Optimising adaptivity in online learning environments

OPTIMISING ADAPTIVITY IN ONLINE LEARNING ENVIRONMENTS

Susanne M.M. de Mooij

Thesis submitted for the degree of
Doctor of Philosophy

Department of Psychological Sciences
Birkbeck College, University of London

January 2021

DECLARATION

I, Susanne M.M. de Mooij, hereby declare that the work presented in this thesis is my own. Where explicit attribution is made, I confirm that this has been indicated explicitly in the thesis.

ACKNOWLEDGEMENTS

Starting my PhD in London was an adventure: to live in a new country without any friends, explore a huge international lab and meet my two supervisors, Iroise Dumontheil and Natasha Kirkham. From the first day onwards, it was clear to me that I would get along with my supervisors on a friendly basis, but at the same time have fierce discussions. Their supervision and criticism, as well as the freedom to explore ideas, have shaped me in the best way possible. I was lucky to have support and supervision from familiar faces, Han van der Maas and Maartje Raijmakers, as well. Even though they had the official role of industrial partners, they have exceeded this role as experienced professors by being involved in every step along the PhD.

The staff from the CBCD lab made me very welcome during my stay in London with memorable Christmas parties (with compulsory karaoke sessions), wine and cheese nights and cultural events. All the seminars and research encounters have expanded my knowledge on babies (in particular) and children as well as all shown me the ins and outs of running a study. Leslie, Berta, Denis, Fred, Tamsin and Isabel to name a few, were always there to smooth over any problems. I am grateful for Birkbeck College to host the PhD position and my funding from the European Union's Horizon 2020 Marie Sklodowska-Curie Innovative Training Network for the opportunity to develop myself as a researcher. One of the neat parts of this scheme was attending and organising workshops all over Europe with my fellow PhD students. Part of this development was also to spend half of my time with the industrial partner Oefenweb (now known as Prowise). Even though I was a visiting researcher, the whole Oefenweb team included me in all the social activities and brainstorm sessions for future directions of the company. They supported my research projects where they could, and took my mind off the projects whenever necessary, with hacky sack or table tennis sessions as one my favourite escapes.

The first few minutes I felt alone in London, but this quickly changed when I got to spend the first two weeks in an airbnb with my fellow PhD students, Cécile, Louisa, Giada and Alicja. From the first night after too much wine, when we were still working on our presentations for the next day, I considered them as my partners in crime on this rollercoaster. Quickly I got to meet other like-minded people that I could add to my favourite crew: Emily, Kathryn, Jessica, Cosmin, Sam, Matt, Georgina and Ana. At the end, we definitely needed big tables for lunch gatherings. Thank you all for being a friend and making my time in London unforgettable.

Lastly, I thank my family and friends at home for their encouragement and support to achieve my potential, but also to reassure me when I had done enough, and to strive for a good work and life balance. Of course, I should thank my partner Raisa as well, although she probably hates being mentioned here. The last few months were intense with the COVID lockdown causing us to be in the same house all day all night. But you have helped me get through the writing with hugs, good food and kicks in my butt.

ABSTRACT

Over the past two decades, the demand for online adaptive educational technology has been rising due to the promise that learning can be tailored to students' needs. To tailor and optimise adaptivity to individual differences in learning effectively, we need to investigate what predicts students' performance. This thesis includes experiments conducted in an online adaptive learning environment, called Learn, played daily by 180,000 primary school children. Adaptive learning is fun and challenging for children because, by definition, it adapts the difficulty of mathematical and language problems to their ability, preventing boredom or frustration. This adaptivity can be optimised in two ways: (1) by using theories and insights from cognitive research on executive function, strategy use and error monitoring to identify key individual cognitive differences when learning; (2) by inferring the involvement of specific cognitive processes or thought patterns using online measures. In **Chapter 2,** two online measures – eye and mouse tracking - are discussed and compared. These measures were used in two of the three studies described in this dissertation, both in a small sample in the lab and a large study in the Learn platform. **Chapter 3** reports a lab study with primary school children assessing whether individual differences in cognitive abilities (working memory and inhibitory control) predict arithmetic performance. This study showed that children's cognitive profile does have an impact on arithmetic performance, and interacts with features of the gamified environment, namely the visibility of time pressure, a key aspect of many online educational tasks. Eye fixations revealed differences in children's attention towards the question and distracting errors between when time pressure was visible and when it was not. Both eye and mouse tracking gave some insight into children's thinking and strategy use during their performance of the arithmetic task. **Chapter 4** presents the analysis of mouse movements of children playing in a learning environment. The mouse movements reflected children's arithmetic difficulties during problem-solving. This introduced a promising way to predict false associations children might have without the need for them to make errors, which are typically maintained at a low level in an online learning environment. Children do make mistakes while learning, and a key element of learning is for children to notice their errors and adjust their strategies to improve their performance. **Chapter 5** shows evidence of adaptive behaviour in the form of post-error slowing (PES), the finding that humans slow down their performance after an error. PES was observed in a range of online mathematical and language tasks. Individual and task-related factors influencing the presence or magnitude of PES were identified. Overall, the studies identify key predictors of student's performance as well as suggest new online measures to infer individual differences in these predictors, which is needed to fulfil the promise of tailoring the learning experience to individual children.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

WM	Working memory
IC	Inhibitory control
EF	Executive function
ERN	Error-related negativity
SES	Socio-economic status
MD	Mathematical disability
DD	Developmental dyscalculia
RT	Reaction time
WCST	Wisconsin Card Sort Task
DCCS	Dimensional Change Card Sort task
PE	Error-related positivity
PES	Post-error slowing
CAT	Computer adaptive testing
IRT	Item response theory
MDev	Maximum deviation
AUC	Area under the curve
AOI	Area of interest
TOST	Two one-sided equivalence test
CI	Confidence interval
BF	Bayes Factor

1

BRIDGING ADVANCED COGNITIVE SCIENCE AND 21ST CENTURY LEARNING ENVIRONMENTS

1.1 Introduction

Learning is a complex system that is challenging to capture, as many elements interact with each other (van der Maas et al., 2014). To create a meaningful learning experience for children, it is essential to understand how, when and why a person learns. However, every child has a unique trajectory of cognitive development and learning, which makes it difficult for educational systems to optimise learning for everyone. The cognitive differences between children are considerable as early as pre-school (Bjorklund & Causey, 2017). Education aims to reduce differences due to inequality. At the same time, good education should also adapt to children's differences in cognitive abilities. A focus on individualised learning, i.e., learning tailored to the preferences and requirements of the individual student, requires great effort using traditional approaches when one teacher faces 30 children with varied prior knowledge and cognitive abilities (Corno & Snow, 1986). With the recent advancements in educational technology and particularly online computer-based systems, computerised learning activities can be carried out individually rather than as a whole classroom.

School closings, caused by the COVID-19 pandemic, have led to a significant demand for online learning software. Consequently, high-quality online education is more important than ever. Even before the start of the pandemic, there was a lot of growth and development in education technology, with global investments reaching $18.66 billion in 2019 (Adkins, 2020). This rising demand for technological solutions also brings new research opportunities, as technological innovations allow access to new methods of learning but also influence the child's response to a learning intervention.

Developmental and cognitive sciences have delivered theories of what benefits learning. This is essential to consider when designing the 21st-century learning environment. An online learning environment is a great experimental platform to test cognitive theories and methods, providing the opportunity for individualised interventions in the learning experience.

1.2 Aims and objectives

In this thesis, the aim is to integrate certain developmental psychological methods and theories, to uncover how and when to effectively adapt online practice systems to individual learners. In a conventional classroom setting with 30 children, it is difficult to get a good sense of the learner's individual needs. When children are tutored one-on-one this is much easier and therefore more effective for performance, with an effect size of d = 2.0 relative to classroom teaching without tutoring (Bloom, 1984). The advantage of an online learning context is that individual differences can be monitored through one-on-one tutoring. Once the key parameters of interest have been identified, the system can be customised automatically for large numbers. To accomplish this customisation, we need to identify key parameters for individual differences. During my PhD, I collected data (small samples) in the lab, where the conditions could be tightly controlled. I also conducted applied research with large samples in an online adaptive learning environment. The ultimate aim of collecting from both resources is to answer the question of what can predict children's performance. Many learning systems nowadays already adapt exercises and instruction to the general ability level of the student. By extending this approach with the focus on other individual differences in learning, we can optimise the adaptivity of the system. On the other hand, it is also a way to investigate if findings in learning theories can be generalised to a 'real-world' setting.

To summarise, this PhD work explores how cognitive research can inform optimisation of the design of an online learning environment and conversely how large online environments can be used to test cognitive theories.

This first chapter gives an overview of the relevant literature. I will first provide evidence that learning is different for each child and discuss the specificities of learning through an online platform (section 1.3). Next, I will focus on the development of arithmetic skills, the main skill that I investigated in my experimental studies (section 1.4). Individual differences in executive functions are discussed (section 1.5), as well as how these differences relate to mathematics performance (section 1.6). In section 1.7, the role of cognitive control and strategy use is emphasised in the development of executive function and mathematics skills. Finally, an outline is given of how the framework of adaptive environments is set up to tailor to individual differences (section 1.8). In this section, I showcase the first results of the online learning environment of Math Garden (more details in Chapter 2), which is the platform I used for the experiments in my PhD.

It should be noted that this review includes correlational studies and that some studies in this thesis included data from a single time point. Therefore, these studies will not inform our understanding of how individual differences and task features affect learning per se, i.e., changes over time. However, a better understanding of how performance in online tasks may be affected by these factors could allow tailoring of the environment to individual learners, making sure that the task challenges, and therefore trains their skills optimally.

1.3 Learning (in virtual environments)

'The illiterate of the 21st century will not be those who cannot read and write, but those who cannot learn, unlearn, and relearn" (Alvin Toffler, Future Shock, 1970).

According to learning theories, learning is an enduring change in behaviour, or in the capacity to behave in a given fashion, resulting from practise or other forms of experience (Schunk, 2012). One of the main tenets of learning theories is constructivism. Constructivism is based on the idea that a learner creates its vision of the world, shaped by its individual experiences and knowledge (Bruning et al., 2004, pp. 194–195). Therefore, the best way to learn is to actively construct one's own understanding and skills. Two ground-breaking researchers in constructivism that focused on cognitive development, were Piaget (1964) and Vygotsky (1978). A key component in Piaget's work was the idea that children's cognitive development progresses through stages, where each stage builds on the accomplishments of preceding stages. In his view, it is important to monitor a child's stage and adjust the learning materials to their developmental level accordingly, but that children should also discover on their own. According to Vygotsky, a learner needs to be challenged as well as be encouraged in their so-called "zone of proximal development" to advantage their skills and knowledge. He also believed that interacting with knowledgeable others (e.g., teachers), helps to progress.

1.3.1 *Unique learning experience*

Despite their differences, Piaget and Vygotsky agreed that a person's acquired knowledge and learning experience are unique because we all interpret information in a different way

(Ormrod, 2014). The individual experience is affected by various features of the learner that interact with the particular task that needs to be learned (Park & Lee, 2003). A first feature to consider is the intelligence of a child. Intelligence interacts with how effective instruction can be (Sternberg & Kaufman, 2013, Chapter 6). For example, students with low intellectual ability are known to be more supported by structure and less complex instruction, whereas children with high intellectual ability thrive with less structure and can handle more complex instruction, also known as discovery learning (Snow & Lohman, 1984). Also, the prior knowledge of the child predicts their immediate learning success. Children with a higher level of prior knowledge seem to require less instructional support to accomplish the task (e.g. Hailikari et al., 2008; Koedinger & Corbett, 2012; Tobias, 1994). The difficulty of the learning materials should be adapted to their prior achievement and intelligence to make it challenging for any child.

At the same time, the pace of practising should not only be challenging but also remain fun. Otherwise, children would stop practising their skills. Ideally, students experience a state of flow, where they are concentrated and absorbed in mastering skills that they are intrinsically interested in (Csikszentmihalyi, 1990). Many researchers emphasize the positive impact of motivation on learning (for a review, see: Gopalan et al., 2017). Especially intrinsically motivated students achieve higher performance, not necessarily students that are extrinsically motivated through rewards (Larson & Rusk, 2011). However, a lack of motivation, translated to boredom or frustration, is detrimental to a child's productivity and their will to practice (D'Mello & Graesser, 2012).

A child's motivation to work on their academic achievements can be impacted by their personality characteristics, such as self-perceived competence and level of test anxiety (Elliot & Dweck, 2013; Ferla et al., 2010). Students with high test anxiety tend to perform poorly on tests in comparison to children with low test anxiety, although it is not always clear whether this difference is due to learning (Sieber et al., 1977, Chapter 5). The circumstances of how tests are administered can make an impact on the learning experience (Sorvo et al., 2019). For example, where a test has time limit or not, as well as whether tasks are adapted to the performance level to minimise failure. Here, the impact of the child's cognitive ability on learning needs to be also considered, which will be discussed in section 1.5. All these factors contribute to individual variance in mathematical skills, the main focus in this thesis, but also in reading and science (Ackerman, 1990).

Moreover, the context in which students learn, such as their culture, school, and teacher's style, contribute to their unique experience (Cormier & Hagman, 1987). Culture can explain differences in learning styles, such as the preference for abstract learning in high-in-group collectivism cultures, as well as avoidance of uncertainty (Joy & Kolb, 2009; Lim, 2004; Nisbett, 2004). Evidently, the school setting and teachers play an important role in students' involvement and engagement, which in turn contribute to learning (Klem & Connell, 2004; Skinner et al., 1993, 2008). This will be discussed in more detail in section 1.4.3. Since the rise of online learning platforms, there is a new context to consider and that is whether the materials are presented online.

1.3.2 *Online learning setting*

The new generation of students, the so-called "Net generation", are radically different from those of the past because they grew up with digital technology (Prensky, 2001). Children and adolescents spend an increasing time behind a screen, and this caused a discussion about its impact on student's learning (Thompson, 2013). In the UK in 2019, 5-15 year-olds spent around 2 hours 11 minutes per day online and 1 hour 52 minutes watching TV (Ofcom, 2020). Some worry that being surrounded by and immersed in technology makes the new generation incapable of productive work or deep learning and that they are less keen on exploring new information and knowledge (e.g. Bauerlein, 2008). Others are more optimistic and claim that the new generation is much more able to respond fast, be creative, multitask and be involved in their learning when given appropriate instruction (Palfrey et al., 2009; Rosen, 2010). In truth, there is not enough evidence to favour either of these views. Furthermore, as Lai and Hong (2015) pointed out, it cannot be claimed that a whole generation has homogeneous characteristics simply due to their technology use.

Nonetheless, more education takes place online nowadays. Numerous studies have looked at differences between learning in a classroom and online, mostly using serious games. Serious games are games that engage users and contribute to the achievement of a defined purpose other than pure entertainment (Susi et al., 2007). These studies focus mainly on the question of whether serious games are effective tools for learning and are motivating. Meta-analyses (Cheung & Slavin, 2013; Wouters et al., 2013) have shown that game-based learning has a positive effect on learning and retention compared to traditional teaching methods. More recently, Castellar et al. (2015; 2014) compared online maths games and traditional paper exercises in second graders (7-8 year-olds). Results showed an enhancement in terms of speed and accuracy performance through playing educational games, and that children rated learning in an educational game as more enjoyable. One of the rare large-scale longitudinal experiments had a similar setup by comparing online vs. paper maths games from grade 1 to grade 3 (Bakker et al., 2015). They also found that online game-based tasks positively affected children's mathematical skills and insights, but that it is most effective when supplemented with other instructional methods, for example a classroom discussion about strategy use (see also, Wouters et al., 2013). So far, a large majority of the serious games literature have focused on the learning potential of mathematics skills.

1.4 Mathematics skills development in children

Mathematics skills are arguably one of the most important skills for a child to master. It is also a subject in which children generally tend to underachieve. A large proportion of children (13% in the UK) does not reach the minimum required level at the end of primary school (Department for Education, 2016). A similar trend is reported in the US, where 12% of the children perform below the basic-proficiency level in mathematics (OECD, 2012). Examining the process of mathematics learning in more detail is essential for our understanding of what drives this large variation in performance and which factors may be limiting the children that are incapable of achieving basic maths proficiency.

1.4.1 *Arithmetic abilities*

Mathematics learning is a complex process. It is a subject that contains a lot of different fields, such as geometry, algebra and arithmetic. Within each field, there is a range of skills that need to be learned and connected. The earliest skill to practice at primary school is arithmetic, i.e., the ability to add, subtract, multiply and divide numbers. In turn, there are numerous abilities to master in arithmetic for a child to become an expert. As Dowker (2005, Chapter 1) noted about arithmetic: "there is no such thing as arithmetical ability: only arithmetical abilities". There are three main arithmetic abilities: (1) Factual knowledge, the knowledge about names for numbers and operations and arithmetic facts; (2) Procedural knowledge, the knowledge of the steps involved to carry out arithmetical operations; (3) Conceptual knowledge, which refers to knowledge of arithmetical principles, such as associativity, as well as word problems and approximate arithmetic (Dowker, 2005). These abilities can influence each other in a positive manner (Hofman, Kievit, et al., 2018). They are acquired with age through education and facilitated by the maturation of the brain and cognitive abilities. The mastering of mathematical skills requires motivation and many hours of practice, for example through repetition of arithmetical procedures (e.g., complex division) and rehearsal of arithmetical facts (e.g., multiplication tables).

1.4.2 *Development of arithmetic strategies*

How do children learn what strategy to use to solve a given problem, since there is such a variety to choose from? To acquire more arithmetic knowledge and skills, children need to know and use a lot of different strategies. During development, they gradually change their mix of strategies to calculate the answers to a problem (for reviews, see: Peters & de Smedt, 2018; Siegler, 1996). Children usually start with counting simple sums on their fingers, but gradually, a child employs more developmentally mature strategies (Geary et al., 1993). So, instead of counting the entire set, they start to count from the first number to counting from the larger number (Geary & Brown, 1991). By repeating these steps, they start to store the answers, such that the answers become arithmetic facts. The development of direct retrieval of arithmetic facts from memory is important to make problem-solving more efficient and less error-prone than counting strategies.

Different models of arithmetic fact retrieval exist (for an evaluation, see McCloskey et al. 1991), such as the well-known associations model introduced by Siegler & Shrager (1984). According to this associations model, when retrieving the answer from memory to a particular number problem (e.g., 3 + 6) associative links of related problems and potential answers are automatically activated, so not only 9 but also 8, 7, 3 and 18 (see also, Siegler, 1996). Depending on the associative strengths of these potential answers, so how differentiated the strengths are, it is more difficult for the child to retrieve the correct answer. Similarly, the network interference model of Campbell (1987, 1995) and the network retrieval model from Ashcraft & Battaglia (1978) postulate that children develop an arithmetic network of associations between arithmetic problems and potential answers, both correct and incorrect answers. During fact retrieval, these multiple associations compete, but through practice, the correct fact becomes most strongly activated.

Another example of arithmetic development can be found in the context of how children enumerate small or large number of items. Studies have found that there are three kinds of strategies that become evident: subitising, counting, and estimating – or a combination of these processes (e.g. Schleifer & Landerl, 2011). Subitising refers to the rapid and accurate enumeration of a set of elements, typically between 1 and 3 or 4, using pattern recognition (Kaufman et al., 1949). Jansen et al. (2014) examined this in the online environment of Learn and found that children who subitised small sets of numbers, the more sophisticated strategy, had higher math performance than children who only counted or estimated.

Children’s strategy use is defined by its efficacy. Successful strategies, associated with correct solutions are more likely to be used in the future and become more strengthened in memory, whereas earlier arithmetic strategies are inhibited (Hui & Lee, 2009). Old strategies do remain available across development, even in adulthood (Lefevre et al., 1996). Important to note is that development in strategy use is not an abrupt shift from one strategy to another. Children tend to use a mix of strategies, rather than just one strategy. Following this, Siegler (1988) states that developmental changes can be seen as changes in the distribution of strategy use rather than plain substitutions of one strategy for another. In the “overlapping waves” model, Siegler (1996) posits that strategies are waves that rise and fall during development. Once a new strategy is discovered it does not simply replace the older one: Rather, children slowly change the frequency of using the old and new strategy. Strategy changes are also accompanied by changes in brain activity (Peters & de Smedt, 2018). A study by van der Ven et al. (2012) confirmed the overlapping waves model in Learn using a latent growth model in single-digit multiplication problems.

1.4.3 *Individual differences in arithmetic*

The strategy trajectories contribute to the huge variance between children in factual, procedural and conceptual knowledge, as well as the relative strengths between arithmetic abilities within children (Dowker, 2005). The individual differences are such that a 7-year-old can perform at the same level as the ‘average’ 11-year-old, and at a higher level than some 14-year-olds (Cockcroft, 1982). In the Learn environment these differences in math abilities are can be found in all primary school years (Straatemeier, 2014). **Fig. 1.1** is an example of large individual differences in solving multiplication problems. The graph shows the proportion of times a child was presented with easy problems on the left (e.g., 6 x 1) and more difficult problems on the right (e.g., 19 x 50), separated for each school Year. The width of the distributions illustrates that within each Year group children’s abilities vary considerably, and the overlap between year groups further highlights large individual differences in the learning trajectories of maths. The large individual differences also illustrate how difficult it is for teachers to teach mathematics in a large classroom with varying abilities.

