recognition. Children’s metacognition across the different tasks is assessed during the first year of primary school, i.e., when students first receive formal reading instruction, through the use of retrospective trial-by-trial confidence judgments (tasks assessing metacognitive ability will be referred to as experimental tasks). Signal detection theory measures are used to estimate type-1 task performance and type-2 metacognitive efficiency. Metacognitive efficiency is a bias-free index of metacognition that: a) controls for individual confidence biases, defined as the tendency of a participant to use higher or lower confidence ratings in a cognitive task, which is particularly critical when assessing metacognitive function in young children because they typically show an overconfidence bias (Finn & Metcalfe, 2014), and b) takes into account the type-1 sensitivity bias, meaning that participants who perform better in the type-1 task may erroneously appear to also have better metacognitive sensitivity compared to their peers (Fleming & Lau, 2014; Masson & Rotello, 2009; for details see also Method section). Finally, standardized reading tasks will be used to assess participants' *general reading level*, indexing their on-paper reading performance in more ecological settings, in terms of accuracy and speed.

Within the above framework, the present study addresses the following research questions:

Is metacognitive efficiency in linguistic tasks assessing orthographic processing associated with individual general reading level?

Previous studies in typically developing populations pinpoint the role of metacognition in reading comprehension, especially after middle childhood, when children usually develop fluent reading (e.g., Roebers et al., 2009; de Bruin et al., 2011). In addition to reading, metacognitive ability has been considered fundamental for learning in other domains such as mathematics, memory and perception (Kuhn, 2000; Schoenfeld, 2016). Educational studies suggest that individuals with higher performance monitoring skills tend to be better learners (Metcalfe & Kornell, 2007; Rawson et al., 2011). The present study is, to our knowledge, the first to examine the role of metacognition in skills that develop early on during reading development, such as visual word recognition and orthographic processing. Moreover, to date, research in development is mainly based on self-report questionnaires and there is a lack of robust metrics of metacognition that can be comparable across tasks.

Only recently, Bellon et al.'s (2019) study in early childhood (7-9 years of age) showed that metacognitive processing, in the context of spelling and arithmetic task performance, correlated with standardised tasks examining the level of performance in these domains (Bellon et al., 2019). Vo et al.'s (2014) examined metacognition in the numerical and emotion domain in the ages of 5-8, and found that metacognitive sensitivity on numerical judgments was positively correlated with math ability, but this was not the case with metacognitive sensitivity in the emotion domain (Vo et al., 2014). However, the abovementioned studies used metacognitive indexes which do not avoid confounding effects stemming from confidence bias (Bellon et al., 2019) or type-1 performance (Vo et al., 2014).

In the present study, we hypothesize that, if metacognitive efficiency, a free of biases metacognitive index, is related to reading and its orthographic prerequisites, participants with higher metacognitive efficiency on the experimental tasks indexing orthographic knowledge (e.g., VA span) will exhibit higher type-1 performance across related tasks (e.g., visual statistical learning and orthographic lexical decision), and also higher performance in the standardised reading tasks, assessing participants' general reading level.

Is metacognitive efficiency supported by domain-general mechanisms?

The question of whether metacognition is supported by domain-general or specific mechanisms remains highly debated. A domain-general model predicts that an individual with poor/good metacognitive ability in one domain (e.g., orthographic lexical processing), will have poor/good metacognitive skill in a different unrelated domain (e.g., recognizing emotions), hence supporting the view that a single metacognitive system monitors performance across different domains. Investigating how this system works during the first year of reading acquisition will help researchers and educators understand whether metacognitive ability can be boosted holistically, i.e., across domains, or whether the development of metacognitive strategies related to reading acquisition should be assisted separately.

Developmental studies assessing the domain-generality/specificity issue are still very limited. The few existing studies suggest that metacognitive processing is domain-specific in early childhood and becomes domain-general later during development. However, the age in which this shift takes place during development still remains under debate, ranging between 8 and 10 years of age among the different studies (Geurten et al., 2018; Lyons & Ghetti, 2010; Vo et al., 2014). However, these studies are limited in the usage of metacognitive indexes that are sensitive to one's confidence biases or that may be confounded by the level of type-1 task performance (see Fleming & Lau, 2014).

The present study will address the domain-generality issue by using state-of-the-art measures of metacognitive efficiency (see: Fleming & Lau, 2014), in combination with Bayesian correlational analysis, which provides evidence in respect to the null hypothesis. As previous literature has suggested that domain-specific mechanisms are likely to support metacognition early in development before the age of 8 (Geurten et al., 2018; Lyons & Ghetti, 2010; Vo et al., 2014), we hypothesize that metacognitive efficiency in the experimental tasks will likely not correlate across tasks in our children participants attending Grade 1 (6-7 years old).

Can metacognitive efficiency predict future learning in different domains?

A key question beyond the interplay between metacognition and other cognitive systems, is the role of metacognition in regulating one's learning across time. Understanding whether the efficiency of children's metacognitive system predicts the development of their cognitive abilities over time will shed light on the importance of enhancing this ability in classroom and clinical settings. Only few longitudinal studies have assessed children's language abilities and theory of mind before entering primary school in connection to their metacognition in the memory domain during the early years of primary school (Lecce et al., 2015; Lockl & Schneider, 2007). Both studies revealed a relationship between theory of mind in pre-school ages and metacognition evaluated in the first years of primary school. This relationship was independent of students' language skills. However, to evaluate this relationship these studies were based on second-order false belief and metamemory tasks whose scoring did not control for confidence or type-1 biases of participants.

Roebers' research group (2017) was among the first ones to longitudinally assess online metacognitive ability in primary school students using measures of type-1 and type-2 performance. They found that performance in a spelling task predicted not only participants' future performance in the task, but also future metacognitive performance, suggesting that type-1 performance can be a driving force in the development of metacognition. However, metacognitive skill was not found to predict future spelling performance (Roebers & Spiess, 2017). On the contrary, Rinne and Mazzocco (2014) suggested that type-2 performance is fueling improvements in future type-1 performance (Rinne & Mazzocco, 2014). These studies also suffer from the lack of control for participants' type-1 performance in the estimation of one's metacognitive skill and other confounding effects of type-2 biases, as noted above.

The present study goes beyond the prior work investigating the connection between metacognition and improvement in participant's performance across time by (i) using bias-free measures of metacognition that also control for type-1 performance and (ii) by providing an objective quantification of learning skill in a linguistic task (orthographic lexical decision) and a non-linguistic task (emotion recognition) across two different time points 10 months apart. The lexical decision task directly assesses participants' orthographic knowledge and reading performance. Hence, the inclusion of the non-linguistic task allows us to determine whether any linkage between metacognition and learning is specific to the linguistic domain. We hypothesize that, if metacognition is an essential skill for learning as previously suggested by educational studies, participants' metacognitive efficiency at time point 1 of assessment will predict improvement in participants' type-1 performance between time 1 and time 2. If the mediating role of metacognition in learning is domain-general, then we expect metacognition to predict longitudinal changes in type-1 performance independently of the domain studied (i.e., linguistic or not).

Method

Participants

Sixty-nine children aged between 6 and 7 years (mean age (±SD): 6.67 ± 0.36, 28 girls) attending Grade 1 (January 2019-March 2019) were initially recruited for this study. They were native Spanish speakers from an urban school in the center of Vitoria, Spain. Participants were recruited from the same urban school to try to match children in their socioeconomic status. However, we did not obtain objective measures of this (e.g., degree of parents' education) as it was out of the scope of the present study. Nine participants were excluded from the analyses due to missing data in one or more of the tasks, due to inability to read, or due to low non-verbal IQ[1].

Participants were asked to participate voluntarily and fully informed consent forms were obtained from the legal tutors of the minors prior to the study. The study was approved by the BCBL Ethics Board and the Bioethics Commission of the University of Barcelona. A subgroup of participants (N=40) of the same cohort was re-tested 10 months later in Grade 2 (October 2019-February 2020) in the orthographic lexical decision task and the emotion recognition task. During retesting, participants underwent simultaneous EEG recordings. The electrophysiological data will be reported as a part of a separate study. Note that it was not possible to re-test the full sample in the context of the EEG study. However, we ensured that the variability in type-1 task performance and type-2 metacognitive efficiency in the sub-group included did not differ from that of the sub-group that did not participate in the EEG study.

[1] Atypically low IQ could be a sign of global developmental delay. We decided to exclude these few participants because in these particular cases it would not be clear whether poor reading skills stem from poor attentional or orthographic skills, or whether they derive from broader learning, cognitive and reasoning differences, and not only reading-related skills. Moreover, in order to perform the type-2 (i.e., confidence rating) task, the participant needs to have access to the relevant type-1 signal. Hence, we consider that kids who were not yet able

to read words, not even by decoding them, would respond randomly both in the type-1 and type-2 task in the reading-related experimental tasks (i.e., VA span, lexical decision task) and were thereby also excluded from the analysis.

Experimental procedure

Participants performed four linguistic tasks related to orthographic lexical processing and orthographic knowledge, and one non-linguistic task in order to assess any domain-specific effects. The tasks were programmed in PsychoPy version 1.83.4(Peirce J & Macaskill, 2019) and were part of a larger task battery, which was administered in counterbalanced order during three sessions of 1 hour in participants' school. The schedule was organized in agreement with the teachers and the director of the school. Each participant was tested individually by the first author in a soundproof room of the school to minimize noise. Participants sat in front of the computer screen at a distance of 70cm without head constraint. In addition to these tasks, participants were also administered two standardised reading tests (word and pseudoword reading) assessing general reading level, as well as a control task aimed at measuring their non-verbal reasoning abilities (non-verbal IQ).

Experimental tasks

Linguistic tasks related to orthographic lexical processing

Visual Attention (VA) Span Task.

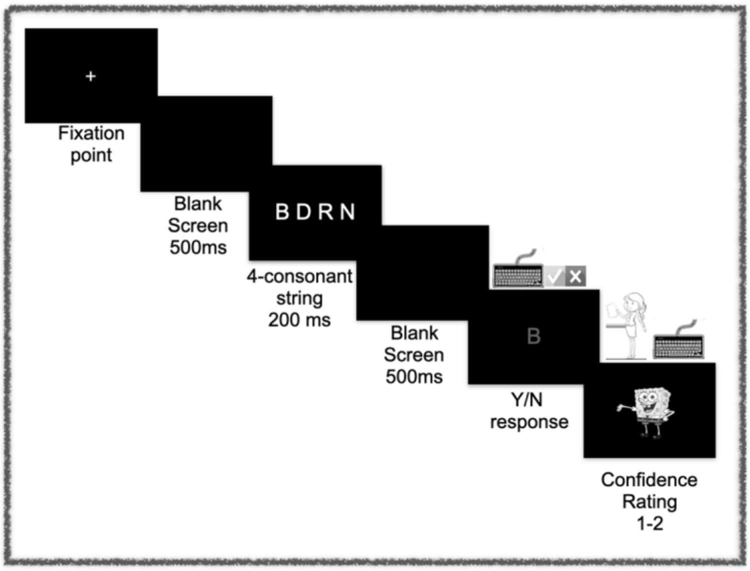
Visual stimuli were composed of 103 distinct 4-consonant strings (e.g., D P N L), created by the use of 13 consonants (B, D, F, G, H, K, L, M, N, P, R, S, T). The present task was designed following well-established experimental protocols used in previous studies to measure VA span (Bosse et al., 2007; Bosse & Valdois, 2009). The criteria for the selection of the consonant strings were that no repetition of the consonants is permitted, that strings do not contain grapheme clusters existing in Spanish language (e.g., ST, TR) and that they do not form word skeletons in Spanish (e.g., P L N T for "PLANETA"; meaning planet in Spanish).

Visual stimuli of the present assessment were displayed on the computer screen using white upper-case Arial font with a black background. The 4-consonant strings occupied a space on the screen of min 5.3° and max 5.55° degrees of visual angle, with a 1.2 centre-to-centre distance between adjacent letters. This stimuli size was chosen in order to minimize lateral masking effects. Each trial started with the onset of a central fixation cross at the center of the screen until the participant reported that they were ready to start the trial. The target 4-consonant string was displayed for 200 ms at the center of the screen. After the appearance of the string and following a blank screen of a 100 ms duration, a single target consonant was displayed in red font either slightly below or above the position that the 4-consonant string occupied (counterbalanced between trials). Participants were then asked to give a YES/NO response on whether the single consonant was part of the string or not by pressing keys on the keyboard labeled as ✓ or Ⅹ. Subsequently, in each trial, participants were asked to rate their confidence on having given a correct response or not. Two options (*1: I have doubts on whether my response was correct or not, 2: I'm sure my response was correct*) were given to the participants that were explicitly explained in each trial. This was recorded through an external keyboard by the researcher (see Fig. 1).