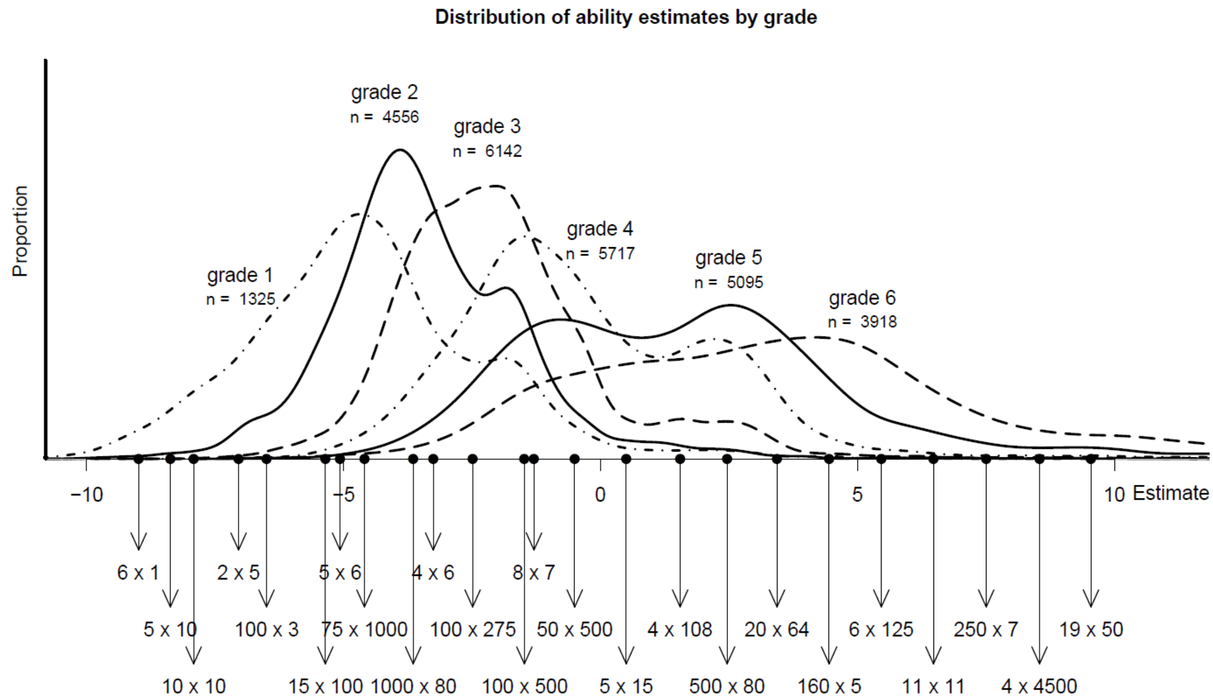


Figure 1.1*. Distribution of the difficulty of problems presented to children in the adaptive e-learning environment called Learn (Straatemeier, 2014) as a function of school Year. The horizontal axis displays the difficulty of certain multiplication items (e.g., '8 x 7'), calculated based on the performance of the whole sample. The distributions represent the proportion of times this item is presented to children per school Year to maintain the accuracy of around 75% for each child.*

Studies have pointed out different sources responsible for the individual differences in arithmetic performance. Children from low socio-economic status (SES) backgrounds start school behind their peers in the early grades, mostly in the verbal aspects of mathematics (Cheadle, 2008; Demir-lira et al., 2016; Duncan & Magnuson, 2011; L. Elliott & Bachman, 2018). Duncan and Magnuson (2011) found that low-SES children scored 1.4 standard deviations lower in their math skills than high-SES children when entering preschool and that this gap persists through childhood. Similarly, the effectiveness of the teacher can make a difference. Already in pre-school, children with less effective teachers in promoting academic achievement, score 0.2-0.48 standard deviation lower than children with more effective teachers (Clotfelter et al., 2006; Nye et al., 2004; Rockoff, 2004). In addition, Nye et al. (2004) showed that these teacher effects are much larger in low-SES schools, suggesting that in low-SES schools it weighs heavier which teacher is giving the lessons than in high-SES schools.

In addition to the effects of the quality of teaching and social background, 3-7% of the school-aged children have some form of cognitive impairment interfering with their ability to learn knowledge and procedures in the mathematics domains (Fletcher et al., 2007; Geary, 2004). Children with dysfunctions in their mathematical learning broadly fall under the term mathematical disability (MD), for a review see Soares et al. (2018). A more specific neuropsychological deficit in MD is also known as developmental dyscalculia (DD), caused by deficient processing of basic numerical concepts (Manor & Gross-tsur, 2000; Moeller et al., 2012). Both genetic and environmental factors contribute to the development of learning disabilities in children. For example, for siblings of children with DD the prevalence of a math disability is ten times higher than in the general population (Shalev et al., 1998).

Both in typical and atypical development of mathematic abilities, executive function plays an important role (e.g., Bull & Lee, 2014; Raghubar et al., 2010; van der Ven, Kroesbergen, et al., 2012), which will be outlined below.

1.5 The development of executive function

1.5.1 *Unity or diversity: Executive function*

Executive function (EF) is a broad term referring to various high-level cognitive processes necessary to regulate and control thoughts and actions (Friedman & Miyake, 2017). A key finding in most cognitive developmental studies is the consistent association between executive functions skills and learning outcomes, such as academic achievement (e.g., Best, Miller, & Naglieri, 2011). This link is found for various ages and in children with and without specific learning disabilities (for a review, see: Müller et al., 2008). Three subcomponents of EF are most commonly studied, namely working memory, shifting and inhibitory control (Miyake et al., 2000). These are thought to form the building blocks for more complex EF skills such as problem-solving, planning, error monitoring, and meta-cognition (Diamond, 2013). The role of the executive system is to handle novel situations when there are no ‘automatic’ psychological processes in place (Miyake et al., 2000; Shallice & Burgess, 1996).

One of the major discussions in the EF literature is the unity and/or diversity of the underlying cognitive abilities (Friedman & Miyake, 2017). Some argue that EF is unitary, in the sense that it is domain-general with a common neural substrate, i.e., the frontal-parietal network (Niendam et al., 2012). Others argue that the low correlations between the tasks are a sign of fractionality of the EF tasks (Duncan et al., 1997; Miyake et al., 2000). Both unity and diversity should be given a place in the study of executive function and the frontal functions (Collette et al., 2005; Duncan et al., 1997). The unity and diversity of EF tasks are also reflected in development: basic EF skills, such as inhibition and working memory, emerge already in infancy, from the age of 7 to 12 months (P. Anderson, 2002; V. Anderson et al., 2011). different With age, EF components all seem to have different developmental trajectories that continue to mature throughout childhood and in some cases into adolescence (Huizinga et al., 2006). Therefore, some researchers also argue that the structure (i.e., unity or diversity) of the EF model changes with age, such that there is a common factor underlying EF in early childhood, and becomes more differentiated with age (Bardikoff & Sabbagh, 2017; Brydges et al., 2014; Wiebe et al., 2011).

Another discussion related to the unity of the function is whether EF can be reduced to a few components that support other developmental processes (Doebel, 2020). One argument that Doebel (2020) gives against defining EF as domain-general components is that far transfer through cognitive training has been limited so far. Numerous studies have tried to improve skills in other related domains through training EF skills with laboratory tasks, but there has not been much evidence of far transfer to date (for a meta-analysis, see: Kassai et al., 2019). Importantly, Karr et al. (2019) argue that there may not be only three components to consider as core executive functions. Miyake et al. (2000) that introduced the most dominant model in the literature, decided to focus on working memory, inhibitory control and shifting for practical reasons. If other tasks had been added, such as tasks of planning, or prospective memory, there is a big chance other results would have been found, resulting in different latent variables/components. Often, three-factor models of EF are rejected as a good model fit (Karr et al., 2019). It is therefore essential for developmental research studying EF to consider the specific goals and context that were asked in the study to make supporting claims for the development of specific cognitive abilities.

Therefore, the three core functions of EF, namely working memory, inhibitory control and shifting will be discussed in more detail. In section 1.6, the relationship between these core components and mathematics performance is discussed.

1.5.2 *Working memory*

Working memory (WM) is the ability to control, regulate, and actively maintain relevant information in mind over short periods without relying on external aids; it is essential for complex cognitive tasks (A. D. Baddeley, 1992; Best & Miller, 2010). WM can be assessed by a variety of tasks. The most commonly used instrument for WM is span tasks, such as the digit or word span task (for a review, see: Conway et al., 2005). Here, a participant hears or sees a sequence of digits or words and is asked to recall the sequence. This can be either forwards or backwards, meaning that the participant is asked to either recall the sequence in normal or in reverse order (which is considered to be more difficult). The task requires both remembering and manipulating the information.

It is widely recognised that working memory is closely linked to learning and school performance. A rudimentary form of WM emerges as early as infancy (Diamond & Goldman-Rakic, 1989). Throughout childhood, WM capacity increases and it continues to do so into adolescence (Crone et al., 2006; Gathercole et al., 2004; Montez et al., 2017).

1.5.3 *Inhibitory control*

Inhibitory control (IC) is the ability to prevent a response that is not relevant to the current task or situation (i.e., distracting stimuli or thoughts) and to control one's attention, focusing on what we choose and resist interference (Diamond, 2013; Miyake et al., 2000).

In its most basic form, inhibition is already present in the first year of life, including the inhibition of neonatal reflexes (Diamond, 1989). Inhibitory control continues to develop mostly throughout childhood and into adolescence (Durston et al., 2002; Garon et al., 2008; Welsh et al., 2006; Williams et al., 1999).

The most common tasks used to measure inhibitory control are tests such as the Stroop task (MacLeod, 1992), the go/no-go task (Donders, 1969), and the stop-signal task (Lappin & Eriksen, 1966). These all test the ability of a person to suppress the dominant response to a stimulus to achieve more adaptive goal-oriented behaviours. For a measure of interference conflict, i.e., the ability to suppress task-irrelevant information, the Flanker Task (Eriksen & Eriksen, 1974) and Simon task (Simon, 1969) are often used. Most studies make use of one task to measure IC. What is concerning is that these IC measures often show poor validity (for a meta-analysis, see: Duckworth & Kern, 2011) and reliability (Gärtner & Strobel, 2019; Hedge et al., 2018). Moreover, when studies do examine more tasks, there is often low or non-significant correlations between the commonly used inhibitory control tests (Enge et al., 2014; Gärtner & Strobel, 2019). For example, Paap and Greenberg (2013) observed only a correlation of r = -0.01 between the Simon and Flanker effect. The lack of converging validity can also explain why some associations cannot be found. For example, de Bruin and Sala (2018) did find effects of age on inhibition for the Simon task, but not for two flanker tasks. Effects can be very task-specific and an experimenter should consider the tasks carefully.

1.5.4 *Shifting*

Shifting, also known as task or set-switching, is the cognitive flexibility to shift attention between tasks and mental states (Miyake et al., 2000). Shifting involves two phases. In the first phase, a participant forms a representation of a task rule (e.g., press on the side of where the dot appears) and this rule is maintained in WM. In the second phase, the task rule changes (e.g., press on the opposite side of where the dot appears) and the participant has to shift to maintaining this rule in WM. These two rules can be more or less in conflict with each other, depending on the amount of conflict that the participant has to overcome (Garon et al., 2008). A common measure of shifting is the Wisconsin Card Sort Task (WCST), or the Dimensional Change Card Sort (DCCS) for younger children (Blair et al., 2010; Garon et al., 2008; Kirkham et al., 2003).

In the next section is discussed how these specific EF skills relate to children's performance in mathematics.

1.6 Individual differences in executive function and mathematics performance

1.6.1 *The role of WM load*

Previous studies have described the role of WM in children's abilities for reading (Gathercole & Pickering, 2000) as well as mathematics (for a review, see: Raghubar et al., 2010). In the particular for mathematics, numerous studies have shown that individual differences in WM capacity in various domains (verbal, numerical and visuospatial) are important predictors of mathematics achievement with age (Bull & Lee, 2014; Dumontheil & Klingberg, 2012; Friso-van den Bos et al., 2013; Peng et al., 2015; Raghubar et al., 2010). In addition, longitudinal studies showed that working memory predicts later mathematics achievement (de Smedt et al., 2009; Mazzocco & Kover, 2007; van der Ven, Kroesbergen, et al., 2012). Further evidence for the role of WM stems from children with WM deficits. A meta-analysis showed that children with WM deficits also display poor mathematics performance (David, 2012).

WM capacity can place an important constraint on both the acquisition of reasoning skills and the acquisition of knowledge (Baddeley, 1992; Eylon & Linn, 1988). This constraint is also known as the working memory overload hypothesis (Niaz & Logie, 1993), posing that too much information processing takes up working memory capacity and can overload the system. WM load is usually studied with a dual-task method, in which children have to solve a primary task, such as addition problems, while their WM is (over)loaded with a secondary task (Huang & Mercer, 2001). Children, in particular those with low WM capacity, are hindered by WM overload in their learning activities (Gathercole & Alloway, 2008). In online game-based learning environments, the dual-processing system is required when there are too many visual or noise distractions through multimedia elements (Huang, 2011; Kiili, 2005; Moreno & Mayer, 2003). For example, a dual system is created when there is both visual and verbal information that the learner needs to process. It becomes very difficult to keep both information streams in their working memory at the same time (Mayer & Moreno, 1998). Such a split-attention effect could create a cognitive overload on the WM capacity constraining their performance (A. D. Baddeley, 1992; Eylon & Linn, 1988)

1.6.2 *Associations with inhibitory control and mathematics performance*

Fewer studies have also looked at the role of IC in mathematics performance. The majority of these studies have found that IC skills predict mathematics performance in typically developing children, particularly in pre-school and primary school children (Bull, Johnston, & Roy, 1999; Bull & Scerif, 2001; Espy et al., 2004; St Clair-Thompson & Gathercole, 2006). Recent literature suggests that IC is the common mechanism through which you can suppress prior knowledge (misconceptions) and learn new theories or concepts (Brookman-Byrne, Mareschal, Tolmie, & Dumontheil, 2018). When faced with misconceptions, children need to suppress conflicting information on their knowledge to acquire new concepts (Mareschal, 2016). Inhibition is also involved in suppressing inappropriate strategies, such as addition when multiplication is required.

Due to these positive associations, researchers thought that improving EF skills will contribute to better academic performance. Therefore, studies started to investigate training inhibition skills (and EF skills in general), but little evidence has been found so far that IC training transfers to academic skills (Wass, Scerif, & Johnson, 2012). Bull and Scerif (2001) have also argued that inhibition can be domain-specific, so for example only number-related inhibition would be relevant for mathematics performance. Encouraging children to use their inhibitory control by delaying a response during problem-solving could still potentially benefit their performance (Roy et al., 2019; Wilkinson et al., 2020).

1.6.3 *Shifting costs with age*

In comparison to WM and IC, shifting has not been consistently related to academic achievement (Espy et al., 2004). A meta-analysis of Yeniad et al. (2013) does link shifting ability to arithmetic performance, as well as reading performance. Some studies also find that general mixing costs tend to improve in school-age children, but that specific switch costs remain relatively stable with age (Kray et al., 2004; Reimers & Maylor, 2005). For a more recent study of the same finding, see Peng, Kirkham & Mareschal (2018). Other studies did find that task-switching improved during childhood (Cragg & Nation, 2009; Huizinga & van der Molen, 2011). It is not yet clear whether shifting is an independent predictor of mathematics, or whether intelligence moderates the association (Yeniad et al., 2013)

In sum, the three most often studied “core” executive functions show development during childhood and their individual differences associated with arithmetic performance to a certain extent. These functions are also implicated in cognitive control functions, for adapting behaviour in changing task demands and environmental circumstances (Botvinick et al., 2001; Ridderinkhof et al., 2004). One of these situations occurs when children are monitoring their errors.

1.6.4 *Error Monitoring*

Children develop error monitoring abilities, i.e., the ability to detect and evaluate errors as well as adjust their performance accordingly, between 6 and 11 years of age (for reviews, see: Ferdinand & Kray, 2014; Tamnes et al., 2013). The ability to adapt behaviour to errors is one of the key elements of cognitive control, by alerting the cognitive system to increase control to achieve appropriate performance (Mcguire & Botvinick, 2010). The common neural marker for the error monitoring system is error-related negativity or ERN, which is automatically generated after all errors. This can be followed by an amplitude of error-related positivity (PE) typically after 500 ms post-response (Lyons & Zelazo, 2011). The amplitude of the PE is thought to be indicative of a person recognising and evaluating the error (Ficarella et al., 2019; Grützmann et al., 2014; Nieuwenhuis et al., 2001). A common behavioural metric for the ability to monitor the error and adjust their behaviour is the slowing down of the response following errors, also called post-error slowing (Rabbitt & Rodgers, 1977). This metric is the focus of the study in Chapter 5.

1.7 Cognitive control and strategy use

1.7.1 *Problem-solving strategies*

To understand the cognitive processes behind learning and to adapt feedback and instruction to individual needs, studies have looked at what kind of problem-solving strategies children use in a variety of different learning domains. And what is striking is that children know and use multiple strategies depending on the task characteristics (for an overview, see: Siegler, 1996). For example, to solve an arithmetic problem such as 8 + 5 children could use their memory to retrieve that the answer is 13; use a transformation, e.g., 8 + 2 + 3 = 13; or just count: 8 + 5 = 8..9..10..11). This variability in strategy choice allows them to be flexible in facing the challenge. The choice of strategy is adaptive, where people aim to choose the most efficient strategy, in terms of speed and accuracy, to solve each problem (Geary et al., 1993; Siegler, 1988).

The individual's choice for a strategy use is affected by task- and context related variables as well as personal features (Payne et al., 1988). An important determinant for strategy selection is the difficulty of the problem. In general, increasing the difficulty of problems promotes the use of more sophisticated and complex strategies, in order to increase efficiency while maintaining accuracy (Siegler, 1996). Rather than accuracy, someone can also decide to focus more on speed, for example due to time pressure. When only limited time is available, people tend to use more rapid heuristic strategies that require fewer operations (Payne et al., 1988; Rieskamp & Hoffrage, 2008). Even though heuristics use less information from the environment, they can still be efficient for certain problems and perform very well, sometimes even better than complex strategies (Bobadilla-suarez & Love, 2018; Gigerenzer & Gaissmaier, 2011)

1.7.2 *Cognitive control*

How strategies are chosen and how cognitive self-regulatory mechanisms in general function are some of the bigger questions in cognitive development. The cognitive capacity of the solver can be a limitation. In general, EF plays a role in deciding which strategy to use and to allocate the attentional resources. Especially when cognitive processes are not automated yet and effort is required (Imbo & Vandierendonck, 2007; Wu et al., 2008). We know that WM influences maths achievement by helping to keep track of relevant information during problem-solving, but WM is also involved in selecting and switching to the most efficient arithmetic strategy (Barrouillet & Lépine, 2005; Cragg & Gilmore, 2014; Siegler & Lemaire, 1997; Wu et al., 2008). Increasing the difficulty of a problem, e.g., more digits in a sum, often results in greater WM demands. For example, Barrouillet and Lépine (2005) have found that children with higher WM capacities solve addition problems more efficient in terms of speed, and had a higher percentage of retrieval use than children with low WM capacity. This difference in performance between children with low and high WM became more pronounced with a larger minimum addend. Further evidence of an association between individual differences in WM and optimal strategy use was provided by van der Ven et al. (2012). They showed that children with poor WM tend to choose immature strategies that require a high load on WM resources and are therefore more error-prone.

1.7.3 *Measures of strategy use*

A difficulty with studying strategy use is how to measure it. One method is to ask participants what kind of strategy they used to solve the problem (Carpenter & Moser, 1984). This is quite successful with adults but becomes more problematic with children since it relies on their ability to express themselves verbally as well as their metacognitive skills. Another approach is to observe the child during problem-solving, e.g., looking at signs of counting through lip movement and use of fingers. This becomes less useful with increasing age since children's behaviour during problem-solving becomes more internalised. A more quantitative approach is to use the reaction time (RT) needed to solve a problem. Mental processes differ in the amount of time they need. For instance, research has shown that adding 7 + 6 is significantly slower than adding 4 + 3 (Ashcraft, 1982), and these time differences can reveal information on mental operations. But this also assumes that children always use the same strategy on particular problems, which is not the case (Siegler, 1989). More promising ways to look at strategy use is through eye tracking and mouse tracking, which are discussed in Chapter 2 (section 2.2 and 2.3).

1.7.4 *Error processing*

A related process to strategy use is error processing. When solving problems, the students' error responses are thought to reflect their cognitive process and/or applied strategies (Ben-Zeev, 1998; Buwalda et al., 2016; Savi et al., 2018). In other words, a wrong type of strategy use can lead to errors. At the same time, making an error can influence strategy use since switching to different strategies after an error can lead to better performance (Borght et al., 2016). An understanding of someone's errors can allow the provision of better feedback adapted to the learner's needs. It is therefore essential to understand whether there is a pattern in someone's errors. An error can be categorised as systematic or unsystematic. An unsystematic error is a mistake or slip, defined to be an action that was not intended (Norman, 1981). This could be the result of being inattentive, sloppy or careless. The more interesting errors are the systematic ones, also known as rational errors. These are errors that are logically consistent and rule-based rather than being random (Ben-Zeev, 1998). Rational errors reflect the student consistently applying an incorrect procedure (J. S. Brown & Burton, 1978). Rational errors can be due to misconceptions a child has, for example related to a poor understanding of fractions. Rational errors can also be due to incorrect knowledge, such as a multiplication incorrectly remembered, $6 \times 8 = 46$. Being able to diagnose misconceptions or incorrect knowledge by analysing systematic difficulties can ultimately help in individualizing education to students' state of knowledge.

In a learning system, it is interesting to examine error patterns for adaptive feedback and instruction. Computerised adaptive learning environments may be used to identify patterns of errors at an individual level.

1.8 Adaptive learning environments

1.8.1 *Two sigma problem*

What is unique about adaptive learning is that the focus is on the use of technology to personalise learning. The technology is based on expertise in various fields of study such as computer science, psychology, education and neuroscience. Adaptive learning is a branch of the wider area called personalised learning. The computer adapts the educational material according to the student's learning needs, by monitoring their responses. A computerised system creates the chance to optimise learning gains at a large scale, in a flexible and dynamic way, based on the generated high-frequency response data. Adaptive learning is therefore sometimes boldly called a way to solve the 2 sigma problem: the educational phenomenon that learners who are tutored one-on-one perform two full standard deviations higher than learners taught via conventional instructional methods (Bloom, 1984), see **Fig. 1.2**. The thought was that "if a human tutor can improve the performance so drastically, then an automated system might be able to also deliver some (but not all) benefits of tutoring" (Savi et al., 2017). Numerous e-learning applications have tried to deliver this promise. Some of the more well-known intelligent tutoring systems that adapt materials to mathematical abilities are ALEKS (Canfield, 2001; Yilmaz, 2017), Knewton (Wilson & Nichols, 2015), i-Ready (Curriculum Associates, 2020) and Learn (Straatemeier, 2014). These systems differ in which algorithm they use to predict the knowledge state of the learner and adapt the learning environment accordingly. For instance, ALEKS is built on the psychometric engine of Knowledge Space Theory; Knewton on Item Response Theory; and i-Ready on Computer Adaptive Testing. Although the algorithms are different, they rely on knowing the difficulty of the problems beforehand. Learn makes use of an on-the-fly Elo-algorithm, which makes the process of pretesting not necessary. For more details on this algorithm, see section 2.1.5.

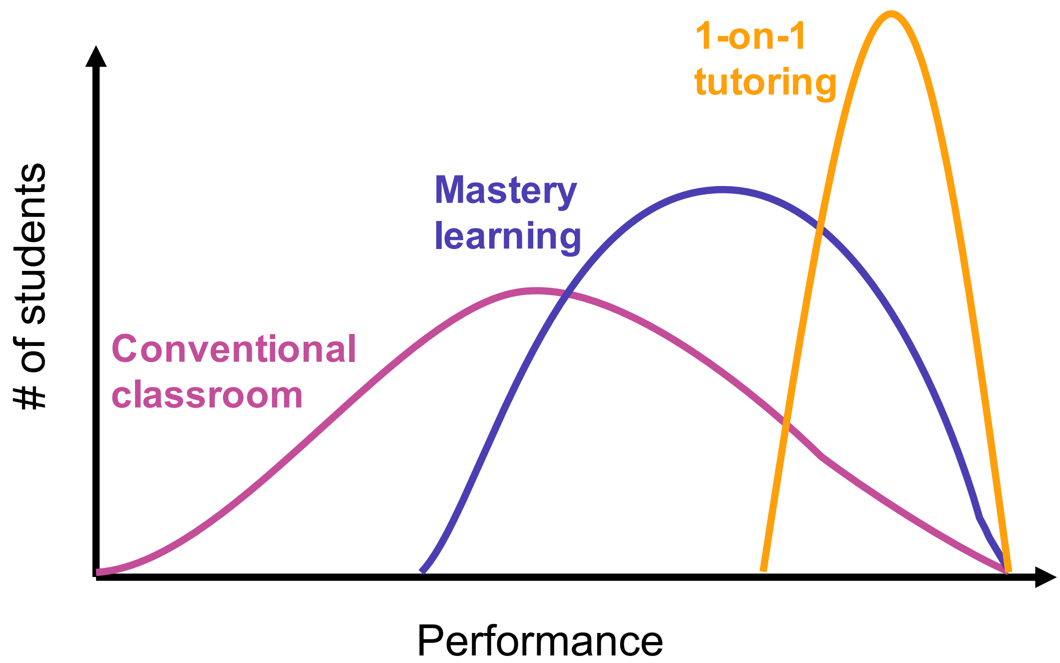


Figure 1.2*. The performance distribution for students under conventional, master learning, and one-on-one tutorial instruction.*

1.8.2 *Idiographic approach to development*

E-learning applications allow studying high quality and highly frequent measurements of maths abilities. Rich intraindividual time series are needed to study individual development and to adapt the item difficulty optimally. The importance of frequent sampling to collect longitudinal data was stressed by Siegler (1991). Indeed, Siegler's (1991) microgenetic method shares many ideas with the idiographic approach (Molenaar, 2004), although the latter has more focus on intra-individual data. Both of these approaches suggest that (1) observations should span the whole period of the developmental process; (2) observations should be collected at a high density; (3) and analysis should focus on inferring the process of quantitative and qualitative changes in development. Focussing on developmental changes is important because developmental processes can be quite sudden (Case, 2013) and are easily missed when there are not enough observations.

I contributed to the writing of the publication by Hofman et al. (2018), which reports intra-individual analyses of the development of mathematical skills in Math Garden. The data consisted of a large set of individual participants' time series of responses to single addition and multiplication items, see **Fig. 1.3** for three examples. When we look at these time series it is clear that the trajectories of mastery differ a lot, where some problems are mastered as expected, with a clear increase in the probability of correct response (**Fig. 1.3A**). Other problems are already mastered (**Fig. 1.3B**) or are not learned at all (**Fig. 1.3C**). For some problems, the extent of learning remains unclear since a student consistently switches between giving a correct and incorrect response.

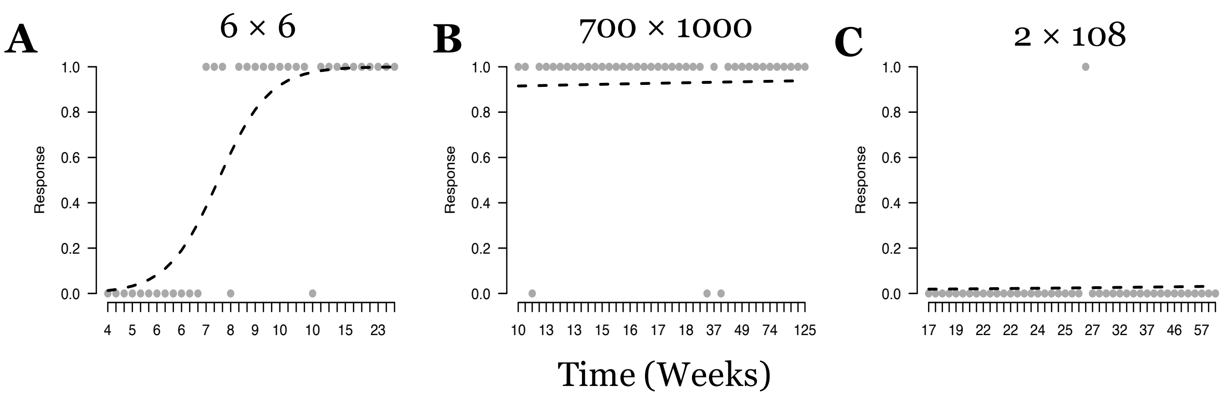


Figure 1.3. *Three examples of response patterns over time of the same student for different multiplication problems.*

In a first study of Hofman et al. (2018), learning patterns were investigated by fitting learning curves with logistic regression models. We found that these patterns were irregular, with both relapses to lower ability and sudden changes to improved ability. Only a minority of the learning curves showed a positive learning curve, so an increase in mastery. The variation in learning curves can be explained by the adaptive nature of the system since the child is always presented with problems tailored to their ability. The adaptivity removes the tails of the learning curve, i.e., the part when the item is still too difficult or becomes too easy. But it is also due to the variation in strategy use that children exhibit, where they sometimes simply forget the learned strategy or switch to less optimal strategies (Lemaire, 2010; Siegler & Crowley, 1991).

In a second study, we looked at the dimensionality or clustering of the mathematical problems. In other words, is there a general skill to master (i.e., multiplication skill) that enables a child to solve all the multiplication problems or are these skills more local to a particular problem (e.g., 3 x 4?) What the results showed was that for instance multiplication table problems and x 100 problems are not clustered and represent two different skills. In general, mathematical skills can be global or local, and learning one local maths skill is not necessarily linked to learning another local skill.

In sum, these studies demonstrated considerable differences within and between children with regards to improvement in the resolution of specific problems. This exemplifies the multidimensionality of arithmetic development. For example, similar items such as 4 × 6 and 12 × 3 are expected to be unidimensional (same strategy use, underlying the global skill of multiplication) but this not the case. The results of these two studies are a useful reminder that learning is a challenging complex system to capture and that individual differences are considerable.

1.9 Thesis overview

In this project, I will use the findings of the previously mentioned literature to investigate how individual differences in cognition (e.g., strategy choice, executive function skills, adaptivity) predict children's performance in an online learning environment and how we can tailor individuals' learning environment based on these differences. To personalise learning in an e-learning application, we also need to infer the cognitive process involved with online measures above and beyond response data. Therefore, I have used online measures such as eye and mouse tracking to investigate how online measure map onto strategy use, task-specific errors and level of executive functioning. These methods will be discussed and compared in **Chapter 2.** In this chapter, I also provide details on the specific online learning environment that was used in the experiments.

This thesis described three studies. The first study, discussed in **Chapter 3**, was a lab study testing primary-school children aged between 8 and 11 years, conducted at the Centre for Brain and Cognitive Development in London. I discuss our finding that individual differences in cognitive abilities (i.e., working memory and inhibitory control) impact arithmetic performance and the number of eye fixations. I show that the associations depended on whether time pressure was visible, a key aspect in most gamified educational settings. In this study, I also explored how eye and mouse movements co-occur during arithmetic problem-solving. These findings were used to develop strategies for the next study that was conducted in the online learning environment of Learn. This second study aimed to get a better understanding of children's uncertainties and strategy use when facing mathematical problems even in the absence of making errors. To examine this, I collected data over 6 weeks from 1,500 children (aged between 5 and 13 years) in Learn on an arithmetic task similar to the task used in the first study. We found that mouse movements indeed related to children's difficulties underlying false associations (**Chapter 4**).

In the third study, I collected response data in Learn from 150,000 children aged 5 to 13 years old over the course of six months. The aim was to investigate whether the behavioural metric for the ability to monitor the error and behaviour adjustment, also called post-error slowing, can be used in a learning environment to predict performance. In **Chapter 5**, I show evidence of post-error slowing, the finding that humans slow down their performance after an error. This was in 21 different learning activities related to mathematical and language skills. Here, I also describe factors influencing the presence or magnitude of post-error slowing.

Finally, the main findings will be summarised in **Chapter 6** and discussed. I will discuss three insights that can be drawn, related to the role of executive function in online learning; how eye and mouse tracking can inform us on the underlying cognitive process; and how post-error slowing can be used as a measure of the zone of proximal development. Finally, limitations related to the used measures for learning, and more specifically the framework of an online learning environment, as well as future directions are addressed.

2 METHODS

2.1 *Learn*, an online learning platform

Learn is an online adaptive learning environment for practising mathematics, (Dutch) language skills and learning English as a second language (www.oefenweb.com). The platform started as a research project called Math Garden at the University of Amsterdam in 2007, in a collaboration between the departments of developmental psychology and psychometrics. The initial idea of this project was to study the mathematical development of children with high-frequency measurements by testing children daily in an educational context. This led to the development of an instrument that would meet both educational and scientific aims: Math Garden would be used to study the development of children and serve as an educational tool for children to play at their level. The platform quickly became so popular with schools that it was commercialised under the name of Oefenweb. Rapidly, other adaptive practice systems were added, such as practising the Dutch language (Language Sea) and learning the English language (Words & birds). Now, the platforms are owned by a bigger company, called Prowise, and is known as the *Learn* environment. The Learn platform has recently been made available in English and German (https://www.prowise.com/en/learn/). The data described in Chapter 4 and 5 were collected in the Learn environment.

2.1.1 *Users of the Learn environment*

The participants of the online learning environment are mainly pupils from primary schools that have bought accounts for their students. In this online learning environment, students can play at home and/or at school; they are also free to choose which games and how many problems of a particular game they want to play. There are currently, as of December 2020, 200,412 active users (active meaning 'logged in as a user in the past three months') playing in *Learn* from across 5,000 schools in the Netherlands. Children can start to play from the age of 4 years and continue throughout their primary school, which is in the Netherlands until the age of 12 years. **Fig. 2.1** shows a distribution of the age of pupils that play in *Learn*, with an average age of 9.2 years. Since the data in *Learn* is privacy sensitive, most data are anonymised. The only information known about the users is their birthdate, gender and which school they go to. Children (their parents or schools) can opt out of being part of the research done in the learning environment, in which case they were not included in any of the studies described in this thesis.

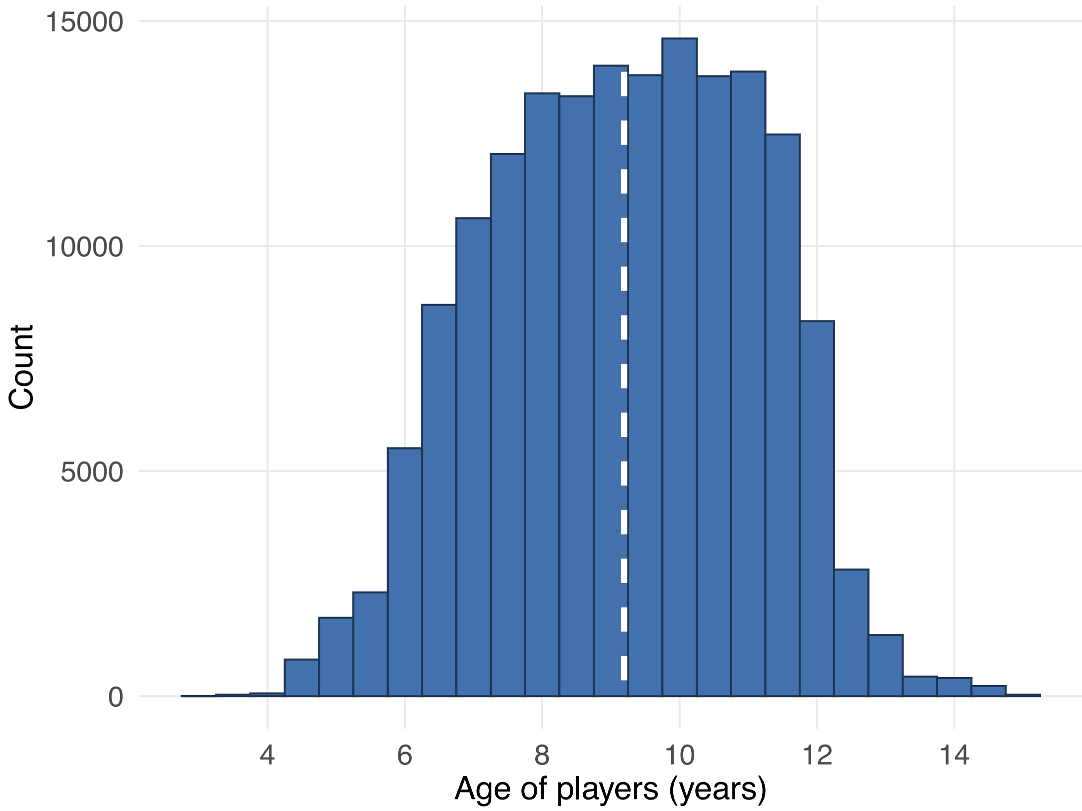


***Figure 2.1**. Distribution of the age of children that actively play in Learn. The dashed line is the average age, M = 9.2 years.*

2.1.2 *Engaging and rewarding activities*

The learning system is designed so that children practice their skills at school and home voluntarily, with playful and adaptive games. The engagement is managed by a few game principles. First, the homepage of *Learn* shows the student's personal play environment. Depending on which platform, this is either a garden for practising mathematics ("Math Garden"), a sea for language ("Language Sea") or a forest with birds for learning English ("Words & Birds"). For example, in Math Garden, each plant represents a game from a different domain, such as addition, multiplication or percentages (**Fig. 2.2**). The state of the play environment represents the persistence of student's practice: When they practice a lot and their ability increases, their plants will flourish, but when a child is not practising enough, their plants will wither. The idea is to motivate children to visit the games regularly. Second, the problem difficulty matches the student's proficiency, a key feature of the adaptivity of the online learning environment, see section 2.1.5. Third, children receive direct feedback after every response and correct performance is rewarded with coins. These coins can be collected to buy virtual prizes such as trophies. Since the mean accuracy rate is controlled by adapting the difficulty of items, trophies and coins are independent of the child's overall ability and only rely on the frequency of playing and their chosen difficulty level (section 2.1.6).

Figure 2.2. *The homepage of Math Garden. Each plant represents a game, such as subtraction or multiplication, and the score underneath the plant indicates the ability of the student (on a scale of 1 to 1,000) in that domain. The student can select the difficulty level of the games by clicking on the right figure, with either one (easy level), two (medium level), or three (hard level) sweat drops.*

2.1.3 *Math Garden and Language Sea games*

During my PhD, I've collected data from various games that are played regularly by children in Math Garden (Chapter 3-5) but also Language Sea (Chapter 5). In the Appendix section A.1 is a description of all the games that were used in the experiments.

Language Sea offers more than 35,000 exercises in a total of 18 games to practise Dutch language skills such as vocabulary, spelling (**Fig. 2.3**), reading, proverbs and dictation. Language Sea is designed for children starting in Year 1 until the end of primary school. Math Garden has a total of 34 games with more than 25,000 exercises, intended for pre-schoolers and primary school children. Children can practice in Math Garden their basic arithmetic, but also the multiplication tables, telling time, counting money and fractions, etc.

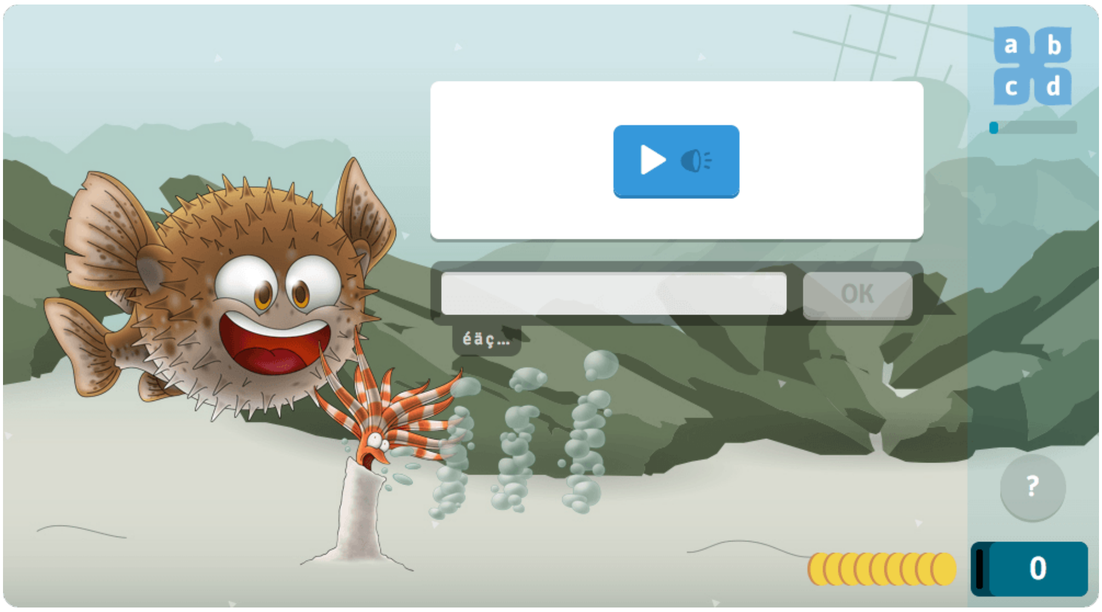


Figure 2.3. *The layout of the game "Dictee" is available in Language Sea for practising children's spelling. The student hears a word and is asked to write down the word in the answer box.*

2.1.4 *Game procedure*

In each game session, a child is asked to solve a problem and either has answer options to choose from or is asked to type in the answer (**Fig. 2.4**). There is a question mark that children can use when they do not know the answer and a stop sign to end the game session. All games are done under time pressure. The remaining time to answer is visualised as virtual coins counting down in the bottom right corner of the screen. The student is rewarded with the coins that remain on the screen after giving a correct response; an incorrect response leads to the remaining coins being subtracted from their total points. No coins are gained or lost when failing to answer within the time limit. The rationale of this scoring rule is explained in section 2.1.5.

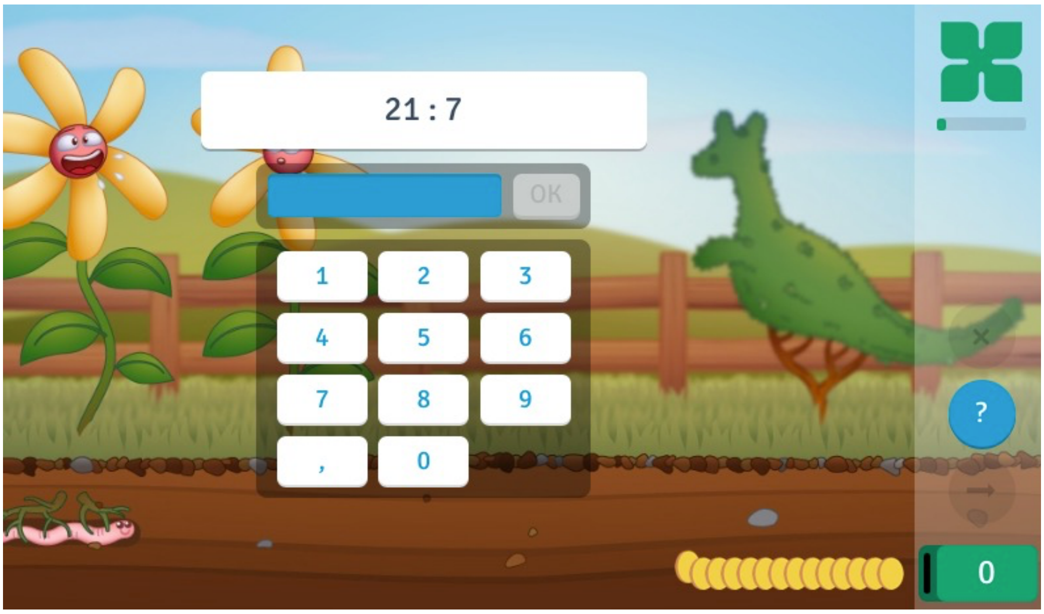


Figure 2.4. *An example of a mixed operations trial in Math Garden. Players have to use their keyboard or click on the numbers in this keypad to submit their answer. If they are unsure of what to answer, they can use the blue question mark button on the right. As a result, the right answer will be given, and no coins are lost or won. The coins at the bottom of the screen represent the remaining time to respond. After a correct response, the remaining coins are rewarded and subtracted after an incorrect response.*

2.1.5 *Computer Adaptive Practice*

In general, adaptive systems use a form of Computer Adaptive Testing (CAT), derived from item response theory (IRT). CAT is based on specifying the probability of a correct answer given the characteristics of the items and the person (e.g., van der Linden & Glas, 2000). The advantage of using CAT is that a person does not have to solve all items to determine their level, but that a more/less difficult item is chosen based on the person's responses to previous items. This way, the test can be tailored to their ability, sparing them from having to solve uninformative (too easy, or too hard) items. In Learn, the CAT technique is extended to allow computer adaptive practise by (1) using both speed and accuracy to score the responses and (2) updating children's scores and item difficulties in real-time using an Elo-algorithm.