A binary 2-points scale of confidence was elected in this and all the experimental tasks described below for the following reasons: a) to simplify the confidence scale for children in early childhood so that they are able to understand and use with ease the different points of the scale (see for e.g., Lyons & Ghetti, 2011 or Filevich et al., 2020) and b) because the signal detection theoretic models benefit from the simplicity of using binary confidence responses to compute metacognitive efficiency estimates; this is also relevant when the amount of trials that can be collected is limited (i.e., with children populations) because including more confidence ratings would also imply the acquisition of more trials per participant.

Figure 1

Task design of the VA span task



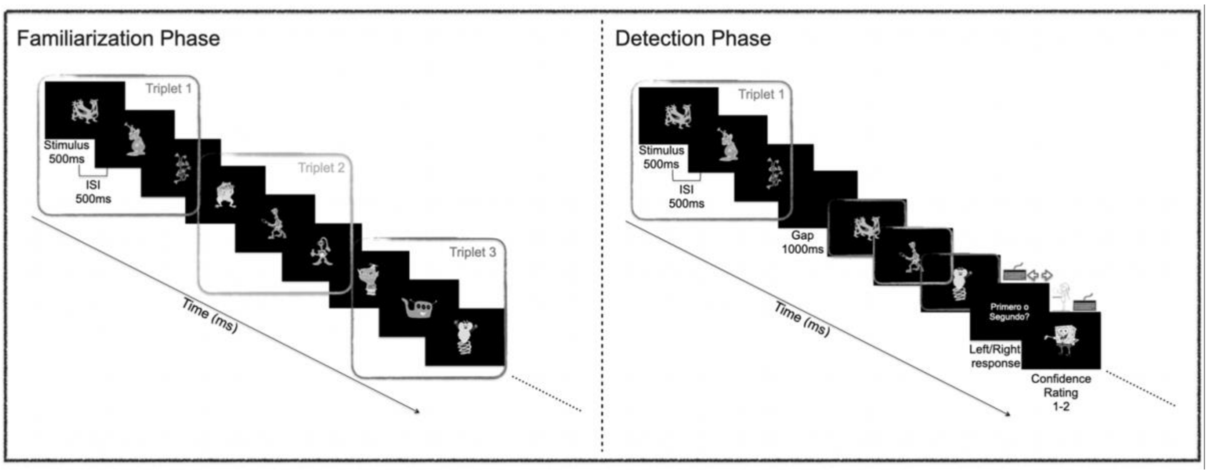
Note. Participants saw a briefly presented sequence composed of 4 consonants, followed by the presentation of a single target consonant in red. Participants had to decide whether the target consonant was part of the sequence (type-1 task) and rate their confidence upon this response (type-2 task).

Visual Statistical Learning Task.

Participants' ability to implicitly detect transitional regularities in a stream of visual stimuli appearing multiple times in specific order on the screen was assessed. The stream was formed of visual sequences composed of 12 alien cartoon figures, that have been previously used in statistical learning studies and that were novel to participants in the sense of not resembling any popular cartoon characters or animals that children were familiar with. The current task was composed of two phases, the familiarisation phase, in which participants were exposed to this stream of novel cartoon figures and the detection phase, in which participants were tested on whether they extracted the triplet regularities occurring in the stream. The exact structure of the type-1 task can be found in (Arciuli & Simpson, 2012). The only modification made was that during the detection phase, in each trial, participants were asked to rate their confidence on having given a correct response or not. Two options (*1: I have doubts on whether my response was correct or not, 2: I'm sure my response was correct*) were given to the participants that were explicitly explained in each trial. This was recorded through an external keyboard by the researcher (see Fig. 2).

Figure 2

Task design of the Visual Statistical Learning task



Note. For details on the structure of the presented task see: Arciuli & Simpson, 2012.

Orthographic Statistical Learning Task.

Participants' ability to implicitly detect ngram orthographic regularities of various grain sizes (unigrams, bigrams, trigrams) was assessed. Pairs of ngrams of the same category were presented on a computer screen, in a white Arial font and black background, isolated from a word/pseudoword context. Pairs were composed of a high-frequency and a low-frequency unit as per the frequency they appear in the Spanish orthography (e.g., C vs Z, ST vs GT, NTK vs SCH). In each trial, participants were asked to point out the ngram appearing on the screen that they thought was more useful to make up new words. Trials were presented in a blocked order (6 unigram trials, 10 bigram trials, 10 tigram trials), but within the blocks they were randomized (for details, see: Mano & Kloos, 2018). In each trial, after giving a response, participants were asked to rate their confidence on having given a correct response or not.

Two options (*1: I have doubts on whether my response was correct or not, 2: I'm sure my response was correct*) were given to the participants that were explicitly explained in each trial. This was recorded through an external keyboard by the researcher (see Supplementary Material: Fig. S1).

Orthographic Lexical Decision Task.

This involved the visual presentation of 40 words and 40 pseudowords. Stimuli were selected from EsPaL, a Spanish lexical database. The selected high-frequency words were 4-letters long. Orthographically legal pseudowords were created by changing one letter of words that were not those presented in the task, but with the same characteristics. The task was divided into two blocks of 40 items (counterbalanced between words and pseudowords) performed in separate sessions in order to ensure participants' attention to the task. Each trial started with the onset of a central fixation cross at the centre of the screen until the participant reported that they were ready to start the trial. Following a blank screen of 500 ms, the target stimulus was presented. Target duration was calibrated for each participant prior to the experimental trials (see below). A backward mask (######) was then presented until participants made their type-1 response, reporting whether the stimulus was a word or a non-word by pressing keys on the keyboard labelled as ✓ or ✗. Participants were then asked to rate their confidence on the response by using two options (i) I *have doubts on whether my response was correct or not*, and (ii) *I'm sure my response was correct*. This was recorded through an external keyboard by the researcher.

In this task, a continuous staircase procedure was used to adjust stimulus presentation duration in order to avoid ceiling effects in type-1 accuracy and to equate participants' type-1 accuracy at around 70%. Such procedure allows for more accurate estimates of type-2 metacognitive efficiency. In order to define the starting duration of each participant, a calibration phase resembling the characteristics of the main task was carried out first, in which the cumulative accuracy on the task was calculated after each trial. The stimulus presentation duration in which participants achieved approximately 70% accuracy on the task was used as a starting duration for the main task. During the experimental trials, the stimulus duration for the next trials was calibrated using a 2 down-1 up staircase, adapting to whether participants responded correctly or not. In the case of two sequential correct responses, the stimulus duration was decreasing by 1 frame, while in the case of one incorrect response, the stimulus duration was increasing by 1 frame (see Supplementary Material: Fig. S2).

The group of participants, which was tested again 10 months later as part of an EEG study, repeated the orthographic lexical decision task using a different set of stimuli (words and pseudowords) with the same characteristics. The experimental procedure described above was followed, except the following modifications, which were essential to maximize the quality of the EEG recordings: (i) the delay following the fixation point was 1000 ms instead of 500 ms; (ii) Participants gave the type-1 response by pointing with their finger on a symbol ✓ on the screen if they thought the string sequence they saw was a real word and ✗ if they thought it was a pseudoword- the symbols (✓ and ✗) were randomly displayed on the left or right side of the screen in each trial; (iii) Participants rated their confidence pointing with their finger on the screen at a cartoon saying “Sure” if they were confident that they responded correctly, and at a cartoon saying “I don’t know” if they were unsure if they responded correctly or not. The cartoons (“Sure” and “I don’t know”) were randomly displayed on the left or right side of the screen in each trial. These modifications aimed to minimise motor activation effects in the EEG recordings that could conflate the confidence signals.

Non-linguistic task unrelated to reading skills

This task was added for comparative reasons, in order to allow us to disentangle whether any observed pattern of results related to the relationship between students’ metacognitive efficiency and their performance in the different tasks is limited to reading skills.

Emotion Recognition Task.

An emotion recognition task was developed including the presentation of human face pictures for which participants were asked to make a decision on whether the face was happy or neutral. A total of 48 stimuli was presented, selected from the "Developmental Emotional Faces Stimulus Set" (DEFSS, Meuwissen et al., 2017), which were balanced for mood, gender, and age (child or adult) of the face. Each trial started with a central fixation cross till the participant was ready to start the trial. The experimenter then initiated the trial. Following a blank screen of 500 ms, a forward mask (composed of different colours and cartoon robot pieces) appeared for 100 ms, and then the target stimulus appeared with a duration that was pre-calibrated prior to the experimental trials using a similar procedure to the orthographic lexical decision task above. After the offset of the stimuli, a backward mask of 100 ms was displayed on the screen and participants were asked to give a HAPPY/NEUTRAL response (by pressing keys on the keyboard labelled as ☺ or 😐) on whether the face they saw on the screen was depicting a happy or a neutral emotion. Subsequently, in each trial, participants were asked to rate their confidence in the same manner as outlined above (see Supplementary Material: Fig. S3).

The group of participants, which was tested again 10 months later for the EEG, repeated the current task as described above, except similar modifications to those noted above regarding the execution of the type-1 and type-2 responses.

Standardized reading tests

In order to obtain a standardised score for tracking participants' reading skill, the single-item reading subtests of the PROLEC-R battery were used, which have been considered to provide accurate estimates of the general reading level of children (see details in Cuetos, Rodrigues, Ruano, 1996). In this test, participants were given lists of words and pseudowords (40 items per list) and were asked to read them out loud. The time taken to read each list as well as the number of errors was recorded. The reading speed was estimated separately for each list, by dividing the number of words participants read correctly by the total time taken to read each list.

Control task: Non-verbal IQ

Participants' non-verbal IQ was assessed using the matrices subtests of WISC-V (Wechsler, 2014). The raw scores were first converted to scaled scores according to the age band each participant belonged, following the tables of normative samples provided in the WISC-V manual. Participants with a score inferior to the 25th percentile on WISC-V matrices' scaled scores were excluded from the analysis. Scaled non-verbal IQ scores were used as a covariate in all analyses in order to rule out the possibility that given associations are driven by general factors of intelligence.

Data analysis

Prior to the main data analysis, participants performing at chance level in a task were excluded from the analysis of the certain task (accuracy ≤ 0.5). Subsequently, the interquartile range (IQR) criterion was used to screen for outliers. Values that are two interquartile ranges larger or smaller from the median, were identified as outliers. Accuracy data were screened for outliers, and outliers were excluded separately in each experiment (total included participants after removing outliers in each task: VA span task: N=55, visual statistical learning: N=30, orthographic statistical learning: N=42, lexical decision task: N=55, emotion recognition Task: N=59). Our analyses focus on the estimation of metacognitive efficiency (meta-d'/d') using an SDT framework.

In SDT models, type-1 performance (d') illustrates the ability of an observer to discriminate between two different states of the world (e.g., signal vs noise, word vs pseudoword, happy vs neutral). It is calculated as d' = z (hits) - z (false alarms), where z(p), p $\in$ [0,1] is the inverse of the cumulative distribution function of the normal Gaussian distribution. "Hits" refer to the proportion of trials in which the subject detected 'signal' when the 'signal' was present, while “false alarms” refer to the proportion of trials in which the subject detected ‘signal’ and the ‘signal’ was absent. Type-2 SDT metacognitive performance (meta-d’) refers to the ability of the subject to discriminate between correct vs incorrect responses by means of the confidence ratings. Here in the type-2 analysis, “hits” correspond to the proportion of trials in which the subject responded with high confidence and the type-1 response was correct, while “false alarms” refer to the proportion of trials in which they responded with high confidence and the type-1 response was incorrect. Type-2 meta-d’ is estimated as the type-1 d’ value that would correspond to the observed confidence distributions in a metacognitive “ideal” subject (Brian Maniscalco & Lau, 2012). Hence, type-1 d’ and type-2 meta-d’ are in the same units, thus they can be comparable. In an ideal metacognitive observer: meta-d’=d’. If meta-d’<d’, we can deduce that the subject is not using all the available stimulus information to inform their metacognitive system. In cases where meta-d’>d’, subjects are supposed to further process the stimulus information fruitfully, after having given the type-1 response and before giving their metacognitive judgments. Using meta-d’ as an estimate of type-2 performance is free of confidence bias but it can be affected by the task difficulty. Calculating the ratio of meta-d’/d’ (M-ratio) allows for an estimate of type-2 performance controlling for the subject's type-1 performance and task difficulty (reported in the results as “type-2 metacognitive efficiency”). This measure permits meaningful comparison across subjects or task domains.