Integrating response time in this algorithm ensures additional information is used to estimate children's ability. Especially for easy items, a correct answer is not informative, but the speed of responding is (van der Maas & Wagenmakers, 2005). Therefore, a scoring rule called the "High-Speed High Stakes" rule was developed where accuracy and speed are weighted, following the equation $S = (2X_{pi} - 1)(d - T_{pi})$ (adapted from Maris & van der Maas, 2012). This rule imposes a speed-accuracy trade-off, where fast and correct responses result in a high score and incorrect responses result in a negative score to prevent fast guessing. X_{pi} is 1 for a correct response and 0 for an incorrect response for player p on trial i. T_{pi} is the response time in seconds before the imposed time limit d; and eventually a score S is calculated. To give an example: if the time limit of the game was set to 15 s and a correct response was given in 8 s, this scoring rule would result in a score of +7, whereas an error after 4 s would yield a score of -9.

Students' abilities and item difficulties are estimated with an on-the-fly Elo estimation algorithm based on the accuracy and speed of the students (Klinkenberg et al., 2011). This Elo algorithm is a self-organising system that updates the item difficulties and children's abilities continuously. The advantage of this algorithm is that items do not have to be pre-tested to determine their difficulty, as is normal in CAT, which is very time-consuming when thousands of items are used. What is unique about the algorithm is that people play against items: a person's ability rating increases when they solve the problem correctly and fast but decreases when the answer is incorrect or very slow. The same goes for the item difficulty ratings: items that tend to be answered fast and correct are scored as easier items, whereas items that are answered incorrect or slow are perceived as more difficult. For a more detailed mathematical explanation of this adaptation, see Maris & van der Maas (2012).

2.1.6 *Difficulty Level*

Learn uses children's ability estimates to present them with problems that will lead to a target correct response rate. The success rate is set high, by default at .75, to ensure students remain motivated (Jansen et al., 2013; Straatemeier, 2014). High success rates cause children to feel more competent and less anxious, which enhances the frequency of playing. In turn, practising on a challenging, but comfortable level enhances the performance (Jansen et al., 2013). Players can choose to modify the success rate, by setting it to a probability of .90 (easy level), .75 (medium level), or .60 (hard level) of answering correctly. This set difficulty level can be chosen by the pupil and changed when preferred. To motivate students to choose a more difficult level, students only get half of the coins in the easy compared to the medium level. In the hard level, students receive twice the number of coins with a correct response.

2.1.7 *Research*

Due to the popularity of *Learn*, the platform has served as a rich source of information for research. Numerous experimental and methodological papers, as well as five dissertations, are the result of that. Roughly three research approaches have been used. First, many methodological papers have focused on validating the methods used in creating the adaptive platform. The topics range from looking at understanding differences in item and person parameters and/or compare these with mathematical models (Gierasimczuk et al., 2013; Hofman et al., 2020; Jansen et al., 2014; van der Ven et al., 2015). For example, it was shown that Math Garden performance measures correlate highly with traditional pen-and-paper arithmetic tests (Klinkenberg et al., 2011). The second line of research focuses on understanding the cognitive strategies used by children in Learn. Studies have, for example, looked at fast and slow arithmetic strategies (Hofman, Visser, et al., 2018) or misconceptions (Buwalda et al., 2016; Savi et al., 2018) in multiplication. The study of van der Ven et al. (2013) showed an association between the level of visuospatial working memory, measured in one of the games in Learn, and arithmetic skills, and found the strongest associations with solving addition and subtraction problems. The third category of research investigates developmental processes with longitudinal studies, among others the phenomenon mutualism. Mutualism theory posits that learning one process (skill) supports learning other skills, such that there is a positive manifold with beneficial relations between different abilities during development (van der Maas et al., 2006). Evidence of mutualism was found in the arithmetic abilities practised in Math Garden, such that learning to count and add positively influenced each other as well as multiplication and division skills (Hofman, Kievit, et al., 2018; Ou et al., 2019).

2.2 Eye tracking

In the experiment described in Chapter 3, human eye movements were recorded with an eye tracker. In the context of learning, eye movements offer a direct and high temporal precision measure of the learners' attention (for reviews, see: M. Lai et al., 2013; Orquin & Loose, 2013). It is an often-used method to get a window into the ongoing dynamics of cognitive processing (Holmqvist & Nyström, 2011; Just & Carpenter, 1976). Studies typically analyse eye movements in terms of fixations (i.e., moments when the eyes are relatively stationary) or saccades (i.e., rapid movements between fixations). The aim is to reveal the extent to which attention is directed towards an object, and supposedly the extent to which this object has been processed.

In an educational context, the implementation of measuring eye movements started primarily in reading studies (for a review see Rayner, 1997), and was later applied to study various cognitive processes. For example, eye tracking can reveal information on individual differences in strategy use in mathematical learning (e.g., Bolden et al., 2015; Gomez et al., 2017; van 't Noordende et al., 2016). These studies investigated strategies with a number line estimation task. In this task, children are asked to estimate the position of numbers on a number line. The number of eye fixations on reference points children made to estimate the numbers on the line was found to reflect individual differences in how mature their strategies were. Eye tracking data can also be used to study how learners split their attention during multi-media learning (Hyönä, 2010; Schmidt-weigand et al., 2010; van Gog & Scheiter, 2010). By tracking eye movements, we can investigate how a student interacts with multiple stimuli and how the order and duration of their attention towards these stimuli affects learning (Guerra-Carrillo & Bunge, 2018; Stojić et al., 2020). Some e-learning applications, have tracked eye movements in an attempt to predict the behavioural state of the user in terms of the level of interest and attention (e.g., Asteriadis et al., 2009; Barrios et al., 2004). However, there is not enough research to suggest that this implementation has been successful and was mostly studied in small sample sizes.

Most studies using eye tracking are conducted in the lab, but it is becoming popular to also use eye tracking in more naturalistic settings. Eye tracking devices can be categorised either as a static eye tracker, e.g., *Tobii TX300* and *Eye link 1000*, which are mainly used in the lab, or head-mounted eye tracker for research, such as the *Ergoneers Dikablis* or *Tobii Pro Glasses* (for an overview, see: Cognolato et al., 2018). With head-mounted eye trackers, it is not required for the participant to sit still, making it more naturalistic. Eye-tracking devices can record eye movements as well as gaze allocation at sampling rates in a range of 30 to over 1,000 frames per second (Holmqvist & Nyström, 2011). Most eye tracker manufacturers provide their own software packages to present stimuli and track eye movements (e.g., Tobii studio). However, researchers can also easily program the presentation and analyses of their experiments such as in MATLAB with Psychophysics Toolbox (PTB; Brainard, 1997) or Python with PsychoPy (Peirce, 2007). In my experiment, I used the static eye tracker Tobii TX300 in the lab and programmed the tasks and data collection in Psychophysics.

Due to the costs of eye trackers, it remains difficult to use eye trackers on a large scale, in a school for example. Therefore, some researchers have looked at webcams to track gaze patterns (Valenti et al., 2009). Especially during COVID-19, people have become more interested in doing online research and webcam-based eye tracking would be an ideal research tool. Until recently, it was difficult to combine eye tracking and online research because there were no libraries in JavaScript, the most commonly used language in online research. But now there are algorithms available, such as *WebGazer* (Papoutsaki & Laskey, 2016). What studies have found so far is that webcam-based eye tracking is an option when detailed spatial or temporal information on the fixations is not required (Semmelmann & Weigelt, 2018). As a minimum, webcam-based eye tracking allows categorising if someone is looking to the left or the right side of the screen. As the foundation of webcam-based studies grows, the accuracy of webcam-based eye tracking is likely to improve and can become an exciting method for online research and potentially online learning environments.

2.3 Mouse tracking

2.3.1 *The benefits of using mouse tracking*

In the experiments explained in Chapter 3 and 4, mouse movements were recorded. Mouse tracking is the recording and tracking of computer mouse movements made by participants. This method aims to provide a continuous stream of information during the decision-making process (Dale et al., 2007; Freeman, 2018; Freeman et al., 2011; Hehman et al., 2015; Song & Nakayama, 2009; Spivey & Dale, 2006; Stillman et al., 2018). Mouse tracking, as a research method, was first introduced by Spivey et al. (2005), who used it as a window into the internal cognitive process during language comprehension. Since the free-to-use standalone *Mousetracker* software was made available by Freeman and Ambady (2010), the software has become very popular in diverse domains of social sciences (for recent reviews, see Erb, 2018; Freeman, 2018; Stillman et al., 2018). More recent toolboxes to facilitate mouse tracking are: (1) as a feature in Gorilla, an online experimental testing platform (Anwyl-irvine et al., 2020), and (2) a MATLAB Toolbox called MatMouse (Krassanakis & Kesidis, 2020).

Studies using mouse tracking are mainly interested in whether participants are deviating their mouse trajectory towards alternative options before arriving at the correct response. The deviation towards alternative options by the motor action of a mouse movement can reveal how multiple representations are competing with each other during the problem-solving process (Schulte-Mecklenbeck et al., 2019; Spivey, 2007; Spivey & Dale, 2006). In other words, mouse movements are thought to serve as a proxy of a person's cognitive process. It is an attractive tool because mouse tracking has many practical advantages: it can be collected online, so it does not require a lab setting; it can be used in combination with other methods such as eye tracking and EEG; it is relatively inexpensive and can be conducted and analysed at a large scale.

2.3.2 *Design and analyses of previous mouse tracking studies*

Most previous mouse tracking studies have used a strict two-response options design. In this design, there is a correct response button in one corner of the screen and a distracting response button in the opposite corner. The participants always start with their mouse at the start button in the middle and move their mouse to select a response button. This mouse trajectory is recorded. The standard measures to analyse mouse movements are calculating the maximum deviation (MDev) away from the correct response and the Area under the Curve (AUC) between a straight trajectory towards the correct response and the deviated trajectory (Stillman et al., 2018, **Fig. 2.5 A)**. Other measures focus more on the temporal information, i.e., the nature of movements at different time points (Hehman et al., 2015). For example, the velocity and acceleration of the mouse between time points or the angle of movement (θ) between time points (**Fig. 2.5 B**). These temporal measures are used to study at which point the participant was influenced by the two response boxes in the screen to move their mouse. Some studies have also used designs with four-response choices (see for example Cloutier et al., 2014; Koop & Johnson, 2013), but in the analyses, only the trajectories where there was deviation towards a single one of the alternative options, and not multiple options, were selected, similar to the analyses of a two response design.

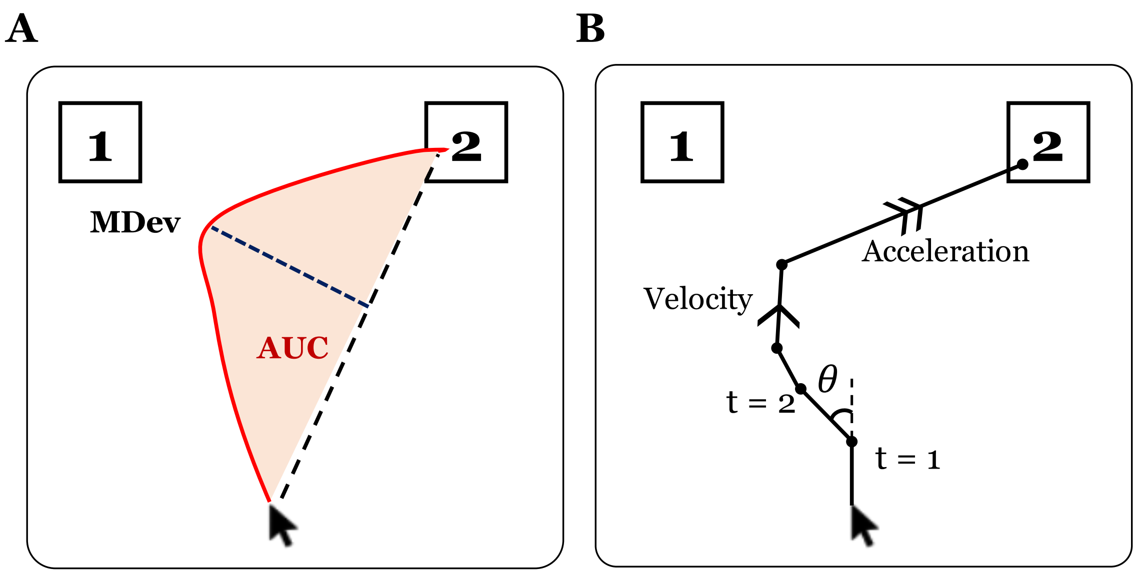


Figure 2.5. Example of a standard mouse tracking experiment design with two answer options. The participant starts the trajectory in the middle at the bottom of the screen and the mouse movements are recorded until one of the answer options (1 or 2) is clicked. Standard descriptive measures of the movement are **(A)** Maximum Deviation (MDev) between the optimal (black dashed line) and real mouse (red line) trajectory as well as the Area under the

Curve (AUC) between the two lines; **(B)** The angle θ, velocity and acceleration between two time points of the mouse trajectory.

2.4 Comparison of mouse and eye tracking

2.4.1 *Literature review*

Previous studies have correlated eye and mouse movements in experiments (Demšar & Çöltekin, 2017; Franco-Watkins & Johnson, 2011; Spivey, 2007), with a large body of research in web searching (Guo & Agichtein, 2010; J. Huang & White, 2012; Milisavljevic et al., 2018). The main reason for this comparison is to determine whether mouse tracking could be a useful alternative to eye tracking as a measure of people's attention (Cooke, 2006). Since mouse tracking can be applied at a large scale, is cheap and does not require much effort it would be an attractive method to deploy on any website or experimental platform if it proves to be useful. Some studies use mouse tracking as an additional tool to eye tracking, for example when there is poor gaze accuracy such as with webcam-based eye tracking (Franco-Watkins & Johnson, 2011; Papoutsaki & Laskey, 2016; Saboundji, 2020). The combination of the methods is also suitable for biometric authentication (Kasprowski & Harezlak, 2018; Rose & Awad, 2017) or to measure a person's stress levels (J. Wang et al., 2017). Connecting the two types of movement is not easy, because they are quite different (Dodge et al., 2009). Eyes move with high speed, in discrete jumps (saccades) between longer stationary stops (fixations), and are ballistic, whereas mouse movements are usually smooth and continuous, and can be changed mid-movement. Another related difference is that mouse movements are thought to be under conscious control, whereas eye movements are in most cases considered to be controlled unconsciously - although, Henderson & Smith (2009) showed mixed eye movement control during scene viewing. However, the two types of movement also have similarities, in the sense that they can both be represented as time series of locations on the screen and are thought to underly the same cognitive process, namely solving the given task.

Some studies have claimed that mouse trajectories can serve as a proxy of gaze trajectories (Demšar & Çöltekin, 2017; J. Huang & White, 2012) and that mouse movements are representative of eye movements (Cox & Silva, 2006; Freeman & Ambady, 2010). A general finding is that the gaze always precedes the hand/mouse movement unless the experimenter explicitly asks otherwise (Demšar & Çöltekin, 2017). This means there is usually a lag between the two time series. Other studies focused on the eye-mouse distance to see whether gaze and mouse aligned, and found that it varied from 0.1 to 1000 pixels depending on the task and the strategy a person holds (Holmqvist & Nyström, 2011). During web browsing for example, Rodden and Fu (2008) found an average distance of 257 pixels between gaze and cursor. The eye-mouse distance can vary a lot depending on the strategy, for example: (1) when the person is reading while keeping the mouse still; (2) when the person is using the mouse as a reading aid (either vertical or horizontal) to guide the eye; (3) or when the person is using the mouse to mark areas that might be interesting to return to, possibly to reduce working memory load. Similar strategies were found in other studies as well (Buscher et al., 2012; Cox & Silva, 2006; J. Huang & White, 2012).

Due to the large variation in the physical eye-mouse distance findings, studies have looked at dwell times in Areas of Interest (AOIs). An AOI is a pre-specified area the person visits with their mouse or eyes. During web browsing, studies have found around 70% chance that the eyes and mouse go to the same AOI on the screen (Chen et al., 2001; Cooke, 2006). Evidently, the height of this overlap depends on the number of AOIs that are specified, with more chance of overlap when defining big and few AOIs. Most of these studies were performed during web searching, but not with cognitive tasks.

In the next section, I will show some exploratory comparisons between eye and mouse movements of 39 primary school-aged children performing two different tasks in a lab-based setting, which is explained in more detail in Chapter 3. Mouse movements were tracked using MATLAB, collecting the cursor's x and y coordinates with a timestamp around 100 times per second. The collection of eye movements is explained in more detail in section 3.2.6.

2.4.2 *Eye and mouse movement in Simon task*

The first task for examination is a Simon task (see section 3.2.5 and **Fig. 3.1.B**), where children were asked to move their mouse as quickly as possible to one of the answer options, either the top left or top right. This set-up is comparable to a standard mouse-study design. In congruent trials, the location of the target was congruent with the correct response and in the incongruent trials, the location of the target was on the opposite side.

In a first analysis, the proportion of eye fixations and mouse points in the two answer option boxes were calculated across the trials per participant. Only the correct trials were used, which was in this case 99% of the trials. We expected that the proportion of eye fixations and mouse deviations towards the incorrect response would differ between the incongruent and congruent trials. **Fig. 2.6** shows the gaze and mouse trajectories in the congruent and incongruent trials on the left, and the proportion of eye fixations and mouse points on the right. As expected, there were more eye fixations in the incorrect answer box for the incongruent trials than for the congruent trials, t (34) = -7.44, p < 0.001, on average a M=-0.15 proportion difference. The proportion of mouse points towards the incorrect answer box also differed significantly between the congruent and incongruent trials, t (38) = -3.69, p = 0.001, but the difference in proportion was smaller M = -0.049.

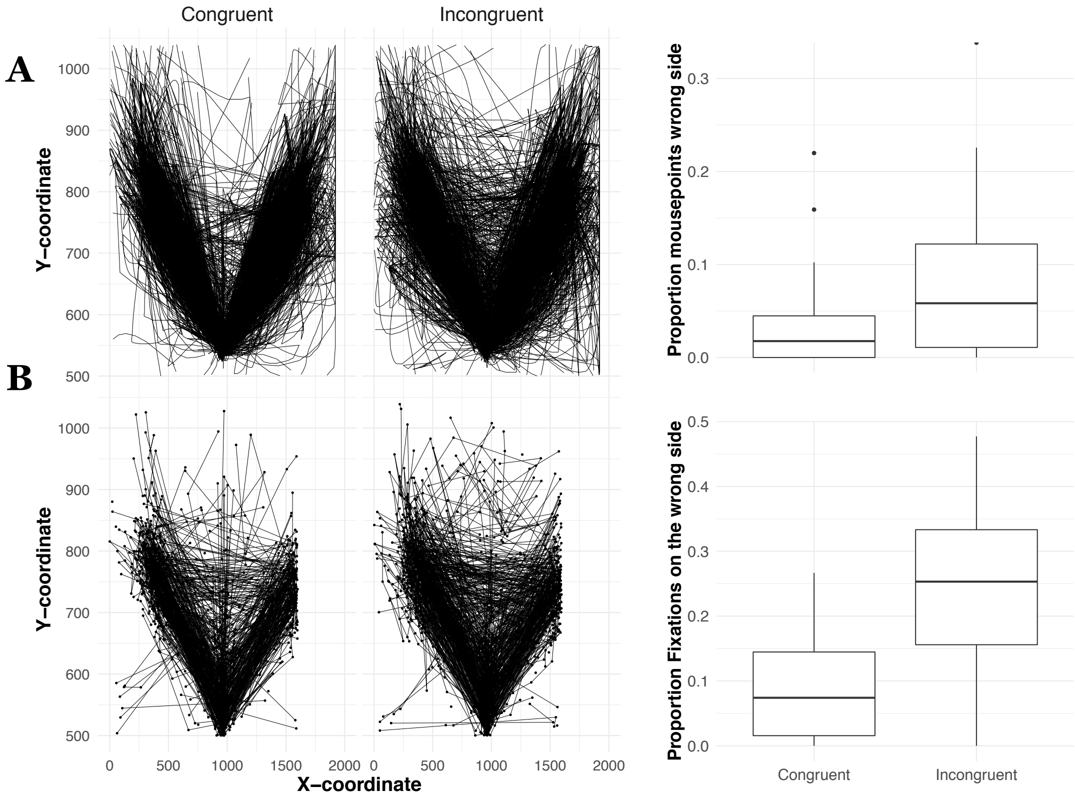


Figure 2.6. (A) *Mouse trajectories on the left and proportion mouse points in the incorrect answer box on the right for the congruent and incongruent trials.* ***(B)*** *Gaze trajectories on the left and proportion eye fixations in the incorrect answer box on the right for the congruent and incongruent trials.*

Second, the purpose of administering the Simon task was to have an individual measure of the interference effect for children, for which the difference in RT between congruent and incongruent trials was calculated (see section 3.2.5). To examine how predictive eye and mouse trajectories were for this RT interference effect, correlation tests were performed. The difference in eye fixations between the congruency of the trials significantly correlated with the RT difference, r = .61, t(30) = 4.17, p < 0.001. The difference in mouse points between congruent and incongruent trials also correlated highly with the RT difference, r = .65, t (37)

= 5.23, $p < 0.001$. The correlation between the eye and mouse congruency differences, was smaller, $r = .47$, but still significant, $t(30) = 2.91$, $p < 0.007$.