In this study, HMeta-d toolbox (https://github.com/metacoglab/HMeta-d), a recently developed SDT hierarchical Bayesian framework (Fleming, 2017), was used to estimate metacognitive efficiency (meta-d’/d’) in all tasks performed by the children. This framework allows for the estimation of metacognitive efficiency both at the single-subject and at the group level. Single-subject estimates are computed separately for each subject based on the corresponding d’ scores (i.e., the ratio between hits and false alarms in the type-1 task) and the meta-d’ scores that reflect how well one’s confidence ratings align with response accuracy in the type-1 task. However, single-subject estimates can be noisy and uncertain (i.e., with reduced trial numbers per participant). Group-level estimates of metacognitive efficiency are useful here because they can incorporate subject-level uncertainty in the estimation, meaning that a participant with high level of uncertainty doesn't contribute equally to the estimation of group-level parameters. These estimates are particularly useful for a direct estimation of covariance in metacognitive efficiency across tasks or groups of participants. Also, HMeta-d Bayesian framework avoids edge correction, handling naturally possible zero cell counts in a certain confidence level. This was crucial in the current study, as in early childhood, several participants have a tendency to respond with a high confidence rating in a high proportion of trials, despite being instructed to use all the confidence ratings accordingly.

Normality tests were conducted in each variable of interest of the experimental tasks. The measure of Skewness was used to evaluate normality. Considering that many values were moderately to highly skewed (Skewness>0.5 or <-0.5), we elected to use non-parametric tests for the frequentist testing.

In order to investigate how single-subject estimates of type-2 metacognitive efficiency relate to participants' performance across the standardized reading tasks and type-1 performance in the experimental tasks administered, Spearman's r correlations were used. For each correlation analysis, False Discovery Rate (FDR) was used for multiple comparison correction and participants' chronological age and intellectual ability were used as covariates.

To identify possible differences in the single-subject estimates of metacognitive efficiency across tasks, the non-parametric Friedman Test was applied with task as a within subject factor. Moreover, under the Bayesian hierarchical framework, we calculated the difference between the posterior distributions of the group-level estimates of metacognitive efficiency, estimated using the Hmeta-d toolbox function: *fit_meta_d_mcmc_group.m*, for each pair of tasks. Significant differences across tasks are indicated when the 95% highest density intervals (HDI) of the difference of the posterior distributions do not overlap with zero.

Next, to examine whether type-2 metacognitive efficiency of the participants correlated across tasks at the single-subject level, we applied Spearman's r correlations. In order to verify these correlations using group-level estimates of type-2 metacognitive efficiency, we applied the HMeta-d toolbox function: *fit_meta_d_mcmc_groupCorr.m* to calculate the 95% highest density intervals (HDIs) on the posterior distributions of the correlation's coefficients (for details see Fleming, 2017). Significance is indicated when the posterior distributions do not overlap with zero.

Finally, in order to estimate participants' longitudinal improvement in performance in the orthographic lexical decision task (linguistic task) and the emotion recognition task (non-linguistic task) between the two time points of the study, we calculated the difference in the stimulus duration need to perform the task between the two timepoints, at the level set by our adaptive staircase. We computed the mean presentation time of the words or pseudowords (lexical decision task) or the face (emotion recognition task) across all trials of each task between the two timepoints. As noted above, the use of the online adaptive staircase at each timepoint adjusted the presentation time of the stimulus based on participants' discrimination accuracy on each trial, converging to a similar level of performance (i.e., around 70%). Hence, we used the difference in stimulus presentation times across the two time points as a measure of longitudinal learning effects.

We then investigated the longitudinal links between participants' metacognitive skills and long-term performance improvement, using linear and Bayesian regression analyses. For the linear regressions, Ordinary Least Squares regression was used, complemented by Huber robust regression to account for outliers. Both results are mentioned in the results' section. For the Bayesian regressions, a default prior of 0.354, as implemented in JASP software (van Doorn et al., 2020) was used and the Bayesian inclusion factor ($BF_{inclusion}$) was estimated for every predictor in the model. $BF_{inclusion}$ is calculated by dividing the prior odds of a model including a predictor of interest by the posterior odds (i.e., BF_{10}) excluding this predictor. When $BF_{inclusion}>1$, it indicates that the model was improved by the addition of this specific predictor.

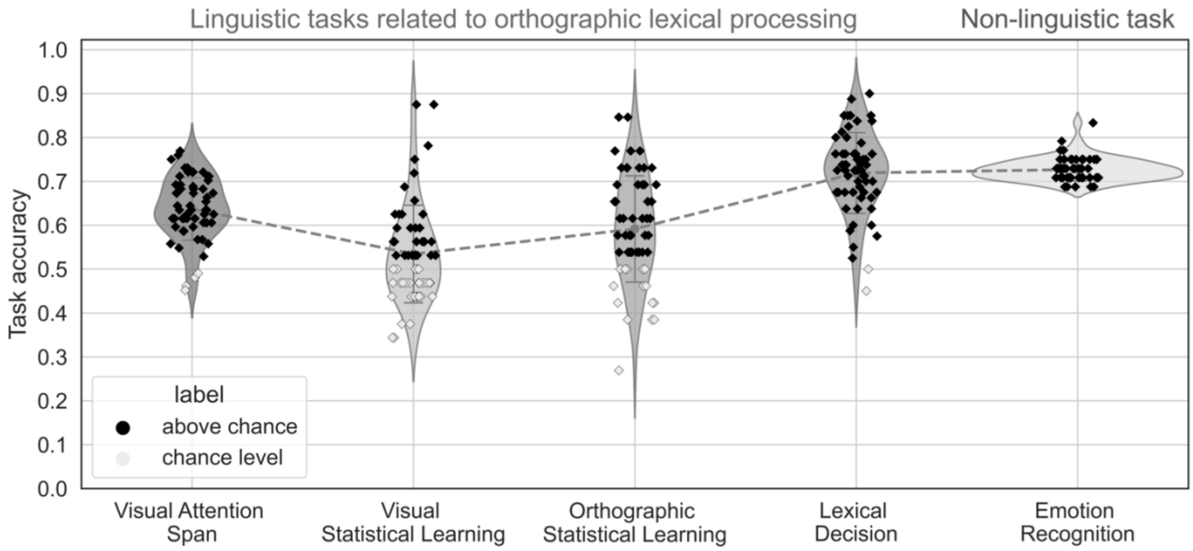
Results

Descriptives

Fig. 3 depicts the distribution of participants' accuracy in each of the experimental tasks. During this pre-processing step of analysis, we found out that almost half of the participants in both statistical learning tasks were performing below chance (see Fig. 3). This suggests that these tasks were quite demanding for 6-7 years old children. Therefore, any results regarding the two statistical learning tasks will be reported separately from the rest of the tasks, and should be viewed cautiously due to the small sample size of children with above chance performance, which would preclude a meaningful assessment of inter-individual differences in metacognitive ability across the different tasks. For the rest of the experimental tasks in which our analyses are going to be focused (VA span, lexical decision, and emotion recognition), the IQR criterion was applied, and outliers were identified for each task separately due to very high or low accuracy on the task (VA span task: N=55, lexical decision task: N=55, emotion recognition Task: N=59).

Figure 3

Distribution of participants' performance accuracy



Note. Jittered scatter plots illustrate the distribution of participants' performance accuracy (0: no correct response - 1: all responses correct) in the 5 experimental tasks (N=60). The dark gray lines connect the means of the distributions on each task. Light gray dots indicate participants whose accuracy was equal or lower than chance level (task accuracy ≤ 0.5) and were excluded from the analysis. Due to the excessive number of participants in the statistical learning tasks performing below chance, these experiments were not analysed further.

Table 1 summarises the descriptive analysis for the measures used to assess type-1 performance (Accuracy, Stimulus Presentation Time, Task sensitivity (d' prime)) and also type-2 performance (mean confidence, single-subject metacognitive efficiency (Mratio)) performance in each task.

Table 1

Descriptive statistics (Mean Score (SD), Range, Skewness) for the different measures of participants' type-1 and type-2 performance in the experimental tasks.

	Linguistic tasks related to orthographic lexical processing		*Non-linguistic task*
	VA Span	Lexical Decision	Emotion Recognition
Task accuracy (% of correct responses)			
Mean (SD)	0.65 (0.06)	0.72 (0.07)	0.73 (0.02)
Range	0.55-0.77	0.55-0.85	0.69-0.79
Skewness	0.207	-0.2354	0.4724
Type-1 task sensitivity (d'prime)			
Mean (SD)	0.84 (0.29)	1.26 (0.47)	1.21 (0.18)
Range	0.29-1.45	0.30-2.37	0.94-2.05
Skewness	0.133	0.282	1.991
Type-2 Metacognitive Efficiency (Mratio)			
Mean (SD)	1.31 (0.87)	1.65 (1.48)	1.44 (0.70)
Range	-0.73-4.10	0.12-9.32	0.09-3.31
Skewness	0.850	3.281	0.433

Next, we present the results of the different analysis dealing with each of the questions addressed in the study.

No evidence for an association between metacognitive efficiency in tasks assessing orthographic processing and performance in standardised reading tests

First, we performed a verification check analysis regarding the relationship between inter-individual differences in type-1 performance on the experimental tasks and participants' general reading level assessed by the standardised reading tasks. Based on previous literature, it was a priori expected that performance on the experimental tasks assessing skills related to orthographic processing would correlate between them and with students' general reading level (e.g.,Ginestet et al., 2021; Valdois et al., 2019). No association was expected between those and performance in the non-linguistic emotion recognition task. In line with this, a strong association was shown between participants' type-1 task sensitivity in the reading-related experimental tasks and their performance in the standardised reading tests (all ps < 0.05). Moreover, a positive correlation was revealed between type-1 task sensitivity in the VA span task and the orthographic lexical decision task. On the contrary, type-1 performance on the non-linguistic task did not correlate with any of our reading-related tasks (all ps > 0.05, see Table 2 for details). These results show a significant variance in the type-1 performance data across participants, hence providing a solid foundation for assessing similar associations in the metacognitive domain.

Next, we addressed whether type-2 metacognitive efficiency across the experimental tasks is associated with variations in participants' performance on the standardised reading tests. We expected participants with higher type-2 metacognitive efficiency to exhibit better performance across all our reading tasks.

No significant correlations were found between type-2 metacognitive efficiency in the experimental tasks and the participants' performance in standardized reading tests (all ps>0.05). Bayes factor provides moderate evidence towards the null hypothesis in most of the cases ($BF_{10}<0.33$, see Table 3).

Table 2

Correlations between type-1 task sensitivity in the experimental tasks and students' reading performance

	Standardised reading tasks		***Linguistic tasks related to orthographic lexical processing***		***Non-linguistic task***
	Words Speed	Pseudowords Speed	VA Span Type-1 Task Sensitivity	Lexical Decision Type-1 Task Sensitivity	Emotion Recognition Type-1 Task Sensitivity
Words Speed	-	r = 0.947*** p < 0.001 BF_{10} >100	r = 0.416** p = 0.005 BF_{10} = 25.630	r = 0.503*** p < 0.001 BF_{10} = 19.030	r = -0.018 p = 0.893 BF_{10} = 0.177
Pseudowords Speed	-	-	r = 0.392** p = 0.007 BF_{10} = 12.788	r= 0.457** p = 0.002 BF_{10} = 11.117	r = -0.035 p = 0.882 BF_{10} = 0.167
VA Span Type-1 Task Sensitivity	-	-	-	r = 0.314* p = 0.049 BF_{10} = 1.061	r = 0.040 p = 0.882 BF_{10} = 0.189
Lexical Decision Type-1 Task Sensitivity	-	-	-	-	r = 0.196 p = 0.235 BF_{10} = 0.548

Note. Spearman's correlations were performed between type-1 task sensitivity (d' prime) in the experimental tasks and: a) students' performance on standardized tasks measuring reading ability, b) type-1 task sensitivity in the rest of the experimental tasks (VA Span Task: N=55, lexical decision task: N=55, emotion recognition task: N=59, *p<0.05, **p<0.01, ***p<0.001, FDR-corrected). Correlations were controlled for participant's age and intellectual ability (non-verbal IQ, Matrices-WISC).

As an exploratory follow-up analysis, we assessed the relationship between type-2 metacognitive efficiency and type-1 performance across the different experimental tasks (this was not done within each task given that type-1 and type-2 performance are likely to be correlated within the same task, see: Fleming & Lau, 2014). In principle, it might be expected that better type-1 performance correlates with better metacognitive efficiency. However, type-2 metacognitive efficiency in the lexical decision task negatively correlated with type-1 performance in both the VA span task (r=-0.355, p=0.027) and the emotion recognition task (r=-0.371, p=0.018; see Table 3 for details and Supplemental Material: Fig. S4 for the visualisation of significant correlations in scatterplots). Moreover, type-2 metacognitive efficiency in the emotion recognition task negatively correlated with performance in the lexical decision task (r=-0.342, p=0.027). On the contrary, type-2 performance in the VA span did not correlate with type-1 performance in any of the tasks (all ps>0.05), with Bayes factor providing moderate evidence towards the null hypothesis (BF_{10}<0.33, see Table 3 and Fig. S4).

Table 3

Correlations between metacognitive efficiency in the experimental tasks and students' reading performance

	Standardised reading tasks		*Linguistic tasks related to orthographic lexical processing*		*Non-linguistic task*
	Words Speed	Pseudowords Speed	VA Span Type-1 Task Sensitivity	Lexical Decision Type-1 Task Sensitivity	Emotion Recognition Type-1 Task Sensitivity
VA Span Type-2 Metacognitive Efficiency	r = -0.013 p = 0.926 BF_{10} = 0.170	r = 0.014 p = 0.926 BF_{10} = 0.171	-	r = 0.029 p = 0.926 BF_{10} = 0.177	r = -0.094 p = 0.645 BF_{10} = 0.215
Lexical Decision Type-2 Metacognitive Efficiency	r = -0.185 p = 0.303 BF_{10} = 0.903	r = -0.195 p = 0.288 BF_{10} = 1.045	r = -0.355* p = 0.031 BF_{10} = 10.721	-	r = -0.371* p = 0.021 BF_{10} = 0.737
Emotion Recognition Type-2 Metacognitive Efficiency	r = -0.097 p = 0.633 BF_{10} = 0.212	r = -0.107 p = 0.602 BF_{10} = 0.226	r = -0.142 p = 0.465 BF_{10} = 0.274	r = -0.342* p = 0.031 BF_{10} = 2.556	-

Note. Spearman's correlations were performed between metacognitive efficiency (Mratio) in the experimental tasks and a) students' performance on standardized tasks measuring reading ability, b) type-1 task sensitivity in the rest of the experimental tasks (VA Span Task: N=55, lexical decision task: N=55, Emotion Recognition Task: N=59, *p<0.05, **p<0.01, ***p<0.001, FDR corrected). Correlations were controlled for participant's age and intellectual ability (non-verbal IQ, Matrices-WISC). Note that in these correlation analysis correlations between type-2 metacognitive efficiency and type-1 performance within each task are not reported, due to the fact that statistical estimation of type-2 metacognitive efficiency is dependent on type-1 performance (Maniscalco and Lau, 2012).

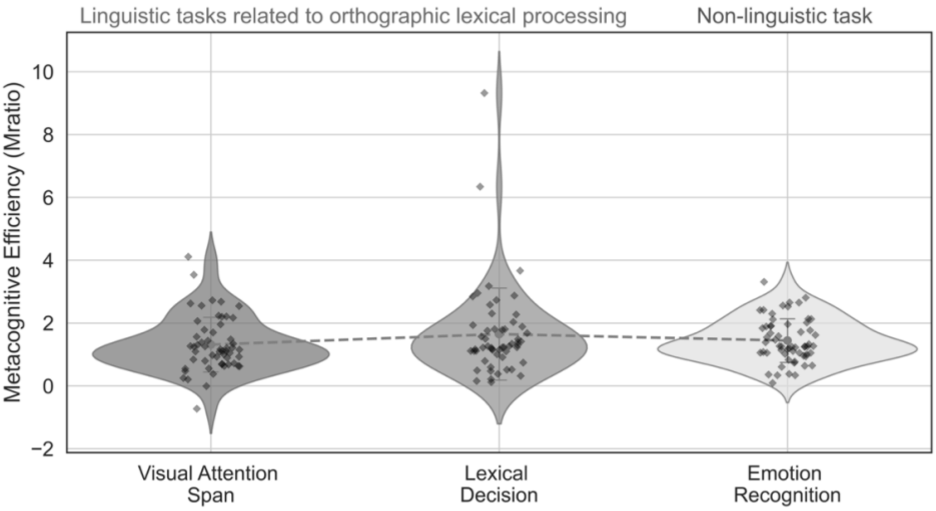
The aforementioned analysis was also performed on the Visual and Orthographic Statistical learning tasks, both including and excluding participants with accuracy lower than 55%. No significant correlation was found between participants' type-1 task sensitivity or type-2 performance and their performance in standardized reading tests (all ps>0.05, see Supplemental Material: Table S1-S3).

Partial support for domain-general mechanisms of metacognition.

Single subject estimates of type-2 metacognitive efficiency (Mratio) were compared across tasks, in order to examine whether metacognition is supported by domain-general or specific mechanisms. A non-parametric Friedman test with task as a within-subjects factor was conducted and rendered a Chi-square value of 2.00, which showed no main effect of task on the metacognitive efficiency of the participants (p=0.368). Fig. 4 shows the distributions of the single subject fits of metacognitive efficiency across tasks. Next, the difference between the posterior distributions of the group-level estimates of metacognition was calculated between each pair of tasks. The same pattern of results was revealed, with none of the HDIs of the difference overlapping 0, indicating no significant difference between the tasks (VA Span-Lexical Decision: 95% HDI = [-0.26, 0.47]; VA Span-Emotion Recognition: 95% HDI = [-0.15, 0.54], Lexical Decision-Emotion Recognition: 95% HDI = [-0.42, 0.17]).

Figure 4

Distribution of students' metacognitive efficiency in the experimental tasks



Note. Jittered scatter plots illustrate the distribution of the single-subject parameter estimates of type-2 metacognitive efficiency of participants in Experiments 1-3 (N=50). The black lines connect the means of the distributions on each task.

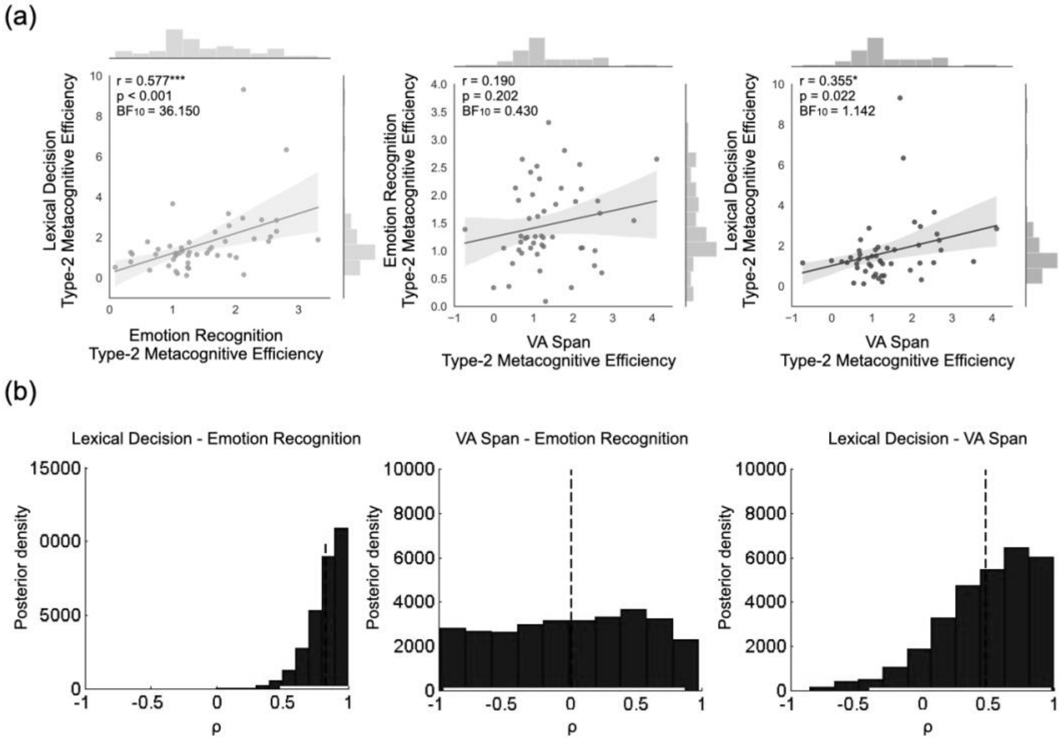
Spearman's correlations coefficients between single-subject estimates of metacognitive efficiency (Mratio) across the different tasks are presented in Fig. 5a. The goal was to examine whether metacognitive efficiency is domain-general. A strong positive correlation was observed between metacognitive efficiency on the orthographic lexical decision task and the emotion recognition task (r=0.577, p<0.001). Bayes factor provides very strong evidence in favour of this hypothesis (BF_{10}=36.15), suggesting the use of a common metacognitive mechanism in these tasks. A significant correlation was also noted between metacognitive efficiency on the VA span task and the orthographic lexical decision task (r=0.355, p=0.022), with the Bayes factor providing only anecdotal evidence towards the alternative hypothesis (BF_{10}< 3).

Next, the covariance of participants' group metacognitive efficiency across tasks was evaluated within the hierarchical model of the Bayesian framework. Substantial covariance is suggested only between the lexical decision task and the emotion recognition task, shown by

95% HDIs on the posterior distributions of the correlation coefficients which do not overlap zero, hence indicating a significant correlation (VA Span-Lexical Decision: ρ=0.491, 95% HDI=[-0.38, 0,99]; VA Span-Emotion Recognition: ρ=-0.063, 95% HDI=[-0.99, 0.74], Lexical Decision-Emotion Recognition: ρ=0.842, 95% HDI=[0.57, 0.99], see Fig. 5b).

Figure 5

Correlations among students' metacognitive efficiency across the experimental tasks



Note. Figure description: (a) Spearman correlations among type-2 metacognitive efficiency in experimental tasks (N=49, *p<0.05, **p<0.01, ***p<0.001, FDR corrected). Correlations were controlled for participant's age and intellectual ability (non-verbal IQ). (b) Posterior distributions over ρ for each correlation pair across Experiments 1-3 determining covariance between metacognitive efficiency across tasks. The white horizontal bar indicates the 95% of high-density intervals (HDIs). The black dotted line indicates the ground-truth correlation between type-2 metacognitive efficiencies. In cases where 95% HDIs on posterior correlation coefficients do not overlap zero, a substantial covariance in metacognitive efficiency across domains is suggested.

Metacognitive efficiency is a predictor of longitudinal learning across different domains

We investigated whether metacognitive efficiency in a sub-group of children (N=40) retested in the context of an EEG study, uniquely and similarly predicts their longitudinal performance improvement on the orthographic lexical decision task and the emotion recognition task during a period of 10 months. In order to ensure that the sub-group tested did not differ from the group of children left out of the EEG study (N=20), the variability in type-1 (d-prime) and type-2 (Mratio) performance variables was compared between these groups using the non-parametric independent samples Mann-Whitney U test, followed by Bayesian independent sample t-test. No significant differences were apparent between groups in these variables in any of the tasks (lexical decision: d-prime (W=354.000, p=0.839, BF_{10}=0.288); Mratio (W=245.000, p=0.088, BF_{10} =0.525) / emotion recognition: d-prime (W=338.000, p=0.502, BF_{10}=0.429); Mratio (W=316.000, p=0.306, BF_{10} =0.526)).

Note that the use of the online adaptive staircase to adjust stimulus duration based on discrimination accuracy, meant that the level of discrimination performance in the lexical decision and emotion recognition tasks was similar (~70% correct) across the two time points. Hence, we assessed the longitudinal improvement in performance (i.e., learning effects) as the difference in the mean presentation duration of words/pseudowords (lexical decision task) or faces (emotion recognition task) between time point 1 (Grade 1) and time point 2 (10 months later, Grade 2) for each of the two tasks. Mean presentation duration was computed as the average duration of the stimuli across all trials within each time point and experimental task. Here, it is important to mention that larger learning effects are indicated by a negative change in stimulus presentation duration, as lower values in duration (msec) indicate better performance (faster in time).