In conclusion, both the eye and mouse trajectories showed more deviation towards the incorrect answer option in incongruent trials than in congruent trials. Both metrics can be used to get a measure of the interference control effect, compared to the commonly used RT difference. The low proportion of mouse points towards the incorrect response could result from the fact that children do not always move their mouse completely inside the answer box but only towards.

2.4.3 *Eye and mouse movement in the arithmetic task*

The second task is much more complicated in design compared to the Simon Task with five incorrect answer options. The task is a multiple-choice arithmetic task, similar to the *speedmix* game in Learn, explained in detail in section 3.2.3, **Fig. 3.1.A**. Children had to click their mouse in the middle of the screen to start a trial in which an arithmetic problem appeared. They had to perform under time pressure, with only eight seconds to respond. The children had to choose between six answer options, appearing in a circle underneath the question. One of these answer options was correct, and the other ones reflected commonly made errors for the problems in Math Garden. For the analyses, we only looked at the trials that were answered correctly. In these trials, we expected children to make mouse movements and eye fixations to frequently made errors, apart from the correct answer.

As a first step, individual trajectories of both eye and mouse movement were explored, with three examples shown in **Fig. 2.7** of three different children that answer the same question. These trajectories gave some insight into the cognitive process and strategies. For example, in **Fig. 2.7A**, the participant seems to mentally calculate/memorise the answer by looking at the question and go straight towards the correct answer with the eyes and mouse. In **Fig. 2.7B**, the participant looks at some of the answer options such as 54 and 64, but only moves the mouse towards 64, the most commonly made error in Math Garden for 9 x 7/ 7 x9. Eventually, the participant moves their mouse and gaze towards the correct answer. In **Fig. 2.7C**, the participant looks and moves their mouse in an anti-clockwise movement towards all answer options and eventually ends at the correct response. Similar to the Simon task, this last example shows that the mouse movements are moving towards an answer option, but do not always end completely inside the answer option. Therefore, for the next analyses, the Areas of Interest, the answer options and the question box, were defined differently for the mouse than for the eye, shown in **Fig. 2.8.**

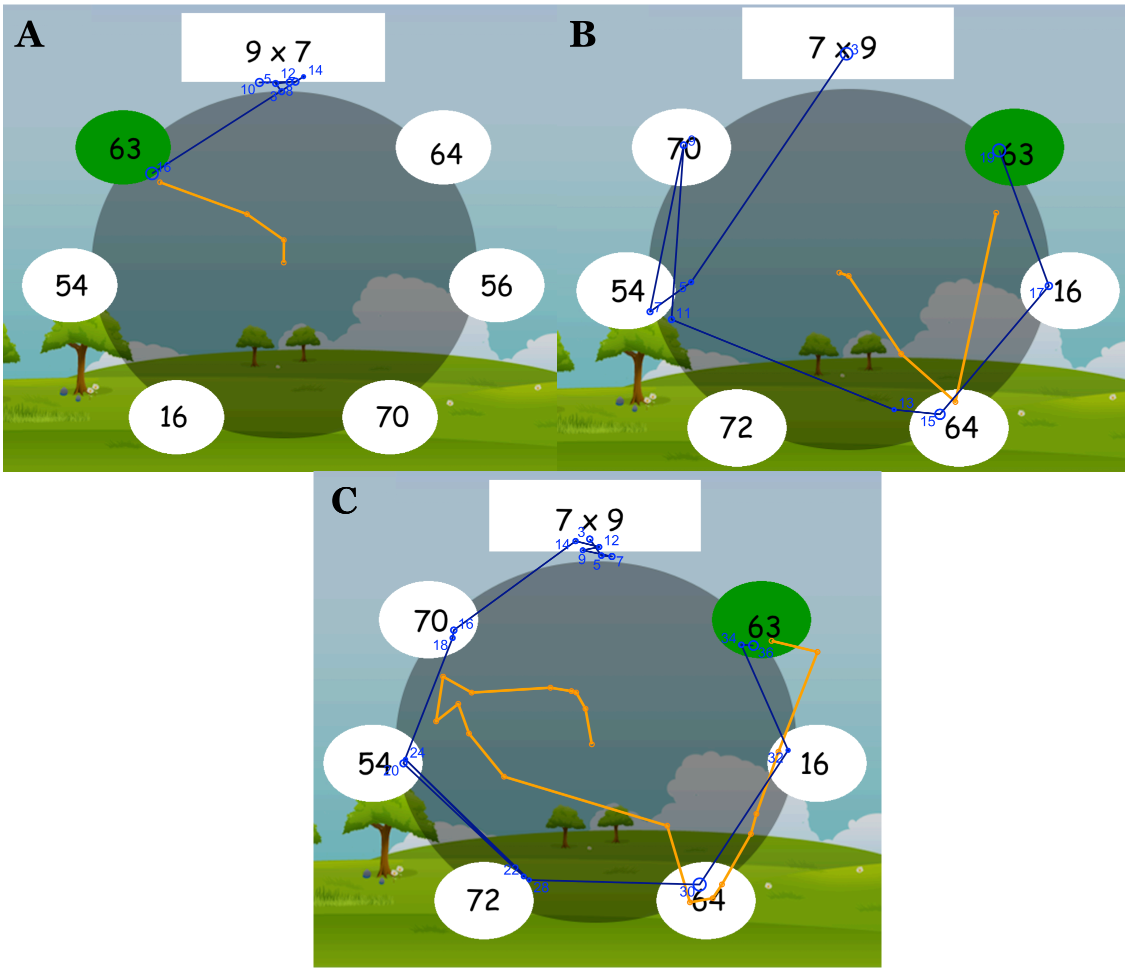


Figure 2.7. *Three examples of children's mouse (yellow line) and gaze trajectories (blue line) when solving the problem 9 x 7/7 x 9. The numbers next to the eye fixations reflect the order of the fixations, and the size of the point indicates the duration of the fixation.*

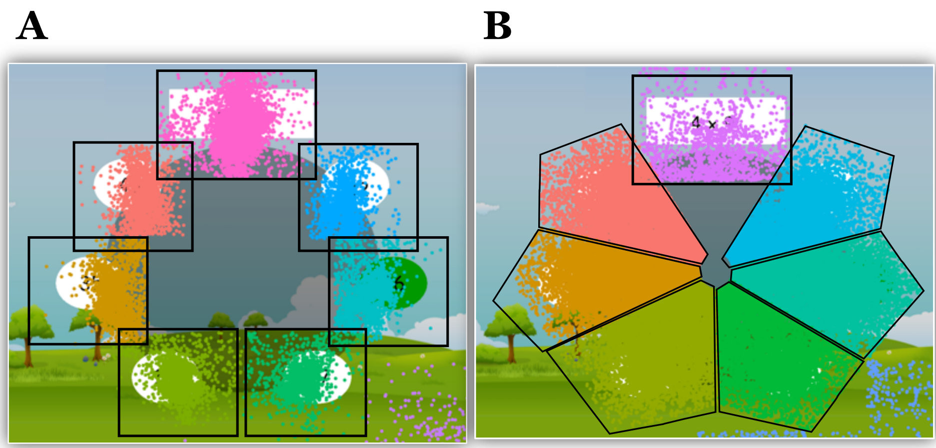


Figure 2.8. (A) *Eye fixations and* ***(B)*** *mouse points, labelled when they were inside the AOIs (black lines). For the mouse points, the AOIs were defined differently, since most mouse trajectories do not reach the answer option completely.*

First, all eye fixations and mouse points were categorised as belonging to one of the AOIs. Depending on the problem, these AOIs were labelled as either the correct answer option, question or an error, labelled E1 to E5 where E1 is the most commonly made error. **Fig. 2.9 A** shows the percentages of eye fixations and **Fig. 2.9 B** the mouse points in the AOIs. A big difference between the two metrics becomes obvious due to the percentage difference for the question box: participants look at the question but do not necessarily move their mouse towards the question. Both metrics have comparable movement towards the commonly made errors. There is not much difference in the percentage of mouse movement from E1 to E5, where you would expect less movement towards the least made error E5 than towards E1.

Finally, the average eye mouse distance was explored, through calculating the Euclidean distance between the x and y coordinates for each time point: Distance = $\sqrt{(X_{eye} - X_{mouse})^2 + (Y_{eye} - Y_{mouse})^2}$. Only the trials across participants were used, where the number of different x and y coordinates of the eyes matched the number of new mouse points, which was in this case 140 trials. These were selected to make a comparison between the two time series possible. In these trials, we found an average eye mouse distance of 141 pixels (*SD* = 20.2), which is relatively small in comparison to the average distance during web searching, which was around 257 pixels (Rodden et al., 2008). The selection of trials may have biased the size of the eye-mouse distance because these are the trials where there is the same amount of mouse and eye movement. This means that trials, where there was little mouse movement and a lot of eye movement or vice versa, were not considered.

To summarise, although the eye and mouse distance in general was not large for the selected trials, the mouse and eye movement were not always the same during the arithmetic task. Children use their eyes to look at the question, but do not move their mouse often to the question; they also explore a lot of answer options with their eyes apart from the correct response. The mouse movement can vary a lot depending on the strategy used, from either moving straight towards the clicked answer; first towards an alternative; or following the eyes along with the answer options. We analysed the mouse movements in terms of whether they were inside an AOI or not, similar to eye fixations. This may not be the best method, since mouse movements can be directed towards an object, but depending on the location they started, do not always end up inside an AOI. Therefore, in Chapter 4 a more dynamic method, irrelevant of the start location of the mouse, is investigated to quantify mouse movement.

First, in Chapter 3 eye tracking was used to infer whether the visibility of time pressure impacts arithmetic performance depending on children's individual cognitive profile.

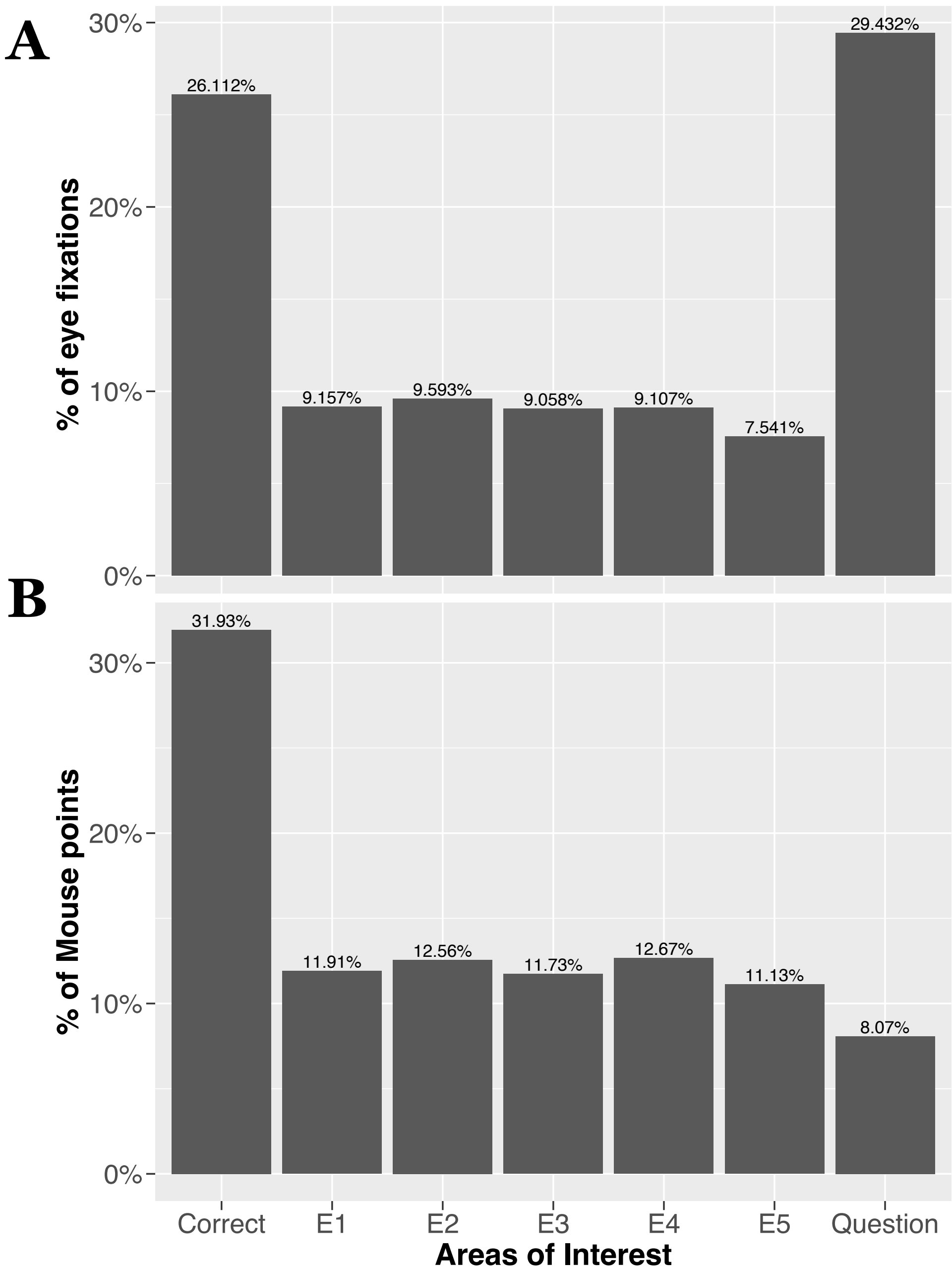


Figure 2.9. *Percentages of **(A)** eye fixations and **(B)** mouse points in the different AOIs. These percentages are not directly comparable since the AOI size was specified differently for the eye and mouse metrics. E1 to E5 are the most frequently made errors in Learn, where E1 is the most frequent error and E5 the least frequent.*

3

THE IMPACT OF VISIBLE TIME PRESSURE AND INDIVIDUAL COGNITIVE ABILITIES ON MATHS PERFORMANCE

ABSTRACT

This study investigated whether learning environments should consider the differential impact of cognitive load on children's maths' performance, depending on their individual verbal WM and IC capacity. An experiment was conducted with 39 children (8-11 years old) performing a multiple-choice computerised arithmetic game. Participants were randomly assigned to two conditions where the visibility of time pressure, a key feature in most gamified learning environments, was manipulated. Results showed that verbal WM was positively associated with arithmetical performance in general, but that higher IC only predicted better performance when the time pressure was not visible. This effect was mostly driven by the younger children. With the use of eye tracking, we showed that when time pressure was visible children attended more often to the question (e.g., 6 × 8). In addition, when time pressure was visible, children with lower IC, in particular younger children, attended more often to answer options representing operant confusion (e.g., 9 × 4 = 13) and visited more answer options before responding. These findings suggest that tailoring the visibility of time pressure, based on a child's individual cognitive profile, could improve arithmetic performance and may in turn improve learning.

The study of this Chapter has been published in the article:

de Mooij, S. M.M., Kirkham, N. Z., Raijmakers, M. E., van der Maas, H. L., & Dumontheil, I. (2020). Should online math learning environments be tailored to individuals' cognitive profiles? *Journal of Experimental Child Psychology, 191*, 104730.

3.1 INTRODUCTION

Extensive individual differences in learning trajectories, described in Chapter 1, show that in education there is no such thing as a one-size-fits-all approach. Emerging e-learning platforms, such as Learn, allow the tailoring of the learning environment to individual students on a larger scale. However, the environmental context in online game-based learning environments, with its interruptions and distractions, poses a risk for the user in terms of sustained attention, engagement, and concentration (Terras & Ramsay, 2012). To maximise the learning potential offered by adaptive e-learning platforms we also need to consider individual differences in the capacities to attend to, process, learn and remember information when designing these technologies (Ramsay & Terras, 2015).

3.1.1 *Cognitive overload on working memory*

Individual differences in WM are known to be associated with arithmetic performance (Bull & Lee, 2014; Dumontheil & Klingberg, 2012; Friso-van den Bos et al., 2013; Peng et al., 2015; Raghubar et al., 2010, as is reviewed in section 1.6.1). When solving mathematical problems, the overall load on an individual's cognitive system, referred to as *cognitive load*, can limit and interfere with performance (Sweller, 1988). This relates particularly to attention and WM. In online game-based learning environments, there is a great risk of overloading a player's WM with the rich number of multimedia elements and gamified features, causing limited capacity for the main task of problem-solving (W. H. Huang, 2011; Kiili, 2005; Moreno & Mayer, 2003). A cognitive overload on WM capacity may constrain both the acquisition of reasoning skills and the acquisition of knowledge (A. D. Baddeley, 1992; Eylon & Linn, 1988).

3.1.2 *Inhibitory control skills and attention in a gamified environment*

The IC level, the ability to selectively attend to relevant stimuli and, therefore, inhibit their attention to irrelevant stimuli (i.e., distracting stimuli or thoughts) and resist interference, is equally important to consider in a gamified environment (Diamond, 2013, see section 1.5.3). IC skills have been found to associate with mathematical performance (Bull, Johnston, & Roy, 1999; Bull & Scerif, 2001; Espy et al., 2004; St Clair-Thompson & Gathercole, 2006, see section 1.6.2). Studies have also looked at the impact of varying cognitive load on maths performance and found that this depends in part on the individual IC level (Avgerinou & Tolmie, 2020). In online game-based learning environments, task-irrelevant distracting stimuli, such as gamified sounds, flashing objects and alternative answer options, can trigger errors. In a laboratory setting, a similar effect occurs in the Simon task (Simon, 1969), where irrelevant information impacts the performance (see also section 1.5.3) Interestingly, individual differences in IC seem to moderate the impact on accuracy and reaction time (Mandich et al., 2003; Tsai et al., 2009). A similar Simon effect needs to be tested in a gamified environment. Furthermore, Bull et al. (1999) and Rourke (1993) suggest that a lack of IC is reflected in the type of errors children tend to make, for example the inability to switch away from addition when multiplication is required (i.e., operant-related error).

3.1.3 *Individual differences in attention and WM and performance under pressure*

Interference and cognitive overload in a learning environment do not always stem from external stimuli, but can also be internal in the form of worries about individual performance or about perceived time pressure (Ashcraft & Kirk, 2001; Mendl, 1999). These stressors can either drive people to use more efficient strategies (i.e., the best speed-accuracy trade-off within the constraints of the new situation) or compete with the attention that is normally allocated to the execution of the task (Caviola et al., 2017; Starcke & Brand, 2012). The latter is also known as the adverse effect of 'choking under pressure', where individuals perform worse than if there were no pressure (Baumeister, 1984; Beilock & DeCaro, 2007; Lewis & Linder, 1997). Critically, studies have found that people with high WM capacity are more affected by such a dual-task environment and suffer more under pressure than those with low WM capacity (Beilock & Carr, 2005; Sattizahn et al., 2016; Z. Wang & Shah, 2014). In addition, Sattizahn et al. (2016) found that individuals' variability in attentional control processes influenced the effect of pressure. Only those with poor attentional processing, measured with a flanker task, suffered decreased performance under pressure. This finding reflects that some individuals can limit the interfering effect of pressure on their performance, whereas others with poorer attentional control cannot. So, although increased WM and IC are generally associated with better maths performance and efficient strategy use, research has shown that it also depends on the stressors in the environment. The purpose of this study was to investigate the impact of stressors in the relatively new context of an online learning environment.

3.1.4 *Particular stressor: Visibility time pressure*

One particular stressor, typical to a lot of online game-based learning environments, is time pressure, usually presented in the form of a gamified visual stimulus. For example, in Learn there is visual time pressure in the form of coins counting down every second which is also incorporated in the game's scoring rule for maths performance (see section 2.1.5). The advantage of using time pressure is that it provides the opportunity to relate the speed of processing to the ability of the child, which is most valuable with easy problems (Klinkenberg et al., 2011; van der Maas & Wagenmakers, 2005). Also, in the case of games (similar to sports), the challenge of acting within a time limit can make the activity more enjoyable (Freedman & Edwards, 1988). Because time pressure itself is invaluable for most game-based learning environments, the current study addresses a different question: should the visibility of the time pressure (in the form of a countdown) be adapted for individuals depending on whether it negatively affects maths performance? Following the interference and overload theory, time pressure in the form of animated visual stimuli could be a distracting component that negatively interferes with solving maths problems, depending on the child's level of IC and WM. However, the alternative situation with no visible reminder of time passing by requires attention to be allocated to time perception. This capacity reduction could result in suboptimal strategies in the speed-accuracy trade-off for the main task (S. W. Brown & Perreault, 2017; Grondin, 2010; Matthews & Meck, 2016; Zakay, 1993)

3.1.5 *The current study*

The purpose of this study was threefold. First, we wanted to replicate the association of individual differences in verbal WM and IC with arithmetic performance of simple addition and multiplication problems in blocks of single or mixed operations in a game-based environment for primary school children (see section 1.6). We expected that both verbal WM and IC would positively associate with arithmetic performance overall. The mixed blocks were added to examine, whether the level of IC associated especially with an inability to switch between the operations, as previous studies have found (Bull et al., 1999; Rourke, 1993). If this is true, it would mean that the impact of the individual cognitive profile in a gamified environment depends on the task required, e.g., single or mixed operation arithmetic task. In line with the previous studies, we expected that higher IC would be associated with a reduced cost of switching between multiplication and addition. Second, we explored whether a particular feature of cognitive load, the visibility of time pressure, would affect arithmetic performance in general. More importantly, we were interested in whether the impact depends on the child's WM and IC skills. We did not have a hypothesis regarding whether visibility or invisibility of time pressure associated with worse arithmetic performance since both features create a dual-task condition. Any effect on arithmetic performance was expected to interact with individual differences in WM and/or IC.

Finally, whether the learner is attending to or actively inhibiting their attention to irrelevant/distracting stimuli was studied through the eye fixations using eye tracking (Duprez et al., 2016; Wijnen & Ridderinkhof, 2007). In general, WM load is thought to increase the number of fixations (Just & Carpenter, 1976). More specific, other studies have found that looking back and forth at the question is positively associated with attentional and WM load

(Droll & Hayhoe, 2007; Orquin & Loose, 2013). With the use of eye tracking, we explored differences in order and duration of gaze to various elements of the display during the arithmetic task, depending on whether time pressure was visible or not and the children's levels of WM and IC.

This study included data from a single time point, and, therefore, and this cannot be stressed enough, will not inform our understanding of how individual differences and task features affect learning over time. However, a better understanding of how performance in online maths tasks may be affected by these factors could allow tailoring of the environment to individual learners, making sure that the task challenges, and therefore trains, their arithmetic skills rather than loading on other aspects of their cognitive capacity.

3.2 METHODS

3.2.1 *Participants*

A total of 42 primary school children aged 8 to 11 years were recruited through a local voluntary participant database and word of mouth. Three children were excluded from all analyses because testing sessions were interrupted due to distress or tiredness. The final sample consisted of 39 children (19 male; *M* = 9.60 years old; SD = 1.02; range = 8.00-11.50). With a sample size of 39, the study had 80%, 90%, and 95% power to detect large eta-square (η^2) effect sizes of 0.18, 0.22 and 0.26 respectively when comparing two groups. Eta-square effect sizes have been classified as follows: small η^2 = 0.02; medium η^2 = 0.13; large η^2 = 0.26 (Cohen, 1988).For three children insufficient eye gaze data were collected, leaving 36 children (18 male; *M* = 9.67 years old; *SD* = 1.00; range = 8.00-11.50) for the eye tracking analyses. The study was approved by Birkbeck's Department of Psychological Sciences ethics Board. Informed consent was given by caregivers, and verbal consent was given by the participants.

3.2.2 *Procedure*

All stimuli were presented in MATLAB using the Psychophysics Toolbox (Brainard, 1997). During the first task, participants performed an arithmetic task on a computer (see **Fig. 3.1A**) similar in design to Math Garden (Straatemeier, 2014). The study took place in a lab setting, and all measures were completed in a single session lasting about 30 min in total. Before data collection started, condition assignment was randomised for a list of 40 participants using MATLAB. An additional two participants were tested to compensate for incomplete or withdrawn participants. Participants first completed the arithmetic task, then the verbal WM task (backward digit span task), and then the IC task (Simon task), which took ~ 30 min. Eye tracking and mouse tracking data were collected during both the arithmetic and Simon task.

3.2.3 *Arithmetic task*

The arithmetic task was divided into three blocks comprising 20 multiplication problems, 20 addition problems, and 22 mixed multiplication and addition problems, respectively. The order of the blocks was counterbalanced across participants. The trials within the blocks were fully randomised. This arithmetic task was based on the speed mix game of Math Garden (see Appendix A.1 for details on this game) and all problems involving single-digit numbers between 1 and 9 stems from Math Garden. For each arithmetic problem, participants were asked to choose one of six answer options, which consisted of the correct answer and the five most frequent errors, made by children of similar age on that arithmetic problem. The frequency was based on Math Garden data previously collected from a large Dutch sample aged between 5 and 13 years old (**Fig. 3.1A**). Participants had a maximum of 8 seconds to click on one of the answers, after which the correct answer was highlighted. In a between-participants manipulation, 19 children were assigned to the visible time pressure condition, where the time limit of 8 seconds was visible in the form of coins counting down at the bottom right of the screen, similarly to Math Garden (**Fig. 3.1A**). The other 20 children needed to respond within the same 8 seconds, but there were no coins on the screen (no visible time pressure condition). After every trial, direct feedback on performance was given; the correct answer was circled in green, and the incorrect answer was circled in red in the case of an incorrect response. The measure of arithmetic performance was calculated with a scoring rule following the equation $S = (2X_{pi} - 1)(d - T_{pi})$ (adapted from Maris & van der Maas, 2012, see also section 2.1.5). This rule imposes a speed-accuracy trade-off, where fast and correct responses result in a high score and incorrect responses result in a negative score. X_{pi} is 1 for correct and 0 for an incorrect response for player p on trial i. $T_{pi.}$ is the response time in seconds; range = 0 to 8) before the imposed time limit d (set to 8 seconds in this study) and eventually, a score S is obtained (range -8 to 8).

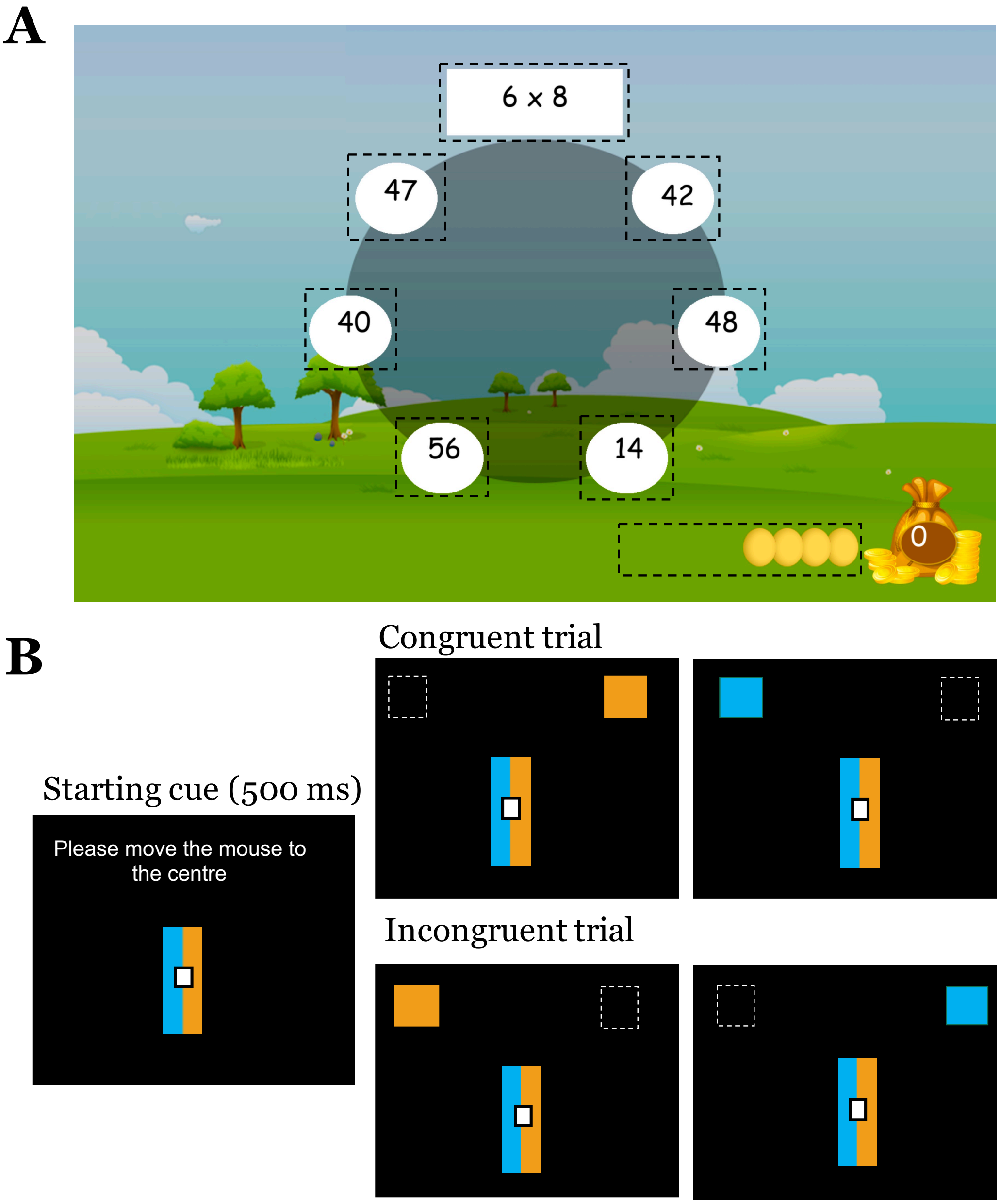


Figure 3.1. *Experimental Tasks.* ***(A)*** *Screenshot of the arithmetic task. Here the problem 6 × 8 is presented at the top of the display with the six answer options underneath. For half of the participants, a visible time pressure was implemented through coins counting down every second. The current total score is depicted in the right bottom corner. The dotted black lines are drawn to represent the areas of interest (AOIs) for the gaze data.* ***(B)*** *Setup of the Simon task. A cue indicating the correct colour–response mapping remained on the screen at all times. In this example, the cue indicates that the left box should be clicked for blue target stimuli and that the right box should be clicked for orange target stimuli. Participants first moved their mouse to the centre of the display (small white square). After 500 ms, a blue or orange target square stimulus was presented in either top corner of the display. Participants were asked to move their mouse toward and click into the box corresponding to the colour of the target stimulus. On congruent trials, the location of the target matched the response associated with the colour of the target (e.g., orange target on the right side). When colour and location did not match (e.g., orange target on the left side), the trial was incongruent and a conflict arose.*

3.2.4 *Backward digit span task*

Participants' verbal WM was then assessed with a backward digit span task since this task measures verbal WM for numerical information. In this task, children were asked to repeat, backwards, lists of single-digit numbers pronounced by the experimenter. After practising with two numbers, the first level consisted of four lists of three numbers; the child moved one level up (with an additional number) when at least three of the four lists were repeated back successfully, otherwise testing stopped. The lists of numbers were fixed across participants. A WM score was computed as the total number of correct answers.

3.2.5 *Simon task*

IC was assessed with a computerised spatial incompatibility Simon task (adapted from Duprez et al., 2016; see **Fig. 3.1B**). This task was used to measure how able children are in blocking irrelevant information. Children were asked to move their mouse to either the top left or top right box depending on the colour of the target square while ignoring its location. When the target was blue children needed to move and click their mouse into the left box. When it was orange, they needed to move their mouse toward and click into the right box. In half of the trials, the location of the target was congruent with the correct response, and in the other half the location was on the opposite side of the correct response (**Fig. 3.1B**). Participants completed 40 trials in a randomised order, which resulted in between 1 and 5 trials of the same type (congruent/incongruent) being repeated in a row. The measure of IC, referred to as the IC interference effect, was computed as the difference between incongruent and congruent trials mean RT divided by congruent trials mean RT, using correct trials only. A high score reflects slower RT on incongruent trials (i.e., difficulty in inhibiting attention to irrelevant information) than on congruent trials (i.e., baseline processing speed).

3.2.6 *Eye tracking*

During the arithmetic and Simon task, children were seated at a distance of 60 cm in front of an eye tracker. The application use of an eye tracker is explained in more detail in section 2.2. Eye movements were recorded using a Tobii TX300 (Tobii, Danderyd, Sweden) at a sampling rate of 120 Hz. The raw data were classified into fixations and saccades using the *gazepath* package in R (Team, 2013; van Renswoude et al., 2017). Gazepath uses an algorithm to categorise the data into fixations and saccades while accounting for individual differences and data quality. Fixations in the arithmetic task were labelled as one of the following three AOIs: (a) the question box, (b) one of the six answer options, or (c) the coins (i.e., the visible countdown of time) (**Fig. 3.1A**).

3.2.7 *Statistical analyses*

Data management and statistical analysis were performed using R Software (Team, 2013). For all independent variables, z scores were generated to standardise the scores for further analyses. In the first set of analyses, the arithmetic performance was averaged over the three blocks (addition, multiplication, and mixed addition and multiplication) and compared between the visible time pressure condition and the not visible time pressure condition. This was analysed using between-participants three-way analyses of covariance (ANCOVAs), covarying for age and WM score or IC interference effect. An additional analysis investigated associations between IC and the cost of needing to switch between operations. We subtracted the average performance of the mixed block trials from the average performance on the trials in the single operation blocks, for multiplication and addition problems separately. These cost measures were entered into ANCOVAs including the IC interference effect, visibility of time pressure, and age for multiplication and addition separately. Assumptions of the ANCOVAs were met, with analyses showing homoscedasticity and normality of the residuals.

Eye tracking analyses (N = 17 in the visible time pressure condition; N = 19 in the not visible time pressure condition) focused on correct trials (excluding 12.7% of trials) and trials where there was at least more than one fixation to ensure high eye tracking data quality (excluding a further 1.2% of trials). The average number of fixations and the proportional duration of fixation on each AOI were calculated for each participant. An additional metric was the average number of answer option AOIs the participant attended to on a trial. We explored, in three-way ANCOVAs, whether these eye-tracking metrics differed according to the visibility of time pressure and whether this interacted with WM score, IC interference effect, or arithmetic performance.

The data were checked for outliers using a criterion of $|z \text{ score}| > 3$ for both the dependent and independent variables. No outliers were identified. In the regression analyses, Cook's distance suggested one to three influential points for some behavioural and eye tracking results. Analyses were repeated excluding these data points, and the results were strengthened except in one case, which is discussed further below.

In addition, Bayesian ANCOVAs were performed post hoc for the results with null effects or p values just under the threshold ($p < .05$) using JASP (JASP Team, 2019). To quantify uncertainty about effect size and to obtain evidence in favour of a null hypothesis (Wagenmakers et al., 2018), we distinguished between experimental insensitivity (Bayes

Factor [BF]$_{10}$ and $BF_{01} < 3$) and robust support for the alternative hypothesis ($BF_{10} > 3$) or null hypothesis ($BF_{01} > 3$; Dienes, 2014)

3.3 RESULTS

3.3.1 *RT and accuracy*

I performed two one-sided equivalence tests with alpha = .05 and no assumption of equal variance. The analyses showed statistical equivalence between the visible and no visible time pressure groups for age, percentage female, verbal WM, and IC (**Table 3.1**).

Table 3.1. *Comparison of demographic and behavioural measures between the visible time pressure group and the no visible time pressure group.*

Variable	Visible time pressure (N = 19) [Mean (*SD*)]	No visible time pressure (N = 20) [Mean (*SD*)]	*TOSTs of equivalence (95% CI)*		*p*
Age	9.46 (0.97)	9.73 (1.08)	-.28	.82	.020
Proportion female	.58	.40	-.09	.45	<.001
WM digit score	8.68 (3.42)	8.75 (3.08)	-.53	.57	.002
IC interference effect	0.10 (0.08)	0.09 (0.12)	-.63	.46	.004
RT arithmetic task	4.08 (0.88)	3.85 (0.94)			
Proportion correct, arithmetic task	0.81 (0.16)	0.83 (0.17)			
Proportion no response, arithmetic task	0.09 (0.08)	0.08 (0.11)			
Arithmetic score	3.19 (1.34)	3.48 (1.35)			

Note. TOST, two one-sided equivalence test; CI, confidence interval

T-tests were run to test whether the visibility of time pressure was associated with arithmetic performance. We did not find any difference between the groups in mean RT, t (38) = 0.82, p = .42; nor for proportion of correct responses, t (38) = 0.45, p = .66; nor for proportion of no response within the time limit, t (38) = 0.15, p = .88; or for mean arithmetic score, t (38) = 0.77, p = .45 (**Table 3.1**). These comparisons indicate that the visibility of time pressure did not affect arithmetic performance. The average overall arithmetic performance was 3.34 (*SD* = 1.34), meaning that on average responses were correct and answered within roughly half the time limit (see section 2.1.5 for the scoring rule). The overall arithmetic performance measure was used for further analyses. In addition, the mean accuracy in the Simon task was high (*M* = .99, *SD* = .05). As expected, RTs differed between congruent and incongruent trials, t (38) = 6.17, $p < .001$. Participants were on average 150 ms slower in incongruent trials (**Fig. 3.2 A**). This individual average IC interference effect was used as a measure of IC for further analyses (**Fig. 3.2 B**).

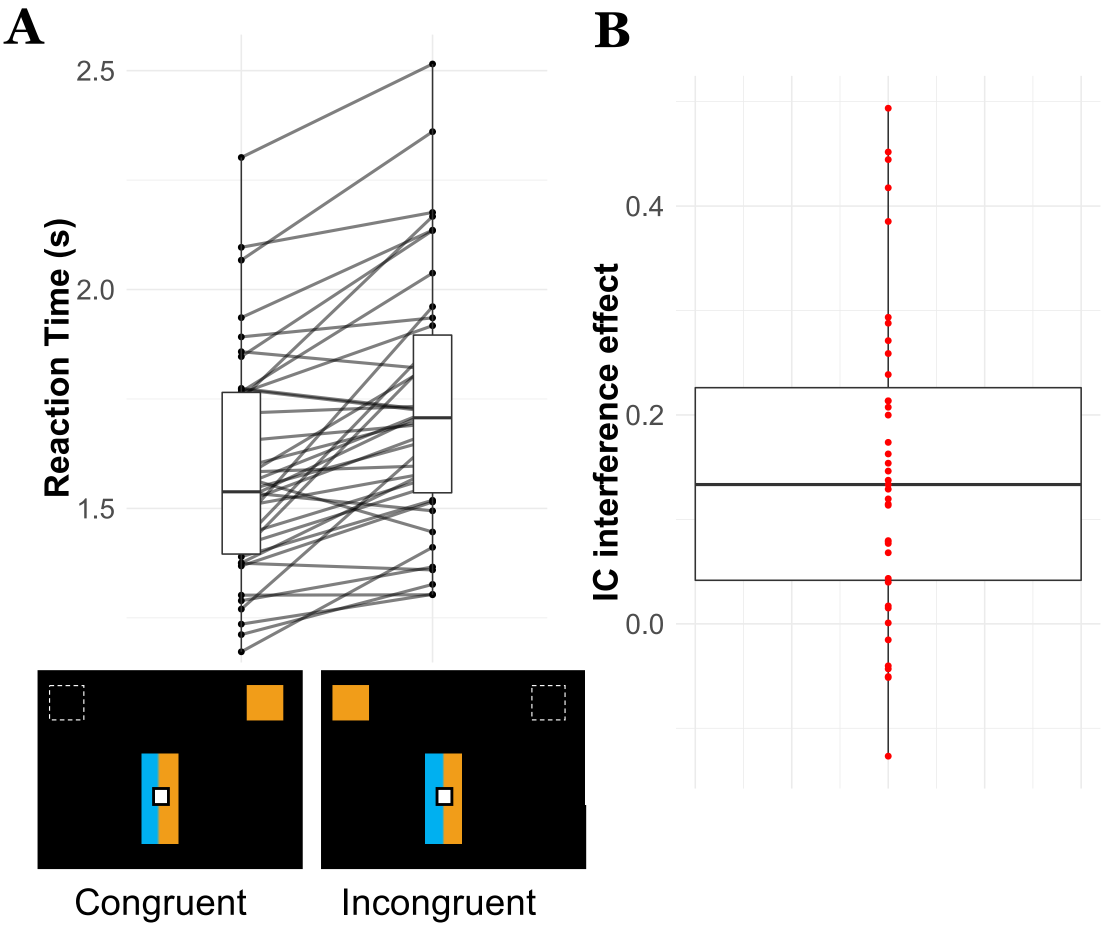


Figure 3.2. *Interference effect on reaction time (RT) in the Simon task of inhibitory control (IC). (A) Box plots of individual mean RTs as a function of trial type (congruent vs. incongruent). (B) Box plot of the IC interference effect, calculated as the difference between the incongruent and congruent trials mean RT divided by congruent trials mean RT for correct trials only.*

3.3.2 *Impact of time pressure on arithmetic performance depending on the level of IC and verbal WM*

The first analysis included only age and visibility of time pressure as predictors of arithmetic performance. The ANCOVA showed a positive association between age and arithmetic performance, $F(1, 35) = 32.05$, $p < .001$, $\eta_p^2 = 0.53$ but not time pressure ($p = .892$, $\eta_p^2 = 0.00$) nor was there an interaction between age and time pressure ($p = .679$, $\eta_p^2 = 0.01$). The second analysis included WM score as a covariate (**Table 3.2**). WM score positively associated with arithmetic performance, $F(1,31) = 13.15$, $p = .001$, $\eta_p^2 = 0.24$ (**Fig. 3.3A**). But there was no interaction with age or time pressure (all p's > .50, $\eta_p^2 < 0.01$). The Bayesian ANCOVA showed that a null model with merely main effects for WM score and age was 11.4 times more likely than including any of the above-mentioned interactions or the main effect of time pressure. The third analysis (**Table 3.2**) included the IC interference effect as a covariate. The individual differences in IC and WM correlated poorly with each other, $r = -0.03$. There was also no main effect of IC interference effect $p = .62$, $\eta_p^2 = 0.01$ with arithmetic performance. The ANCOVA did show a significant two-way interaction between time pressure and IC interference effect, $F(1,31) = 6.59$, $p = .015$, $\eta_p^2 = 0.18$, as well as a three-way interaction between time pressure, age and IC interference effect on arithmetic performance, $F(1,31) = 4.55$, $p = .041$, $\eta_p^2 = 0.13$. Significant evidence for both interaction effects was demonstrated through Bayesian analyses (**Table 3.2**).