We conducted regression analyses to predict performance improvement on type-1 tasks using metacognitive efficiency, age and IQ at time point 1 (t1) as predictors. Analyses were performed using three regression models that all indicated that for both tasks, metacognitive efficiency in t1 was a significant predictor of performance improvement (orthographic lexical decision: Ordinary least squares: p=0.030, Hubert robust regression: p=0.033, Bayesian regression: $BF_{inclusion}$=2.227 (see Table 4, Supplemental Material: Fig. S5a); emotion recognition: Ordinary least squares: p=0.035, Hubert robust regression: p=0.018, Bayesian regression: $BF_{inclusion}$=1.386). In addition, when age and IQ were included in the null model of Bayesian regression as nuisance predictors ($BF_{inclusion}$=1, see details in: van den Bergh et al., 2021), metacognitive efficiency still strongly predicted performance improvement (lexical decision: $BF_{inclusion}$=3.003; emotion recognition: $BF_{inclusion}$=2.639 (see Table 5, Supplemental Material: Fig. S5b). Note that the coefficient associated with metacognitive efficiency as a predictor of learning in both tasks is negative because learning effects are calculated as a difference in stimulus presentation duration (reduced time, better performance). Hence, the more negative the difference is between the two time points of the study, the greater the amount of learning.

Table 4

Regression analysis of longitudinal improvements in the lexical decision task predicted by students' early metacognitive efficiency

	Learning (t2-t1 Mean Stimulus Presentation duration in lexical decision task (ms))				
	β	t	p	BFinclusion	BFinclusion (Age, IQ in null model)
Type-2 Metacognitive Efficiency in t1	-39.136 (-38.965)	-2.276 (-2.212)	0.030 (0.033)	2.227	3.003
Age	-5.255 (-10.492)	-0.085 (-0.166)	0.933 (0.868)	0.446	1.000
Non-verbal IQ	0.443 (0.335)	-0.057 (0.004)	0.955 (0.966)	0.468	1.000

Note. Early metacognitive efficiency in the lexical decision task in t1, age, and non-verbal IQ, were introduced as predictors in the different regression models applied (N=36, 4 outliers for accuracy excluded in t1). Values after applying ordinary least squares regression are reported outside the parenthesis, while values after applying Hubert robust regression to account for outliers are reported in parenthesis. BF$_{inclusion}$ factor represents the change from prior to posterior probabilities of a model when a predictor is added in the equation (BF$_{inclusion}$>1 indicates that the predictor improves the model).

Table 5

Regression analysis of longitudinal improvements in the emotion task predicted by students' early metacognitive efficiency

	Learning (t2-t1 Mean Stimulus Presentation duration in emotion recognition task (ms))				
	β	t	p	BFinclusion	BFinclusion (Age, IQ in null model)
Type-2 Metacognitive Efficiency in t1	-12.720 (-12.986)	-2.187 (-2.533)	0.035 (0.018)	1.386	2.639
Age	-18.265 (-13.206)	1.692 (-1.388)	0.099 (0.163)	0.934	1.000
Non-verbal IQ	0.134 (-0.326)	0.104 (-0.287)	0.918 (0.776)	0.501	1.000

Note. Early metacognitive efficiency in the emotion recognition task in t1, age, and non-verbal IQ, were introduced as predictors in the different regression models applied (N=40). Values after applying ordinary least squares regression are reported outside the parenthesis, while values after applying Hubert robust regression to account for outliers are reported in parenthesis. BF$_{inclusion}$ factor represents the change from prior to posterior probabilities of a model when a predictor is added in the equation (BF$_{inclusion}$>1 indicates that the predictor improves the model).

Discussion

The main goals of this study were threefold. First, we wanted to explore whether students' metacognitive ability whilst performing tasks tapping into orthographic lexical knowledge is related to reading. Second, we sought to determine whether metacognitive ability at this early stage of neurocognitive development is mediated by a domain-general or a domain-specific system. Third, we examined whether metacognitive efficiency in a linguistic task assessing orthographic knowledge and a non-linguistic emotion recognition task can predict long-term learning during the first year of formal reading instruction. Below, we briefly report the main results that will be discussed.

First, type-1 performance in the reading-related tasks (VA span and orthographic lexical decision), but not in the emotion recognition task, was associated with students' performance in the standardised reading tests, assessing the general reading level of students (word and pseudoword reading). There was no evidence for any association between metacognitive efficiency on the three experimental tasks and reading performance in the standardised tests. Interestingly, we found a significant negative correlation between participants' metacognitive efficiency in the lexical decision task and type-1 performance in both the VA span task and the emotion recognition task. Second, the levels of metacognitive efficiency on the lexical decision and the emotion recognition tasks (but not the VA span task) were associated both in an individual and group level analyses, providing some evidence for the existence of domain general metacognitive mechanisms. Last, we showed that metacognitive efficiency was a significant predictor of longitudinal learning in both a linguistic (orthographic lexical decision) and a non-linguistic (emotion recognition) task. Below, we discuss the different results under the scope of the research questions we set in the Introduction:

Metacognitive efficiency for orthographic processing and general reading level

Previous research investigating the relationship between metacognitive ability and performance in standardised tests across different domains reported mixed findings. Positive associations have been reported in several domains, like mathematics (Bellon et al., 2019), emotion recognition (Kelly & Metcalfe 2011), spelling, and text comprehension (Griffin et al. 2008). However, other studies reported no association between metacognitive monitoring and cognitive ability such as memory strategies (Kelly et al., 1976) or text comprehension skills (Griffin et al., 2009). It has been suggested that these results may be attributed to the use of metacognitive indexes which are susceptible to the confounding effects of confidence bias, participants' level of type-1 performance and methods that permit guessing, which can differentially affect the estimation of metacognitive accuracy in high vs poor performers (Vuorre & Metcalfe, 2021). In the present study, we used a bias free signal detection theoretic framework, including Bayesian estimation, to assess participants' metacognitive efficiency, overcoming these issues.

No association was found between participants' metacognitive efficiency in any of the experimental tasks on orthographic processing (VA span and Lexical Decision) and their performance in the standardised reading tests, assessing their general reading level (word and pseudoword reading). Recently, Filevich et al. (2020) have suggested that the age of 6 (like our children participants) is a critical age in the development of metacognitive monitoring. Using tasks in which children had to recognize and report their knowledge certainty in the task, they showed that children's ability to correctly identify and explicitly report that they did not know is associated with key changes in cortical thickness in the medial orbitofrontal cortex (Filevich et al., 2020). Additionally, Brod et al. (2017) suggested that the first year of schooling brings a shift in children's cognitive abilities that may be critical for metacognition. Specifically, during this first year of schooling, students between 5 and 6 years of age showed great improvements in tasks requiring executive control functions, which are also linked to activity changes in parietal cortex regions associated with attention control (Brod et al., 2017).

In t1 of the present study, children belonged to this critical age window, both for the development of cognitive and metacognitive processes, and for the development of reading ability. The first stages of reading development have been previously characterized by an enhanced variability in students' performance (Parrila et al., 2005). Heterogeneous previous reading experience ranges from children who enter primary school having received extended reading instruction at home/kindergarten and who, already from this early stage, rely more on sight-word reading, to others who have received no previous reading instruction and read by decoding. Taking into account this variability in reading ability together with the neurocognitive changes happening in this critical age window, one possibility is that the neurodevelopmental trajectories of the systems that are relevant for acquiring reading are somehow segregated from the systems that support attention and cognitive control, and hence, metacognition. Further investigation in later stages of primary school is needed to determine the factors that mediate the interplay between metacognition and general reading ability.

Third, intriguingly, we observed that metacognitive efficiency in the lexical decision task negatively correlated with type-1 performance in both the VA span and the emotion recognition task. A similar pattern of results was observed when correlating metacognitive efficiency in the emotion recognition task and type-1 performance in the lexical decision task and vice versa. We suggest that at this early age of development, type-2 metacognitive processes may not develop linearly with the type-1 processes, but rather work adaptively to regulate type-1 task performance according to students' abilities. Hence, students with lower type-1 performance may compensate for their difficulties by means of an increase in metacognitive monitoring ability, possibly driven by an increased signalling from error monitoring systems, which would lead to them knowing better when they are incorrect or feel uncertain about their decisions. Accordingly, it has been suggested that the development of metacognitive monitoring in the early ages of primary school is particularly related to the efficient monitoring of incorrect responses (Destan et al., 2014). To support this hypothesis, we ran some additional post hoc correlation analyses between participants' type-1 task sensitivity and the percentage of incorrect responses rated with low confidence on the task. We found that type-1 sensitivity both in the lexical decision task and the VA span task negatively correlated to the proportion of incorrect responses rated with low confidence within each task, so that children with these higher rates had lower type-1 performance (r=-0.463, p=0.004 and r=-0.329, p=0.033 respectively); likewise, type-1 sensitivity in the VA span task negatively correlated to the percentage of low confidence incorrect responses in the lexical decision task (r=-0.431, p=0.005). Moreover, we found that the percentage of incorrect responses rated with low confidence in the lexical decision task was negatively correlated with participants' performance in both of the standardised tests measuring reading ability (word reading: r=-0.333, p=0.033/pseudoword reading: r=-0.283, p=0.073, see details in Supplemental Material: Table S4).

One possible explanation regarding the negative correlation between participants' type-1 performance in the VA span task and error monitoring indexes in the orthographic lexical decision task is that early readers that already have in place the necessary tools for fluent reading such as VA span skills -previously proven to be a cognitive precursor of reading (Valdois et al., 2019) may use more implicit or automatic ways of reading strategies, having less need of monitoring their performance at this stage of reading acquisition. Conversely, students who exhibit lower reading performance and orthographic knowledge do this in a more controlled fashion, and become more able to detect their errors efficiently in the reading related tasks.

Domain-general/specific mechanisms supporting metacognitive efficiency

A further goal of the present study was to investigate whether metacognitive ability in early childhood is supported by domain-general or domain-specific mechanisms. This issue has been scarcely studied in the field of cognitive development. Our data only revealed partial evidence pointing to common underlying mechanisms supporting metacognition. Only in the lexical and emotion recognition tasks, participants' metacognitive efficiency, both in a single-subject and a group level, was highly correlated. We observed a weak correlation between single-subject estimates of metacognitive efficiency on the VA span task and the orthographic lexical decision task, which was not borne out by the analysis of group-level estimates under the hierarchical Bayesian framework. No correlation was found between the VA span and the emotion recognition task. This is in keeping with previous studies pointing to a gradual shift towards a domain general metacognitive system during childhood (Bellon et al., 2019; Geurten et al., 2018; Vo et al., 2014).

First, Vo et al. (2014) suggested the existence of domain specific metacognitive mechanisms supporting numeric and emotional domains in the age of 5-8 (Vo et al., 2014). Geurten and colleagues later evaluated metacognition in different age groups in arithmetic and memory domains and suggested that the shift towards domain general mechanisms underlying metacognition is happening at the age of 10-11 (Geurten et al., 2018). A following study of Bellon et al (2019) found that correlations of metacognitive ability across arithmetic and spelling domains can already be detected from the age of 8-9 (Bellon et al., 2019). These studies are to our knowledge the only developmental studies studying cross-domain metacognition in different tasks using confidence judgments, but are limited by the use of metacognitive indexes which do not control for the effect of metacognitive bias or the level of type-1 performance.

Here, this issue was addressed by using group-level estimates of type-2 metacognitive efficiency (Mratio) under the Bayesian H-metad framework, which revealed the existence of common underlying mechanisms of metacognition even from the age of 6-7. Nevertheless, another explanation for the significant association between metacognitive performance in the lexical decision and the emotion recognition task may well be related to the similar structure of these tasks, both using a 2-alternative discrimination task design and a staircase to adjust type-1 performance. Specifically, the existence of a staircase procedure provided a variety of presentation duration timings of the stimuli during a task, which may have been used as cues to inform childrens’ confidence judgments. Strategic cue utilisation has been suggested to grow especially in younger children (before the age of 11) and to improve monitoring accuracy (Ackerman & Koriat, 2011; Roebers et al., 2019). Hence, the existence of this strong correlation may be related to the fact that participants apply common heuristics in their metacognitive monitoring in these two tasks, rather than indicating the existence of a domain general mechanism supporting metacognition.