To examine the two- and three-way interactions, separate multiple regressions were performed in the visible and no visible time pressure groups. In the group with visible time pressure, the IC interference effect and Age x IC interference effect interaction terms did not significantly predict variance in arithmetic performance ($BF_{01} = 0.57$, i.e., no evidence for either hypothesis) (**Fig. 3.3 B**). In contrast, the group with no visible time pressure showed a negative association between arithmetic performance and IC interference effect, $\beta = -0.42$, $t(16) = 2.77$, $p = .014$ ($BF_{10} = 6.47$, i.e., substantial evidence for including this effect) (**Fig. 3.3B**). Also, a significant interaction between age and IC interference effect was found, $\beta = 0.43$, $t(16) = 2.629$, $p = .018$ ($BF_{10} = 8.84$). The interaction effect showed that the association between arithmetic performance and IC interference effect was mostly driven by the younger children (**Fig. 3.3C).** However, this age IC effect could also be driven by the increased variability in math performance for the younger children, as **Fig.3.3C** shows.

Table 3.2. *Summary of effects observed in analyses of covariance of behavioural and eye tracking data.*

Covariate: WM	Age	Time pressure	WM	Age x Time pressure	WM x Time pressure	WM x Time pressure x Age
1. Behavioural data[a]						
Arithmetic performance	$\eta_p^2 = .53^B$	Null[B]	$\eta_p^2 = .24^B$	Null[B]	Null[B]	Null[B]
2. Eye tracking data (number of fixations)[b]						
Question box	-	$\eta_p^2 = .15^B$	-	-	-	-
Answer options	Null[B]	Null[B]	Null[B]	Null[B]	Null[B]	Null[B]

Covariate: IC	Age	Time pressure	IC	Age x Time pressure	IC x Time pressure	IC x Time pressure x Age
1. Behavioural data[a]						
Arithmetic performance	$\eta_p^2 = .53^B$	Null[B]	Null[B]	Null[B]	$\eta_p^2 = .18^B$	$\eta_p^2 = .13^B$
Operation switch cost on multiplication problems	Null[B]	Null[B]	Null[B]	Null[B]	Null[B]	Null[B]
Operation switch cost on addition problems	-	-	$\eta_p^2 = .13$	-	-	Null[B]

2. Eye tracking data (number of fixations)[b]						
Question box	-	$\eta_p^2 = .15^B$	-	-	-	-
Operation errors on multiplication problems	-	Null^B	-	-	$\eta_p^2 = .16^B$	-
Operation errors on addition problems	-	Null^B	-	-	-	-
Answer options	-	-	-	-	$\eta_p^2 = .13$	-

Note. Effect sizes of significant effects (*p*'s < .05) are reported. Cases, where robust support (Bayes factor [BF] > 3) for the alternative or the null hypothesis was provided by the Bayesian analyses of covariance, are indicated with a superscript letter B. Dashes (–) indicate that the main effect or interaction was not significant but that there was no strong evidence in support of the null hypothesis. [a] df = 31. [b] df = 2

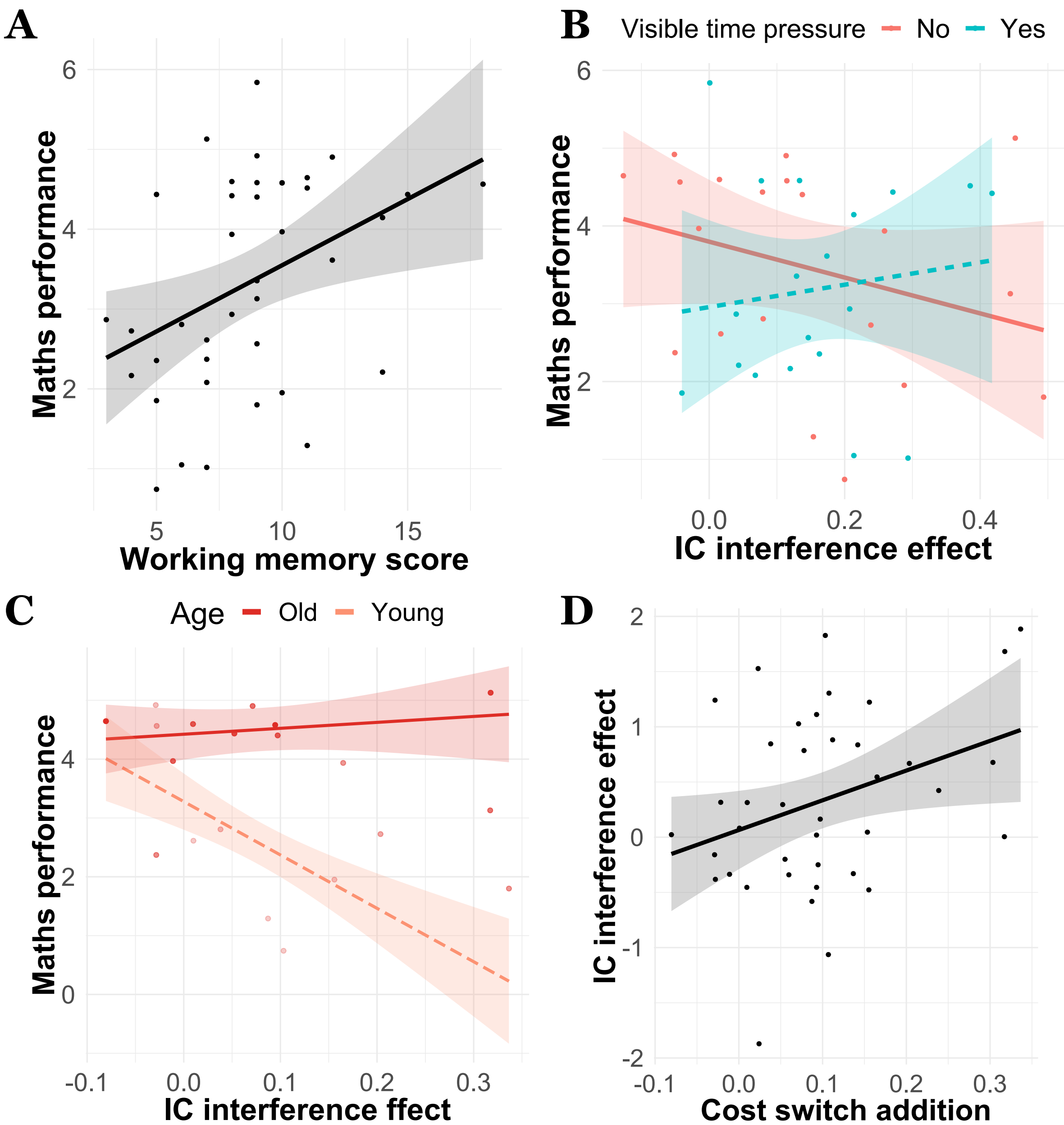


Figure 3.3. *Arithmetic performance (combined accuracy and reaction time score) as a function of verbal working memory (WM) score and inhibitory control (IC) interference effect.* ***(A)*** *WM score was positively associated with maths performance.* ***(B)*** *The association between maths performance and IC interference effect depended on the visibility of the time pressure.* ***(C)*** *Graph illustrating the Age x IC interference effect interaction on maths performance for the no visible time pressure group. A median split was performed for age showing two regression lines for younger (8.0 - 9.5 years) and older children (9.6-11.5 years), but note that age was treated as a continuous variable in the analyses.* ***(D)*** *The cost of mixing operations on performance of the addition problems (mixed operations block score – single operation block score) was positively predicted by the IC interference effect.*

3.3.3 *Operation switch cost*

To investigate whether switching between operations led to a cost in performance, we compared the mean arithmetic scores of single operation versus mixed operation blocks, for multiplication and addition problems separately. Paired *t*-tests showed that children's performance on multiplication problems did not differ between the mixed (M = 2.83) and single (M = 2.76) operation multiplication blocks, t (38) = 0.41, p = .341. For addition, children performed less well on the trials in the mixed block (M = 3.67) than on those in the single operation block (M = 4.00), t (38) = 2.51, p = .008. Therefore, children showed a cost of needing to switch between multiplication and addition on addition problems only.

Because the ability to switch between arithmetic operations has been associated with IC in previous studies (Bull et al., 1999; Rourke, 1993), additional analyses explored whether IC predicted the ability to switch between addition and multiplication operations in the mixed operation blocks compared with the single operation blocks (**Table 3.2**). As no cost of mixing was found for multiplication problems, this analysis was performed for addition problems only. The IC interference effect predicted the performance difference between the mixed and single operation blocks, F (1, 31) = 5.06, p = .031, η_p^2 = .13 (**Fig. 3.3 D**). The Bayesian ANCOVA revealed that a model including IC was 2.68 times more likely than the null model; no interaction with age (p = .302, η_p^2 = 0.03) or time pressure (p = .153, η_p^2 = 0.06) was shown.

3.3.4 *Eye fixations and patterns*

Exploratory analyses investigated whether eye movements during the arithmetic task could give some insight into the behavioural findings. Analyses were performed using the mean number of fixations and proportion of total fixation duration on specific AOIs. The latter did not show any significant effect.

The first analyses looked at the fixations on the question box AOI (e.g., 6 × 8 on **Fig. 3.1 A**). Previous studies have found that fixations on the question relate to attentional and working memory load (Droll & Hayhoe, 2007; Orquin & Loose, 2013), such that these studies found more and longer fixations on the question- used as an external memory space- to reduce working memory demands. Based on these studies we expected an increase in the number of fixations when participants could experience more cognitive load, due to low IC and WM capacity or visibility of time pressure. ANCOVAs were run to test for associations with the visibility of time pressure in interaction with individual differences in IC and WM separately, while covarying for age and arithmetic performance (**Table 3.2**). The analysis was done separately for IC and WM, to minimise the number of predictors in the model considering the limited sample size. A significant main effect for time pressure, $F(1,28) = 12.02$, $p = .003$, $\eta_p^2 = 0.43$ (BF_{10}= 30.88, i.e., very strong evidence), showed that there were more fixations on the question box when time pressure was visible ($M = 2.69$) than when there was no visible time pressure ($M = 2.02$; **Fig. 3.4A**).

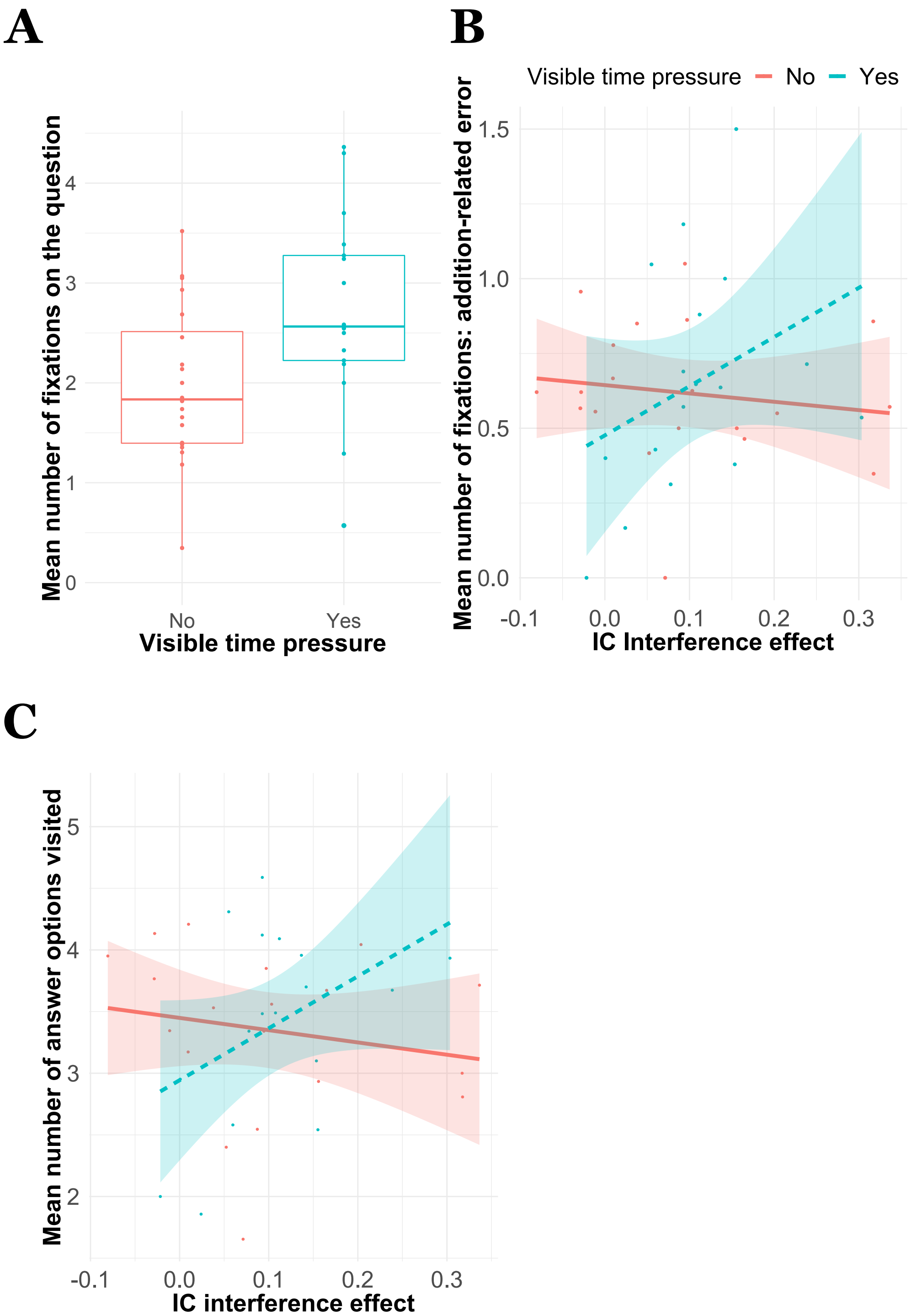


Figure 3.4. *Eye tracking metrics showing significant associations with the visibility of time pressure and the inhibitory control (IC) interference effect.* ***(A)*** *The average number of fixations on the question box was higher in the visible time pressure group compared with the no visible time pressure group.*

(B) *The mean number of fixations on addition-related errors was positively associated with the IC interference effect in the visible time pressure group only.* ***(C)*** *The mean number of attended answer options was positively associated with the IC interference effect when the time pressure was visible.*

Second, because operation-related errors were found to associate with the level of IC (Bull et al., 1999; Rourke, 1993), fixations on the operation-related error answer options were also investigated. Following these studies, we expected an increased number of fixations on this specific error answer for individuals with lower levels of IC. We explored whether the visibility of time pressure, age and maths performance covaried, similar to the other eye tracking analyses. Again, separately for addition and multiplication collapsing the single and mixed blocks. ANCOVAs were run to test for associations with the visibility of time pressure and the IC interference effect, covarying for age and arithmetic performance. First, for the addition problems with multiplication-related errors as answer options, we found no significant predictors (p's > .20; $\eta_p^2 < 0.05$) (**Table 3.2**). For multiplication problems with addition-related errors as answer options, a significant interaction was found between time pressure and IC interference effect, F (1,28) = 5.34, p = .018, $\eta_p^2 = 0.44$ (BF_{10}= 5.21, i.e., substantial evidence). **Fig. 3.4.B** shows this interaction, such that the mean number of fixations on the addition-related error increased with increasing IC interference effect (β = 0.56) only when time pressure was visible.

Finally, analyses were performed to investigate the mean number of answer options participants looked at before giving their answer. The question was whether this was related to WM, IC, and the visibility of time pressure. In general, it was expected that participants who experienced high cognitive load, also showed more fixations (Just & Carpenter, 1976; Orquin & Loose, 2013). An ANCOVA was performed with the average number of answer options attended to as the dependent variable, visibility of time pressure and IC interference effect or WM as independent variables, and age and arithmetic performance as covariates. The analysis with WM as a predictor showed no main or interaction effects, only evidence that a null model was 11 times more likely than including any of the predictors. The analysis with IC as a predictor showed a significant interaction between visibility of time pressure and the IC interference effect, F (1,31) = 4.60, p = .039, $\eta_p^2 = 0.13$. However, Cook's distance highlighted that there was one influential point that drove this interaction. Consistent with this, only anecdotal evidence (BF_{10} = 2.90) for including this interaction to the null model was found in the Bayesian regression (**Fig. 3.4C** and **Table 3.2**). Follow-up regression analysis showed a trend for a positive association for IC interference effect when time pressure was visible, β = .54, t (13) = 2.03, p = .063, but little evidence (BF_{10} = 1.23) in the Bayesian regression. No association between the IC interference effect and the number of answer options visited, was found when time pressure was not visible, β = -.21, t (16) = 0.82, p = .423; Bayesian regression showed anecdotal evidence for null hypothesis (BF_{01} = 2.00).

3.4 DISCUSSION

This study combined behavioural and eye tracking measures to test whether individual differences in verbal WM and IC in primary school children could predict their ability to solve arithmetic problems in different online learning environments, where visibility of time pressure was varied. The behavioural results showed that verbal WM was a positive predictor of arithmetic performance in general and that this association was independent of the visibility of time pressure. In contrast, individual differences in IC predicted arithmetic performance only when the same time pressure was not visibly illustrated by an animation. Also, we found that this association with IC was mostly driven by younger children. Eye tracking results showed that the children fixated on different parts of the stimuli during the arithmetic task depending on the visibility of time pressure, their IC level and age.

As expected and in line with previous studies, WM related to mathematical performance in children, in this case verbal WM (see Raghubar et al., 2010 for a review). For the IC interference effect, a general association with maths performance was expected, but the association was found to depend on factors such as visibility of time pressure and age. Surprisingly, there was no strong relation between the IC interference effect and verbal WM, whereas you would expect these factors to covary, given the underlying model for EF (Friedman & Miyake, 2016, see section 1.5.1). In this study, only one task was administered to measure IC, but for future research it would be interesting to measure other related IC tasks such as the Stroop task (MacLeod, 1992) and Flanker task (Eriksen & Eriksen, 1974).

The study was conducted investigating simple multiplication and addition problems, in blocks of single and mixed operands. This enabled us to investigate whether there was an overall cost of switching between operands and whether this was impacted by the individual cognitive profile and by the visibility of time pressure. Our results showed a switch cost for addition-related problems between mixed and single blocks, but not for multiplication-related problems. During fact retrieval (see section 4.1.2) such as Campbell's network interference theory (1995) suggests, operation-related associations are thought to be activated during fact retrieval, especially when the operation problems are mixed. Our finding for addition-related problems resembles this theory, such that in mixed blocks children make more errors and therefore achieve a lower performance due to the network interference of multiplication facts.

Since there was no switch-cost found for multiplication problems, it could be that the children from this study did not already use retrieval strategies for these problems but procedural strategies, where there is not yet or less interference from operation-related associations. What we also found is that the individual level in IC interference effect predicted the switch-cost in operations for addition. This suggests that our IC interference effect is underlying more cognitive flexibility, another aspect of executive functions, measuring the skill to shift your attention, e.g., to a different operator (Dajani & Uddin, 2015).

Overall, the findings point out that an animated countdown may affect the performance of certain individuals and that possible constraints of attentional control (i.e., the amount of interfering information compromising cognitive resources) should be considered. The presence or absence of a time pressure reminder in a task can both create dual-task environments, leading to less attention to the main task of solving math problems. When there is visible time pressure, the user has a constant physical reminder of timing (in this study in the form of an animated visual stimulus). Adding more visual stimuli and time pressure is suggested by previous studies to contribute to loading on WM capacity, leading to suboptimal strategies and attention (Barrouillet et al., 2007; Caviola et al., 2017; Terras & Ramsay, 2012). The magnitude is thought to be influenced by other individual differences such as maths anxiety (Ashcraft & Krause, 2007; Caviola et al., 2017; Kellogg et al., 1999), engagement and attitude to learning (Barkatsas et al., 2009; Kebritchi et al., 2010). Although the visibility of time pressure did not interact with individual differences in verbal WM in terms of the impact of arithmetic performance, the notion of visible time pressure as an increasing demand on WM resources is reflected in our eye tracking results. Children made more fixations on the question in the visible time pressure condition than in the invisible time pressure condition, suggesting that children may have found it more difficult to keep the question in mind with time pressure present (Orquin & Loose, 2013). Although previous studies suggested that the impact of extra stressors on mathematic performance depends on the ability to resist distractions (i.e., IC; Sattizahn et al., 2016), we showed that the performance of children was not affected by their level of inhibition when time pressure was visible. The higher number of fixations on answer options and operation-related errors did suggest that for children with lower IC the task was more demanding in terms of decision difficulty and/or attentional resources (Orquin & Loose, 2013), but this did not result in lower arithmetic performance.

Time perception is intensively studied (for an overview of recent reviews, see Block & Grondin, 2014) and involves diverse perceptual, motor, cognitive and brain processes (Block & Gruber, 2014). One line of investigation in time perception concerns its bidirectional interference with higher-level executive cognitive processes such as mental arithmetic but also with executive functions (Block et al., 2010; S. W. Brown et al., 2013). Bidirectional interference occurs in a dual-task condition, where time perception competes for the same attentional resources as the other task, leading to cognitive load. Because the interference is bidirectional, studies have also shown that lower IC is associated with less accurate time perception (S. W. Brown & Perreault, 2017; Meaux & Chelonis, 2005). This bidirectionality finding closely aligns with our finding that children with low IC showed lower arithmetic performance when they also needed to estimate time without a reminder. There could be an impairment of time perception, such that the low IC children have trouble in deciding on an optimal speed-accuracy trade-off strategy. Therefore, a visible time countdown might reduce cognitive load for children with low IC, whereas children with high IC seem to be able to estimate time in parallel with solving arithmetic problems.