Another aspect of the study to be considered is that participants' metacognition in the VA span task was assessed through a target detection ('Yes/No') task. Maniscalco and Lau (2011) have reported that in this type of tasks, metacognitive sensitivity (meta d') for "no" responses is lower than for "yes" responses, as if the presence of the key target feature weights more the sensory representation than its absence (Maniscalco & Lau, 2011). Recently, it has been suggested that differences in task structure might hinder the detection of cross domain correlations (Ruby et al., 2017; Samaha & Postle, 2017). Mazancieux and colleagues (2020) showed cross-task correlations in metacognitive efficiency across four different tasks (i.e., semantic memory, episodic memory, executive function, visual perception), and all of the tasks with the same task structure (Mazancieux et al., 2020, preprint). It would be relevant for future studies that re-examine the domain-generality issue during early neurocognitive development by using similar task structures across all cognitive domains.

Metacognitive efficiency as a predictor of learning

Interestingly, despite the observed lack of associations between metacognitive ability and students' general reading level, we found that students' metacognitive efficiency in Grade 1 was a significant predictor of participants' performance improvement in Grade 2, both in a linguistic (orthographic lexical decision) and a non-linguistic context (emotion recognition).

Metacognition has been long considered as a driving force at regulating individuals' learning, by monitoring uncertainty, guiding exploration and controlling performance (Flavell, 1979; Metcalfe, 2009; Narens, 1990). In educational practice, it has been suggested that metacognition can regulate study time allocation for easy vs hard tasks, or direct students' need for information seeking or assistance (Desender et al., 2018; Dunlosky et al., 2011; Son & Metcalfe, 2000). Of note, most of these studies investigating the link between metacognition in learning have focussed on associations within a certain time point. Few studies have investigated the relationship of metacognitive monitoring and type-1 performance longitudinally like the present study. Roebers and Spiess (2017) longitudinally tracked the development of online metacognitive monitoring in early primary school in the spelling domain but did not observe that metacognitive monitoring at the beginning of the study (children's age: 7 y.o.) predicted children's performance in a spelling task 8 months later in Grade 2. These results are in contrast to Rinne and Mazzocco's (2014) study showing that early metacognitive skills can predict later performance in an arithmetic task three years later in primary school (Grade 5 to Grade 8). Differences among studies, including ours, may be attributed to the age difference of the children or to the fact that here we assessed the link between metacognition to the learning effect based on the change in performance across two time points. For example, Roebers and Spiess (2017) merely correlated type-2 metacognitive sensitivity at time point 1 with type-1 performance at time point 2 (Roebers & Spiess, 2017) without quantifying any change in performance across time points as we have done here. Our observation that metacognitive skill is predictive of subsequent learning effects across time is in keeping with prior educational studies suggesting that individuals' monitoring ability of their own performance is fundamental for learning (Metcalfe & Kornell, 2007; Rawson et al., 2011).

Of note, when tackling our first research question regarding how metacognitive monitoring in one task relates to task sensitivity in another task, we observed significant negative correlations between participants' metacognitive efficiency in both the lexical decision task and the emotion recognition task and type-1 sensitivity in the rest of the tasks. Moreover, the trend, although non-significant, indicated a negative relationship also between students' metacognitive efficiency in the lexical decision task and students' general reading level (e.g., lexical decision Mratio-words speed: r = -0.185, lexical decision Mratio-pseudowords speed: r = -0.195). Overall, these results illustrate that children who performed poorly on the reading and reading-related tasks are the ones with higher metacognitive efficiency indexes.

Taken together these results with the finding reported here, that students with higher metacognitive efficiency in t1 were the ones exhibiting larger longitudinal learning effects in the same tasks, we suggest that students who are performing lower during the first months of attending primary school, may rely more on metacognitive strategies and error monitoring processes in order to catch up with their initially more advanced peers. Indeed, regarding the linguistic domain, previous literature has indicated that children who learn how to read later on than their peers (e.g., children entering primary school having received extended reading instruction at home or at the kindergarten), catch up in the first years of primary school by the age of 11 (Suggate et al., 2013). Hence, we propose that explicit metacognitive monitoring may be most beneficial when a skill is developing and the student can use monitoring as a tool to inform and optimise/increase automaticity in their type-1 performance.

Limitations and future directions

We believe that a major strength of the present work is that it examined the relationship between metacognitive ability and crucial skills in early reading development (i.e., visual word recognition, orthographic lexical processing), using bias free measures of metacognitive efficiency. Moreover, the use of a longitudinal within-subject approach to assess metacognition in connection with long term learning demonstrates the role of metacognition in the long run, even when this ability did not correlate with students' performance within a certain time point. Notwithstanding, reading acquisition requires the employment of a number of other complex skills that were not studied in the present study. Therefore, additional work is needed to address whether the findings reported here generalise, for instance, to the development of students' phonological and semantic lexicon, which are equally crucial for an individual's reading development. In assessing the domain-generality of metacognition across the different tasks, it will be important to preclude the existence of additional cues that may influence metacognitive monitoring regardless of self-evaluation. Here, the use of a staircase procedure with varying stimulus durations to match performance (Fleming et al., 2010), meant however that the correlations in metacognitive efficiency in the lexical and the emotion recognition task could be overestimated, given that children may have used the varying stimulus duration as a cue to rate their confidence. It will be also important to contrast the use of different scales for rating confidence. Here, we elected to use a binary scale to facilitate the calculation of the signal detection indices, but we acknowledge that using a confidence scale with 3 or more ratings, may be more informative to distinguish, for instance, cases in which children are aware of their mistake vs when they are unsure on the correctness of their response. However, using confidence scales with more ratings would also mean to increase the number of trials per participant. Moreover, it will be relevant to assess whether training metacognition by providing feedback on the calibration of participants' metacognitive judgments can modulate the role of metacognition in regulating long term learning. Ongoing work in the lab is being directed to test this possibility. Finally, additional studies employing bigger sample sizes using similar longitudinal designs tracking children across critical age ranges (i.e., 8 to 12) should be carried out to define the role of metacognition more generally in the development of non-linguistic skills (i.e., math) and how one's early metacognition predicts subsequent academic achievement.

Understanding the role of metacognition in monitoring and controlling cognitive and behavioural performance, and in guiding learning during childhood can have important implications in designing educational programs for fostering and optimising metacognitive strategies for a given individual and promoting lifelong learning and self-improvement. Further research is needed to understand whether these programs can enhance metacognitive ability as a transferable skill across distinct domains of learning and education, or whether unique domains should be targeted separately.

Declarations

Funding: This project was funded by the “la Caixa” Foundation INPhINIT fellowship program, by the Basque Government through the BERC 2018-2021 program, by the Spanish State Research Agency through the Severo Ochoa programme grant awarded to the BCBL (SEV-2015-0490), project grants PID2019-105494GB-I00 and RTI2018-096242-B-I00, and also by an RYC-2015-17356 grant from MINECO.

Conflict of interest: The authors declare that they have no conflicts of interest.

Ethics approval: The present study involving human participants was approved by the Basque Center on Cognition Brain and Language Ethics Board and the Bioethics Commission of the University of Barcelona.

Availability of data and material: Raw data and material are available in OSF (https://osf.io/gpmv6/).

Code availability: Analysis scripts are available in OSF (https://osf.io/gpmv6/).

Consent to participate: Fully informed consent forms were obtained from the legal tutors of the minors and oral consent from the children prior to the study.

Consent for publication: All authors consent to the publication of the present manuscript.

References

Ackerman, R., & Koriat, A. (2011). Response latency as a predictor of the accuracy of children’s reports. *Journal of Experimental Psychology: Applied*, *17*(4), 406–417. https://doi.org/10.1037/a0025129

Arciuli, J., & Simpson, I. C. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science*, *36*(2), 286–304. https://doi.org/10.1111/j.1551-6709.2011.01200.x

Bellon, E., Fias, W., & De Smedt, B. (2019). More than number sense: The additional role of executive functions and metacognition in arithmetic. *Journal of Experimental Child Psychology*, *182*, 38–60. https://doi.org/10.1016/j.jecp.2019.01.012

Bosse, M.-L., Tainturier, M. J., & Valdois, S. (2007). Developmental dyslexia: The visual attention span deficit hypothesis. *Cognition*, 104(2), 198–230. https://doi.org/10.1016/j.cognition.2006.05.009

Bosse, M.-L., & Valdois, S. (2009). Influence of the visual attention span on child reading performance: a cross-sectional study. *Journal of Research in Reading,* 32 (2), 230–253. https://doi.org/10.1111/j.1467-9817.2008.01387.x

Boukadi, M., Potvin, K., Macoir, J., Jr Laforce, R., Poulin, S., Brambati, S. M., & Wilson, M. A. (2016). Lexical decision with pseudohomophones and reading in the semantic variant of primary progressive aphasia: A double dissociation. *Neuropsychologia*, *86*, 45–56. https://doi.org/10.1016/j.neuropsychologia.2016.04.014

Brod, G., Bunge, S. A., & Shing, Y. L. (2017). Does One Year of Schooling Improve Children’s Cognitive Control and Alter Associated Brain Activation? *Psychological Science*, 28(7), 967–978. https://doi.org/10.1177/0956797617699838 de Bruin, A. B. H., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology*, 109(3), 294–310.

Chetail, F. (2017). What do we do with what we learn? Statistical learning of orthographic regularities impacts written word processing. *Cognition*, *163*, 103–120. https://doi.org/10.1016/j.cognition.2017.02.015

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective Confidence Predicts Information Seeking in Decision Making. *Psychological Science*, *29*(5), 761–778. https://doi.org/10.1177/0956797617744771

Destan, N., Hembacher, E., Ghetti, S., & Roebers, C. M. (2014). Early metacognitive abilities: the interplay of monitoring and control processes in 5- to 7-year-old children. *Journal of Experimental Child Psychology*, *126*, 213–228. https://doi.org/10.1016/j.jecp.2014.04.001

De Vos, T. (1992). Tempo Test Rekenen Nijmegen. *The Netherlands: Berkhout*.

Dunlosky, J., Mueller, M. L., Morehead, K., Tauber, S. K., Thiede, K. W., & Metcalfe, J. (2021). *Why Does Excellent Monitoring Accuracy Not Always Produce Gains In Memory Performance? Zeitschrift für Psychologie, 229(2), 104–119.* https://doi.org/10.1027/2151-2604/a000441

Ehri, L. C. (2014). Orthographic Mapping in the Acquisition of Sight Word Reading, Spelling Memory, and Vocabulary Learning. *Scientific Studies of Reading: The Official Journal of the Society for the Scientific Study of Reading*, *18*(1), 5–21. https://doi.org/10.1080/10888438.2013.819356

Filevich, E., Forlim, C. G., Fehrman, C., Forster, C., Paulus, M., Shing, Y. L., & Kühn, S. (2020). I know that I know nothing: Cortical thickness and functional connectivity underlying meta-ignorance ability in pre-schoolers. *Developmental Cognitive Neuroscience*, *41*, 100738. https://doi.org/10.1016/j.dcn.2019.100738

Finn, B., & Metcalfe, J. (2014). Overconfidence in children's multi-trial judgments of learning. In *Learning and Instruction*, 32, 1–9. https://doi.org/10.1016/j.learninstruc.2014.01.001

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *The American Psychologist*, *34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Fleming, S. M. (2017). HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *2017*(1), nix007. https://doi.org/10.1093/nc/nix007

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. https://doi.org/10.3389/fnhum.2014.00443

Frith, U. (1985). Beneath the surface of developmental dyslexia. *Surface Dyslexia: Neuropsychological and Cognitive Studies of Phonological Reading*. https://ci.nii.ac.jp/naid/10022405906/

Geurten, M., Meulemans, T., & Lemaire, P. (2018). From domain-specific to domain-general? The developmental path of metacognition for strategy selection. *Cognitive Development*, 48, 62–81. https://doi.org/10.1016/j.cogdev.2018.08.002

Ginestet, E., Phénix, T., Diard, J., & Valdois, S. (2019). Modeling the length effect for words in lexical decision: The role of visual attention. *Vision Research*, *159*, 10–20. https://doi.org/10.1016/j.visres.2019.03.003

Ginestet, E., Shadbolt, J., Tucker, R., Bosse, M., & Hélène Deacon, S. (2021). Orthographic learning and transfer of complex words: insights from eye tracking during reading and learning tasks. In *Journal of Research in Reading*, 44(1), 51–69. https://doi.org/10.1111/1467-9817.12341

Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition*, *37*(7), 1001–1013. https://doi.org/10.3758/MC.37.7.1001

Kelly, M., Scholnick, E. K., Travers, S. H., & Johnson, J. W. (1976). Relations among memory, memory appraisal, and memory strategies. *Child Development*. https://www.jstor.org/stable/1128179

Kuhn, D. (2000). Metacognitive Development. *Current Directions in Psychological Science*, 9(5), 178–181. https://doi.org/10.1111/1467-8721.00088

Lecce, S., Caputi, M., & Pagnin, A. (2015). False-belief understanding at age 5 predicts beliefs about learning in year 3 of primary school. *The European Journal of Developmental Psychology*, *12*(1), 40–53. https://doi.org/10.1080/17405629.2014.949665

Lockl, K., & Schneider, W. (2007). Knowledge about the mind: links between theory of mind and later metamemory. *Child Development*, *78*(1), 148–167. https://doi.org/10.1111/j.1467-8624.2007.00990.x

Lyons, K. E., & Ghetti, S. (2010). Metacognitive Development in Early Childhood: New Questions about Old Assumptions. In *Trends and Prospects in Metacognition Research*, 259–278. https://doi.org/10.1007/978-1-4419-6546-2_12

Lyons, K. E., & Ghetti, S. (2011). The development of uncertainty monitoring in early childhood. *Child Development*, 82(6), 1778–1787.

Maniscalco, B., & Lau, H. (2011). On a distinction between detection and discrimination: metacognitive advantage for signal over noise. *Journal of Vision*, 11 (11), 163. https://doi.org/10.1167/11.11.163

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition* 21 (1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Mano, Q. R., & Kloos, H. (2018). Sensitivity to the Regularity of Letter Patterns Within Print Among Preschoolers: Implications for Emerging Literacy. *Journal of Research in Childhood Education*, 32 (4), 379–391. https://doi.org/10.1080/02568543.2018.1497736

Martinet, C., Valdois, S., & Fayol, M. (2004). Lexical orthographic knowledge develops from the beginning of literacy acquisition. *Cognition*, *91*(2), B11–B22. https://doi.org/10.1016/j.cognition.2003.09.002

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: implications for studies of metacognitive processes. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(2), 509–527. https://doi.org/10.1037/a0014876an

Mazancieux, A., Fleming, S. M., Souchay, C., & Moulin, C. (2020). Retrospective confidence judgments across tasks: domain-general processes underlying metacognitive accuracy. *PsyArXiv.* https://doi.org/10.31234/osf.io/dr7ba

Metcalfe, J. (2009). Metacognitive Judgments and Control of Study. *Current Directions in Psychological Science*, *18*(3), 159–163. https://doi.org/10.1111/j.1467-8721.2009.01628.x

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: the effects of generation, errors, and feedback. *Psychonomic Bulletin & Review*, *14*(2), 225–229. https://doi.org/10.3758/BF03194056

Metcalfe, J., & Shimamura, A. P. (1994). *Metacognition: Knowing about Knowing*. MIT Press.

Meuwissen, A. S., Anderson, J. E., & Zelazo, P. D. (2017). The creation and validation of the Developmental Emotional Faces Stimulus Set. *Behavior Research Methods*, *49*(3), 960–966. 10.3758/s13428-016-0756-7

Narens, T. O. (1990). Metamemory: A Theoretical Framework and New Findings. *Psychology of Learning and Motivation*, 26, 125–173. https://doi.org/10.1016/s0079-7421(08)60053-5

Peirce J & Macaskill. (2019). Building Experiments in PsychoPy. *Perception*, 48 (2), 189–190. https://doi.org/10.1177/0301006618823976

Rawson, K. A., O'Neil, R., & Dunlosky, J. (2011). Accurate monitoring leads to effective control and greater learning of patient education materials. *Journal of Experimental Psychology: Applied*, *17*(3), 288–302. https://doi.org/10.1037/a0024749

Rinne, L. F., & Mazzocco, M. M. M. (2014). Knowing right from wrong in mental arithmetic judgments: calibration of confidence predicts the development of accuracy. *PloS One*, *9*(7), e98663. https://doi.org/10.1371/journal.pone.0098663

Roebers, C. M., Mayer, B., Steiner, M., Bayard, N. S., & van Loon, M. H. (2019). The role of children's metacognitive experiences for cue utilization and monitoring accuracy: A longitudinal study. *Developmental Psychology*, *55*(10), 2077–2089. http://dx.doi.org/10.1037/dev0000776

Roebers, C. M., Schmid, C., & Roderer, T. (2009). Metacognitive monitoring and control processes involved in primary school children's test performance. *The British Journal of Educational Psychology*, 79(Pt 4), 749–767.

Roebers, C. M., & Spiess, M. (2017). The Development of Metacognitive Monitoring and Control in Second Graders: A Short-Term Longitudinal Study. *Journal of Cognition and Development*, 18 (1), 110–128. http://dx.doi.org/10.1037/dev0000776

Ruby, E., Giles, N., & Lau, H. (2017). *Finding domain-general metacognitive mechanisms requires using appropriate tasks*. https://doi.org/10.1101/211805

Samaha, J., & Postle, B. R. (2017). Correlated individual differences suggest a common mechanism underlying metacognition in visual perception and visual short-term memory. *Proceedings. Biological Sciences / The Royal Society*, *284*(1867). https://doi.org/10.1098/rspb.2017.2035

Schoenfeld, A. H. (2016). Learning to Think Mathematically: Problem Solving, Metacognition, and Sense Making in Mathematics (Reprint). *Journal of Education*, 196 (2), 1–38. https://doi.org/10.1177/002205741619600202

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*(1), 204–221. https://doi.org/10.1037/0278-7393.26.1.204

Valdois, S., Roulin, J.-L., & Line Bosse, M. (2019). Visual attention modulates reading acquisition. *Vision Research*, *165*, 152–161. https://doi.org/10.1016/j.visres.2019.10.011

van den Bergh, D., Clyde, M. A., Gupta, A. R. K. N., de Jong, T., Gronau, Q. F., Marsman, M., Ly, A., & Wagenmakers, E.-J. (2021). A tutorial on Bayesian multi-model linear regression with BAS and JASP. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01552-2

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Gupta, A. R. K. N., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2020). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-020-01798-5

Vo, V. A., Li, R., Kornell, N., Pouget, A., & Cantlon, J. F. (2014). Young children bet on their numerical skills: metacognition in the numerical domain. *Psychological Science*, *25*(9), 1712–1721. https://doi.org/10.1177/0956797614538458

Vuorre, M., & Metcalfe, J. (2021). Measures of relative metacognitive accuracy are confounded with task performance in tasks that permit guessing. *Metacognition and Learning*. https://doi.org/10.1007/s11409-020-09257-1

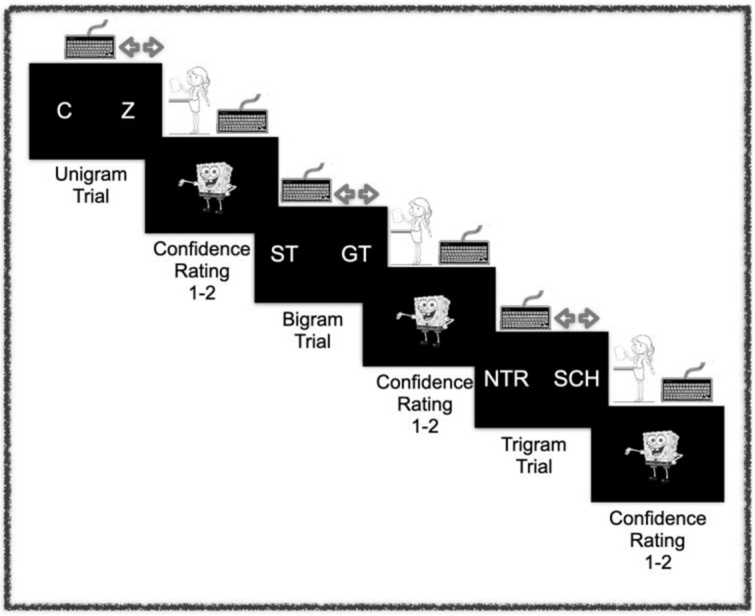
Wechsler, D. (2014). Wechsler intelligence scale for children–Fifth Edition (WISC-V). *Bloomington, MN: Pearson*.

Wickens, T.D. (2001). *Elementary Signal Detection Theory*. https://doi.org/10.1093/acprof:oso/9780195092509.001.0001

Supplemental Material

Supplementary Figure 1

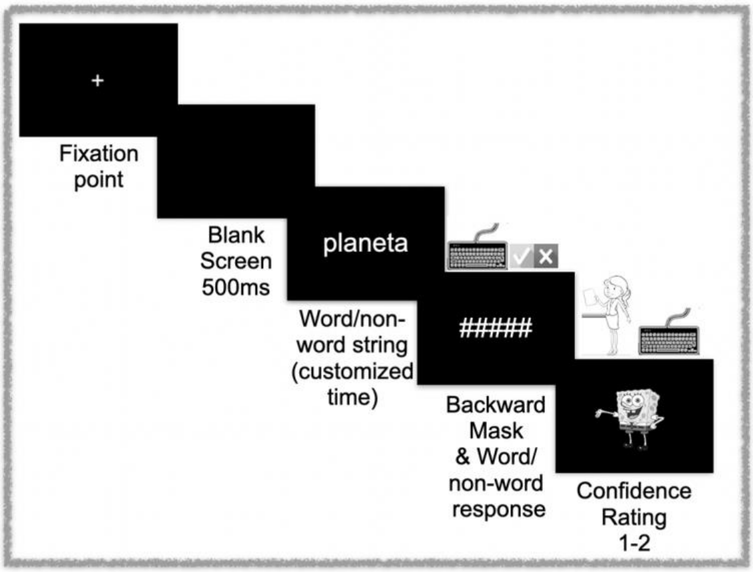
Task design of the Orthographic Statistical Learning task



Note. Participants saw pairs of ngrams (unigrams, bigrams, trigrams) presented on the screen and they had to choose the ngram with which they could form words easier and rate their confidence upon this response (type-2 task).

Supplementary Figure 2

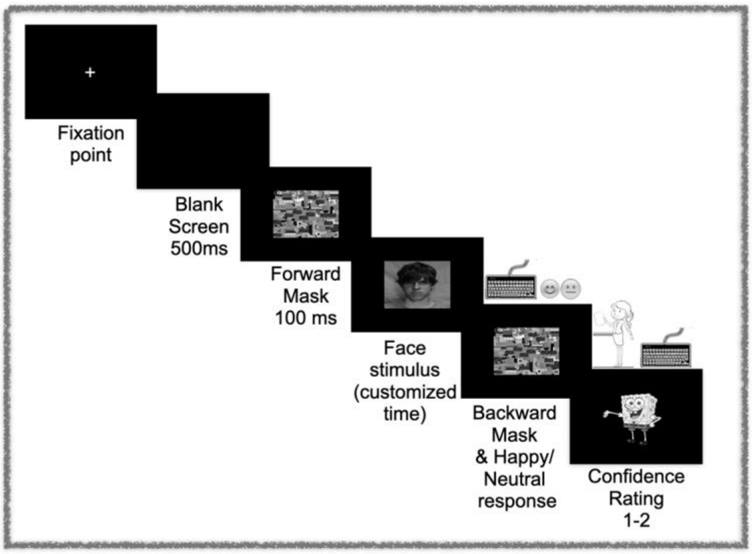
Task design of the Orthographic Lexical Decision task


+
Fixation point
Blank Screen 500ms
planeta
Word/non-word string (customized time)
✓ ✕
#####
Backward Mask & Word/ non-word response
Confidence Rating 1-2

Note. Participants saw a briefly presented sequence of letters that composed words or pseudowords. Participants had to decide on the identity of the sequence (word vs pseudoword, type-1 task) and rate their confidence upon this response (type-2 task).

Supplementary Figure 3

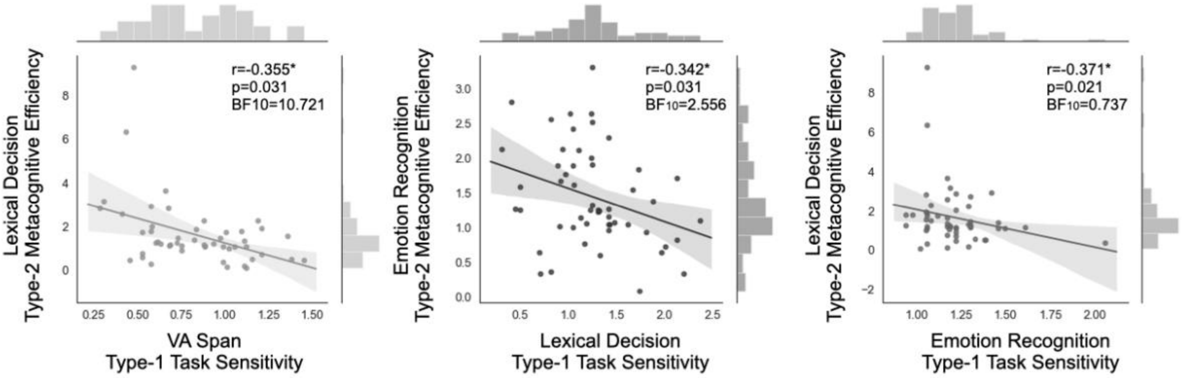
Task design of the Emotion Recognition Task



Note. Participants saw a briefly presented face that expressed a happy or neutral emotion. Participants had to decide on the emotion of the face presented (happy vs neutral, type-1 task) and rate their confidence upon this response (type-2 task).

Supplementary Figure 4

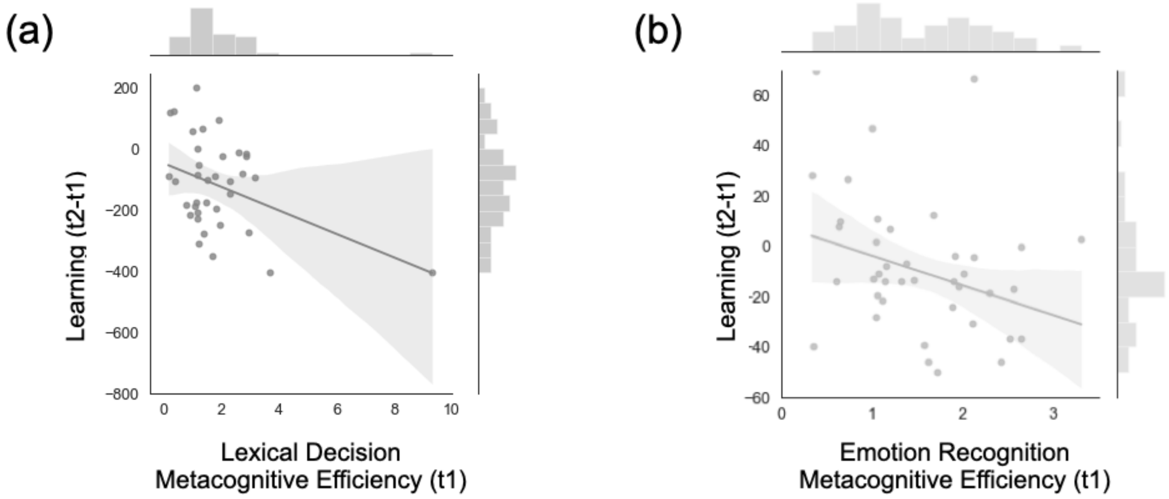
Correlations among students' metacognitive efficiency and type-1 task performance in the experimental tasks



Note. The present scatterplots indicate the significant Spearman correlations among type-2 metacognitive efficiency in experimental tasks and type-1 task sensitivity in the rest of the tasks (*p<0.05, **p<0.01, ***p<0.001, FDR corrected). Correlations were controlled for participant's age and intellectual ability (non-verbal IQ).

Supplementary Figure S5

Linear relationship between students' longitudinal learning and their early metacognitive efficiency within an experimental task



Note. Longitudinal learning was measured as the difference of the mean stimulus presentation duration between the two timepoints, and metacognitive efficiency was estimated in t1 (a: lexical decision task, b: emotion recognition task).

Supplementary Table S1

Descriptive statistics (Mean Score (SD), Range, Skewness) for the different measures of participants' type-1 and type-2 performance in the statistical learning tasks, both including all participants and excluding participants with accuracy less than 55%.

	Orthographic Statistical Learning		Visual Statistical Learning	
	All participants (N=58)	Accuracy≧0.55 (N=34)	All participants (N=58)	Accuracy≧0.55 (N=22)
Task accuracy (% of correct responses)				
Mean (SD)	0.59 (0.12)	0.67 (0.07)	0.53 (0.11)	0.64 (0.09)
Range	0.27-0.84	0.58-0.85	1.122	0.56-0.88
Skewness	–0.136	0.553	0.34-0.87	1.380
Type-1 task sensitivity (d'prime)				
Mean (SD)	0.46 (0.63)	0.88 (0.42)	0.18 (0.59)	0.74 (0.56)
Range	-0.068	0.36-1.92	-0.77-2.17	0.29-2.17
Skewness	-1.33-1.92	0.758	1.43	1.596
Type-2 Metacognitive Efficiency (Mratio)				
Mean (SD)	-102.64 (1501)	0.95 (0.76)	106.32 (1213)	0.33 (1.56)
Range	-7384-6704	-1.04-2.24	-4279.60 (7578.50)	-2.92-4.28
Skewness	-1.194	-0.593	3.624	0.090

Supplementary Table S2

Correlations between type-1 task sensitivity in the statistical learning tasks and students' reading performance

	Orthographic Statistical Learning Type-1 Task Sensitivity		**Visual Statistical Learning Type-1 Task Sensitivity**	
	All participants (N=58)	Accuracy≧0.55 (N=34)	All participants (N=58)	Accuracy≧0.55 (N=22)

Words Speed	r = -0.034
	p = 0.979
	BF10 = 0.203
		r = 0.160
		p = 0.902
		BF10 = 0.323
			r = 0.047
			p = 0.729
			BF10 = 0.187
				r = 0.128
				p = 0.855
				BF10 = 0.409

Pseudowords Speed	r = -0.122
	p = 0.979
	BF10 = 0.280
		r = 0.154
		p = 0.902
		BF10 = 0.344
			r = 0.096
			p = 0.713
			BF10 = 0.231
				r = -0.008
				p = 0.970
				BF10 = 0.304

VA Span Type-1 Task Sensitivity	r = -0.228
	p = 0.352
	BF10 = 0.417
		r = 0.126
		p = 0.902
		BF10 = 0.297
			r = 0.225
			p = 0.352
			BF10 = 0.716
				r = -0.182
				p = 0.975
				BF10 = 0.282

Lexical Decision Type-1 Task Sensitivity	r = -0.029
	p = 0.833
	BF10 = 0.301
		r = -0.009
		p = 0.989
		BF10 = 0.244
			r = 0.088
			p = 0.828
			BF10 = 0.172
				r = -0.102
				p = 0.975
				BF10 = 0.361

Emotion Recognition Type-1 Task Sensitivity	r = -0.029
	p = 0.833
	BF10 = 0.231
		r = -0.128
		p = 0.902
		BF10 = 0.290
			r = 0.075
			p = 0.828
			BF10 = 0.188
				r = -0.065
				p = 0.975
				BF10 = 0.281

Orthographic Statistical Learning Type-1 Task Sensitivity	-	-	r = -0.167
			p = 0.438
			BF10 = 0.213
				r = 0.006
				p = 0.987
				BF10 = 0.511

Visual Statistical Learning Type-1 Task Sensitivity	r = -0.167
	p = 0.438
	BF10 = 0.213
		r = 0.006
		p = 0.989
		BF10 = 0.511
			-	-

Note. Spearman's correlations were performed between type-1 task sensitivity (d'prime) in the Orthographic and Visual Statistical Learning tasks and a) students' performance on standardized tasks measuring reading ability, b) type-1 task sensitivity in the rest of the experimental tasks (*p<0.05, **p<0.01, ***p<0.001, FDR corrected). Correlations were controlled for participant's age and intellectual ability (non-verbal IQ, Matrices-WISC).

Supplementary Table S3

Correlations between metacognitive efficiency in the statistical learning tasks and students' reading performance

	Orthographic Statistical Learning Type-2 Metacognitive Efficiency		**Visual Statistical Learning Type-2 Metacognitive efficiency**	
	All participants (N=58)	Accuracy≧0.55 (N=34)	All participants (N=58)	Accuracy≧0.55 (N=22)

Words Speed	r = -0.003
	p = 0.979
	BF10 = 0.182
		r = -0.132
		p = 0.528
		BF10 = 0.215
			r= 0.129
			p= 0.670
			BF10 = 5.411
				r = -0.055
				p = 0.970
				BF10 = 0.308

Pseudowords Speed	r = 0.009
	p = 0.979
	BF10 = 0.164
		r = -0.112
		p = 0.528
		BF10 = 0.213
			r = 0.148
			p = 0.670
			BF10 = 5.678
				r = -0.149
				p = 0.855
				BF10 = 0.488

VA Span Type-1 Task Sensitivity	r = -0.078
	p = 0.828
	BF10 = 0.411
		r = -0.088
		p = 0.855
		BF10 = 0.258
			r = 0.107
			p = 0.859
			BF10 = 0.266
				r = 0.238
				p = 0.783
				BF10 = 0.292

Lexical Decision Type-1 Task Sensitivity	r = 0.007
	p = 0.958
	BF10 = 0.174
		r = 0.034
		p = 0.855
		BF10 = 0.272
			r = -0.025
			p = 0.859
			BF10 = 0.198
				r = 0.001
				p = 0.977
				BF10 = 0.311

Emotion Recognition Type-1 Task Sensitivity	r = -0.027
	p = 0.933
	BF10 = 0.175
		r = 0.257
		p = 0.521
		BF10 = 0.544
			r = -0.128
			p = 0.852
			BF10 = 0.725
				r = 0.056
				p = 0.997
				BF10 = 0.302

Orthographic Statistical Learning Type-1 Task Sensitivity	-	-	r = -0.039
			p = 0.859
			BF10 = 1.106
				r = 0.576
				p = 0.408
				BF10 = 3.391

Visual Statistical Learning Type-1 Task Sensitivity	r = -0.171
	p = 0.518
	BF10 = 0.184
		r = -0.200
		p = 0.855
		BF10 = 0.394
			-	-

Note. Spearman's correlations were performed between type-2 Metacognitive efficiency (Mratio) in Orthographic and Visual Statistical Learning tasks and a) students' performance on standardized tasks measuring reading ability, b) type-1 task sensitivity in the rest of the experimental tasks (*p<0.05, **p<0.01, ***p<0.001, FDR corrected). Correlations were controlled for participant's age and intellectual ability (non-verbal IQ, Matrices-WISC).

Supplementary Table S4

Correlations between the percentage (%) of incorrect responses rated with low confidence in the experimental tasks and students' reading performance

	Standardised reading tasks		***Linguistic tasks related to orthographic lexical processing***		***Non-linguistic task***
	Words Speed	Pseudowords Speed	VA Span Type-1 Task Sensitivity	Lexical Decision Type-1 Task Sensitivity	Emotion Recognition Type-1 Task Sensitivity
VA Span % of incorrect responses rated with low confidence	r = -0.198 p = 0.242 BF10 = 0.555	r = -0.225 p = 0.172 BF10 = 0.493	r = -0.329* p = 0.033 BF10 = 1.385	r = -0.086 p = 0.706 BF10 = 0.218	r = 0.041 p = 0.851 BF10 = 0.180
Lexical Decision % of incorrect responses rated with low confidence	r = -0.333* p = 0.033 BF10 = 9.574	r = -0.283 p = 0.073 BF10 = 5.632	r = -0.431** p = 0.005 BF10 = 7.087	r = -0.436** p = 0.004 BF10 = 61.060	r = -0.052 p = 0.851 BF10 = 0.222
Emotion Recognition % of incorrect responses rated with low confidence	r = -0.109 p = 0.546 BF10 = 0.286	r = -0.174 p = 0.277 BF10 = 0.377	r = -0.250 p = 0.128 BF10 = 0.743	r = -0.151 p = 0.385 BF10 = 0.234	r = -0.000 p = 1.000 BF10 = 0.162

Note. Spearman's correlations were performed between the percentage (%) of incorrect responses rated with low confidence in the experimental tasks and a) students' performance on standardized tasks measuring reading ability, b) type-1 task sensitivity in the rest of the experimental tasks (VA Span task: N=55, lexical decision task: N=55, Emotion Recognition Task: N=59, *p<0.05, **p<0.01, ***p<0.001, FDR corrected). Correlations were controlled for participant's age and intellectual ability (non-verbal IQ, Matrices-WISC).