One of the limitations of this study was the small sample size. This was due to the use of an eye tracker in a lab setting. The use of a participant volunteer database and testing in a lab setting also likely biased our recruitment toward children from higher socioeconomic backgrounds and with higher cognitive abilities. The next step would be to replicate our findings with a larger heterogeneous sample from online learning environments such as Learn to ensure that the behavioural findings are reliable. Also, we chose a between-participants design that minimises the effect of learning and testing time. A within-participants design would have had more power to detect interactions between the time pressure manipulation and individual differences in WM and IC. Future work should investigate whether learning, rather than performance at a single time point, can be improved based on an adapted environment, informed by the results of the current study. Although the purpose of this study was to implement these findings in an online adaptive environment, the arithmetic problems were standardised to ensure that we could compare arithmetic performance within the sample size. Due to our wide age range (8–11 years), certain arithmetic problems were inevitably less challenging for some children; therefore, all analyses were covaried for age. Note, however, that there are large individual differences within year groups on arithmetic tasks (Straatemeier, 2014), so a more homogeneous sample in terms of age may have still shown considerable variability in arithmetic performance. To further investigate whether the associations among IC, WM, time pressure and arithmetical outcome change with age, the difficulty level of the arithmetic problems should be adapted to the ability of the child. Finally, whereas we considered the coins countdown to reflect time pressure, it also indicated the potential reward to be gained when correctly solving the problem. Although the obtained reward was shown to both groups of participants when a trial was completed, the group with no visible coin countdown did not have a constant reminder of the potential reward. This reward cue difference between the groups may have led to some of the differences observed between the conditions.

In conclusion, we found that the (in)visibility of time pressure, a key feature in a lot of online game-based learning environments and psychological tasks in general, may create cognitive overload and affect the application of knowledge and skills. Specifically, we showed that this aspect of online game-based learning environments may differentially affect children's arithmetic performance as a function of their cognitive abilities. Measuring the individual levels of cognitive functioning, in particular WM and IC, is essential to allow children to perform and practice tasks at their highest level. In addition, the use of an eye tracker in this context allowed an in-depth exploration of how learners interacted with the different elements in the environment above and beyond accuracy and RT. Future work should focus on developing a broader online adaptive framework for learning mathematical skills and knowledge that adapts not only to children's mathematical skills but also to their more general cognitive strengths and weaknesses.

4 ERROR DETECTION THROUGH MOUSE MOVEMENT

ABSTRACT

While response time and accuracy are measures of overall performance, their value in uncovering the cognitive processes underlying problem-solving is limited. A promising online measure designed to track the problem-solving process is computer mouse tracking, where mouse attraction towards different locations may reflect the consideration of incorrect response options. Using a speedy arithmetic multiple-choice game in Learn, we examined whether mouse movements could reflect arithmetic difficulties when error rates are low. Results showed that mouse movements towards incorrect responses in correctly answered problems mapped onto the overall frequency of errors made in this online learning system. This mapping was stronger for the younger children, as well as for easy arithmetic problems. On an individual level, users showed more mouse movement towards their previously made response errors than towards other incorrect options. This finding opens the possibility of adapting feedback and instruction on an individual basis through mouse tracking.

The study of this Chapter has been published in the article:
de Mooij, S. M.M., Raijmakers, M. E., Dumontheil, I., Kirkham, N. Z., & van der Maas, H. L. (2020). Error detection through mouse movement in an online adaptive learning environment. *Journal of Computer Assisted Learning, 191*, 104730.

4.1 INTRODUCTION

To create the optimal learning environment, tailoring the instruction and feedback to the needs of individual learners is necessary (Federico, 2000). However, to individualize learning materials (Bray & Mcclaskey, 2010), we need to understand the underlying thinking processes that drive particular behavioural responses during learning: we need to understand why a child may consistently have difficulties when faced with a particular type of problem. As was investigated in the previous lab study (Chapter 2 and 3), mouse and eye tracking are process-tracing paradigms that serve as indirect measures (to varying degrees) of underlying cognitive processes such as attention, decision making and information processing. In the present study, mouse tracking was used to measure difficulties children might have during arithmetic problem-solving, on a larger scale in the online learning environment of Learn.

4.1.1 *Individual differences in arithmetic difficulties*

Systematic difficulties in arithmetic differ between students because they all have different learning trajectories with a different set of strategies they use. Revealing these systematic difficulties can help diagnose incorrect understanding, sometimes called misconceptions (J. S. Brown & Burton, 1978). Students' errors can be used to track individual systematic difficulties and to give insight into students' state of knowledge (see section 1.7.4). The downside of this approach is that it requires students to make enough errors that a pattern can be identified.

Computer adaptive testing has a big advantage that the item selection algorithm can create a challenging and yet motivating situation for students to practice their skills. To keep students motivated, adaptive learning systems usually choose a high success rate and, as a consequence, few errors are made (Eggen & Verschoor, 2006; Jansen et al., 2013). Having to rely on the limited number of incorrect responses recorded per student makes it almost impossible to rapidly detect underlying systematic difficulties in adaptive learning systems.

However, there is another way to track down misconceptions, or systematic difficulties. Different fact retrieval models, such as the associations model, the network interference model and network retrieval model (see section 2.4.1) have shown that during the retrieval of information in mental arithmetic, in addition to the mental representation of the correct answer being activated (e.g., $3 \times 2 = 6$), mental representations of incorrect answers may also be activated and compete with the correct answer during the decision-making process and response selection (e.g. $3 \times 2 = 5$; Ashcraft & Battaglia, 1978; Campbell, 1987; Domahs et al., 2006; Siegler & Shrager, 1984). Another example of such a wrong association is $8 \times 4= 36$, where a multiplicand is added (8 × 5), or 8 × 4 = 33, with a small addition error of + 1 (Siegler, 1988). In general, the fact retrieval models have argued that the more difficult problems have a greater distribution of associative strengths to wrong associations. In conclusion, although a correct answer is eventually given, the student might still contemplate these incorrect answers, possibly associated with consistent misconceptions or consistent confusions.

4.1.2 *Different ways to track arithmetic cognitive processes*

Measures of neural activity (e.g., electroencephalography) or eye movements (eye tracking) have provided insights into the dynamics of the learner's cognitive processes in mathematics (see chapter 3; and i.e. Artemenko et al., 2019; Hinault & Lemaire, 2016; Huebner & LeFevre, 2018; M. Lai et al., 2013; Spüler et al., 2016). However, these measures are also laborious and expensive and therefore difficult to scale to large samples outside research laboratories, such as an online learning environment. An emerging addition to these methods is mouse tracking, which is also thought to provide insight into the decision-making process (see section 2.3). Due to its many practical advantages (explained in section 2.3.1), mouse tracking is a great additional tool to record online mental processing in young children in their natural environment (S. E. Anderson et al., 2011).

In section 2.3, mouse experiments in the social sciences in general were described. Studies have also employed hand or mouse trajectories during mental arithmetic to study numerical processing in a variety of ways (Dotan & Dehaene, 2013; Faulkenberry et al., 2018; Fischer & Hartmann, 2014; Marghetis et al., 2014; Santens et al., 2011). For example, Marghetis et al. (Marghetis et al., 2014) presented their participants with addition and subtraction problems (e.g., 6 +2) and asked them to move the mouse cursor from the lower centre to one of the top corners of the screen with an answer option (e.g., 8 and 9). The authors found a systematic leftward deflection of the mouse for subtraction problems and a rightward deflection for addition problems, irrespective of where the correct response was. Another example is a study by Faulkenberry et al. (2014; 2016), who examined mouse movements during a numerical comparison task. In this task, the numerical distance between two digits needed to be judged while ignoring the physical size of the digits. This study showed that in incongruent trials - when these two variables differed (e.g., 2 8) - the mouse path was curving more towards the incorrect response than in congruent trials (e.g., 2 8). The greater attraction towards the incorrect response due to size congruity interference is thought to reflect response competition (see section 2.3). In these studies, mouse-tracking allowed an examination of the strength of the attraction towards incorrect options, without the need for an error to be made.

4.1.3 *The current study*

In the current study, we implemented mouse tracking outside the laboratory. Specifically, mouse movements of primary school children were tracked while doing arithmetic exercises in the online adaptive practice environment of Learn (Straatemeier, 2014 section 2.1). Previous mouse tracking studies in mathematic tasks focused on the incongruency effect, either with the physical location on the screen or the size of the number, but this was not of interest in this study. The aim of this study was twofold. First, the study serves as a validation that mouse movements can reflect the competition, at the cognitive level, between multiple answers during the resolution of arithmetic problems, in an ecologically valid setting (e.g., sitting in their bedroom, or living room, working on their own computer). Critically, we hypothesized that the extent of mouse movements towards non-selected incorrect options would relate positively to how frequently these errors are made in a speedy arithmetic multiple-choice game. This hypothesis is based on the assumption that, if mouse movement underlies mental processing and deviates towards distracting options (see section 4.1.2 and 2.3); and children are thinking of the representation of incorrect responses during retrieval of mental arithmetic (see section 4.1.1), mouse trajectories should deviate towards frequently made errors. Second, we examined whether we could detect systematic patterns of mouse attraction towards certain errors, that were made in the past, at the individual level. I will also explore whether mouse tracking may reflect common errors across age and difficulty levels. The ultimate aim of this study is to develop a measure of attraction towards errors that would reveal underlying systematic difficulties.

For this measure of attraction, we developed a mouse tracking method to analyse mouse movement in complex multiple-choice tasks, such as the games in Learn. In the current study, we have designed a method with five incorrect response options (in contrast to previous studies, see section 2.3.2) where attraction towards multiple answer options is allowed. During the process of solving a mathematical problem, individual children can have different error-related associations. Second, one child can also have multiple error-related associations. By presenting more than one incorrect option, we can track the whole problem-solving process, where a variation of error-related associations is possible.

4.2 METHODS

4.2.1 *Participants*

For this study, 90,000 children, aged between 5 and 13 years old (*M* = 10.2 y), were randomly selected from a pool of users playing actively in Learn (*N* = 180,000 users). For this selection, participants had to have logged in to the environment in the last three months before data collection, to increase the chance that these children would play the game during the six weeks of data collection. Not all students were selected to be tracked through mouse movements to limit the load on the database. After recording data for a total of six weeks, 1,590 different users (*M* = 10.3 y; *SD* = 1.46; 46% female) had played the selected arithmetic problems in our task and their mouse trajectories were used for analyses. This seems like a small sample in comparison to the designated pool, but there are thousands of items to play in this game, and children are not obliged to play the particular game. As can be seen in **Table 4.1**, users were predominantly eight years of age and older, since the selected problems required basic knowledge of, and practice with, all mathematical operations.

Table 4.1. *Distribution of age and gender in the sample.*

Age	Proportion of the sample (%)	Female %
5 - 7 years	3.3	43%
8 years	10.8	43.1
9 years	15.0	44.4
10 years	23.1	44.8
11 years	23.4	46.7
12 years	20.7	48.4
13 years	3.7	50.0

4.2.2 *Arithmetic speed mix game*

Speed mix, one of the 24 games available in the Math Garden platform of Learn, was used for the current study. This is the same game as the one that was adapted for the lab study in Chapter 3. In each game session, ten problems with a mix of four different operations (i.e., addition, subtraction, multiplication and division) are presented. In the game, students are asked to click one of the six answer options within eight seconds, following the high-speed high stakes rule (see section 2.1.5). This task was chosen for this study because (1) the students practice core arithmetic skills, basic tools essential for solving more complex maths problems;

(2) it has a multiple-choice design instead of giving a response through a keypad so that the mouse trajectory towards the different multiple-choice response options can be investigated; (3) the students are under time pressure to answer (i.e., eight seconds), which promotes movement of the mouse before reaching a decision (Kieslich & Henninger, 2017; Scherbaum & Kieslich, 2018).

The original speed mix game design was adapted for this mouse tracking study. At the start of each problem, a start screen was added, with a blue button in the middle that needed to be clicked before the arithmetic problem and answer options are shown. This ensured that every mouse trajectory started in the same position and at an equal distance from all the answer options (**Fig. 4.1**). The answer options for a given arithmetic problem were the same for every participant but were randomly placed across the six location boxes every time the problem was presented to a participant. A trial ended with the participant either clicking on one of the answer options, or on the question mark (**Fig. 4.1**) to skip the trial (i.e., when the student does not know the answer) or when the time limit was exceeded.

The speed mix game contains over a thousand different problems. To reduce the server storage load and simplify the analyses, we selected 36 problems to be tracked. Nine problems were selected per operand use (i.e., addition, subtraction, multiplication and division). The problems were chosen based on the frequency of errors made for each problem and the problem difficulty, see **Table 4.2**. The problem difficulty stemmed from the Elo-rating, which was averaged over two years of data collection within the speed mix game before the start of this study. Similarly, the frequency of the errors was calculated as the proportion of times a particular incorrect response was given to a problem over the course of two years in the speed mix game before starting the study.

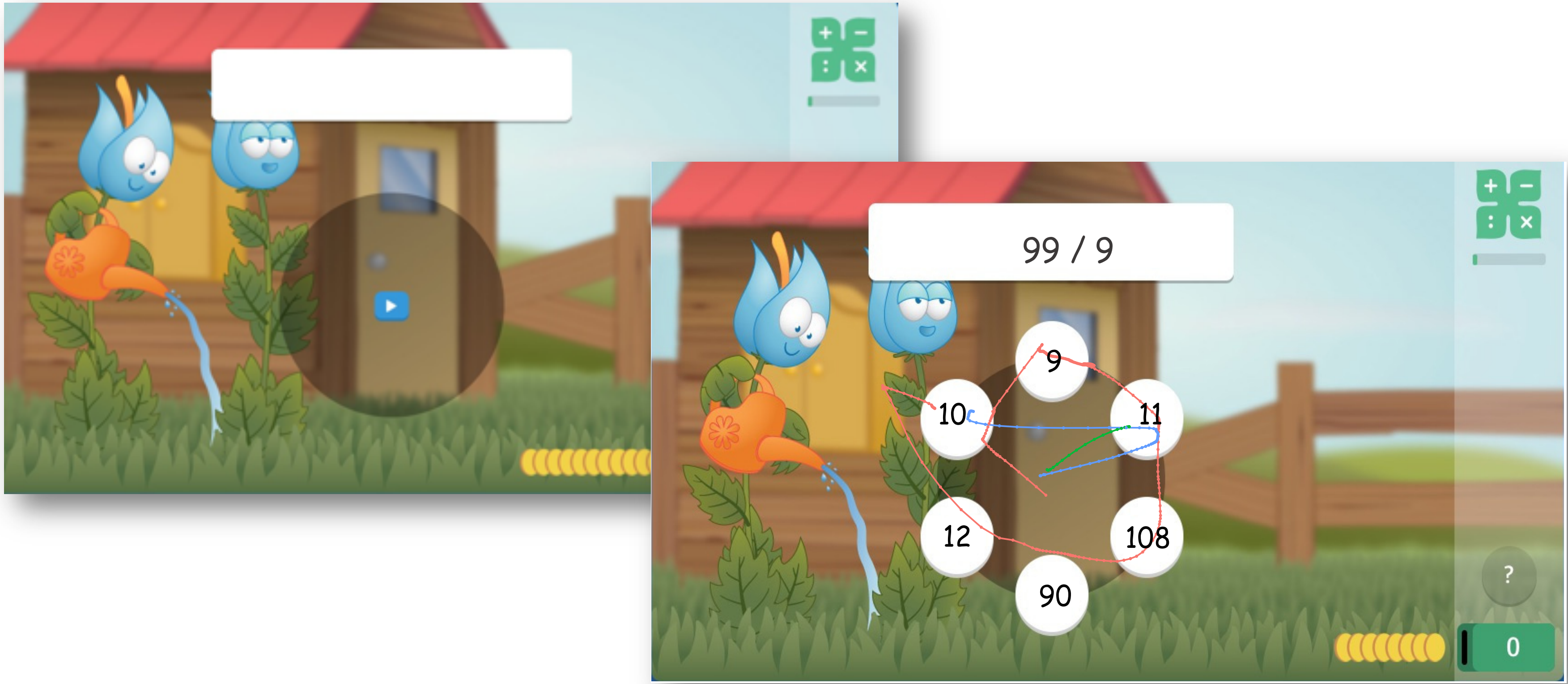


Figure 4.1. *Design of the arithmetic speed mix task| Left: Start screen shown before every problem administration. Students needed to click the blue button in the middle to reveal the problem. Right: Example arithmetic problem: "99/9". The students clicked one of six answer options (40 pixels) including one correct and five incorrect options. The answer options were placed randomly in one of six locations every time the problem was administered. The question mark on the bottom right can be clicked when the student does not know the answer to the problem. Also, three examples of mouse trajectories are shown.*

Table 4.2. *Description of a subset (12/36) of the arithmetic problems tracked in this study. These values were calculated using data from all the users that played the Speed Mix game in the last two years before the start of this study. Difficulty level, the overall number of errors and frequency of errors are provided for each problem, as well as the frequency of correct answers and the five most frequent errors. The arithmetic problems and the corresponding answer option choices are highlighted in bold.*

Arithmetic problems	Difficulty level [-15 15]	Number of errors % of trials	Correct answer % of trials	Error 1 % of trials	Error 2 % of trials	Error 3 % of trials	Error 4 % of trials	Error 5 % of trials
5 + 508	0.8	**93406**	**513**	**512**	**503**	**515**	**523**	**558**
		28%	78.5%	13.5%	3.3%	2.7%	1.4%	0.6%
7 + 80	-5.4	**74775**	**87**	**78**	**88**	**807**	**150**	**15**
		26%	86.8%	10.3%	1.1%	0.9%	0.5%	0.4%
55 + 66	4.6	**33276**	**121**	**111**	**131**	**120**	**116**	**115**
		27%	64.9%	28.2%	3.1%	1.5%	1.3%	1%
9 – 7	-5.9	**89467**	**2**	**3**	**1**	**4**	**16**	**0**
		25%	84.5%	8.8%	2.6%	2.3%	1.2%	0.6%
14 – 5	-5.2	**90022**	**9**	**11**	**10**	**19**	**8**	**7**
		27%	83.4%	9.4%	2.3%	2%	1.8%	1.1%
55 – 8	0.9	**91703**	**47**	**46**	**57**	**17**	**53**	**50**
		26%	80.2%	10.9%	5.1%	1.3%	1.9%	0.6%
5 x 5	-6.4	**69766**	**25**	**10**	**35**	**20**	**30**	**24**
		25%	82.9%	6.8%	4.1%	3.6%	2%	0.6%

7 x 9	-1.0	**76903**	**63**	**64**	**72**	**54**	**70**	**16**
		29%	76.5%	10.2%	6.7%	5.1%	0.9%	0.6%
50 x 12	2.4	**60520**	**600**	**550**	**580**	**60**	**500**	**58**
		26%	77.3%	9.6%	5%	4.1%	3.2%	0.8%
18 / 2	-1.8	**78647**	**9**	**6**	**8**	**7**	**10**	**20**
		28%	71.6%	17.4%	8.2%	1.6%	0.6%	0.6%
36 / 4	-0.6	**92387**	**9**	**8**	**7**	**32**	**40**	**10**
		27%	74.3%	16.3%	6.6%	1.2%	0.8%	0.8%
99 / 9	-2.7	**83306**	**11**	**10**	**9**	**90**	**12**	**108**
		29%	67.9%	14%	11.1%	4.8%	1.3%	0.9%

4.2.3 *Mouse tracking*

The Learn team built a mouse tracking tool in JavaScript for this study and adapted the design of the speed mix game in the learning platform. On every administered problem, the x- and y coordinates of the mouse, the timestamp and the locations of the answer options were recorded at a sample rate of approximately 35 Hz (35 measurements per second). This was stored on the data server of Learn.

4.2.4 *Procedure*

For six weeks, participants were tracked when playing the speed mix game. Students played in their natural environment independently, either in- or outside school. No teachers or parents were involved in this study. Schools and families with accounts are informed that Learn collects diverse categories of data, such as mouse tracking data, for research purposes. Children (their parents or schools) can opt out of being part of the research done and are therefore not included in this study. All data were anonymized before analysis. This procedure was approved by the University of Amsterdam's department's Ethics Review Board. Learn can be played on different types of devices, including touchscreens and tablets. Sessions tracked in this study that were played using touchscreen devices were excluded from analyses (11%). Since Learn is a child-directed setting, children could decide for themselves whether they would play the speed mix game and for how long they would play. This means that during our data collection, not every child has seen (or answered) all the possible problems; some children may have only seen one problem, and some children may have performed the same problem multiple times.

4.2.5 *Measures*

4.2.5.1 Response Errors

First, the frequency of the response errors was calculated. To determine in a stable and robust way which errors are made most often, the responses administered for the last two years by all users in Math Garden were examined for the chosen 36 problems (see **Table 4.2**). From the N = 607,125 collected responses in the speed mix game, 70% were answered correctly and less than 1% were not answered on time. In the remaining N = 128,000 the student chose one of the five incorrect responses. The proportion of trials where children selected one of the incorrect responses, was calculated across all users. For example, the incorrect answer 512 was given 13.5% of the time for the problem 5 + 508, which was the most frequent error children made (**Table 4.2**). These response values were used to compute associations with mouse movement attraction. For the individual mapping analyses, the proportion of error occurrence per incorrect response was calculated for every individual participant instead of across participants.

4.2.5.2 Mouse movements

The number of trajectories collected per participant across all problems in this study varied from one to 48, including repetitions of the same administered problems (M = 2.89, SD = 3.82). This average is low since the chance is also low that a participant comes across multiple problems of the 36 selected problems within six weeks, out of a pool of thousands of items. In six weeks, a total of 6,443 trajectories were collected. First, trajectories with more than three mouse locations (N = 5,653) were selected, to remove trajectories with minimal movement. Second, trajectories with uncontrolled movements, (i.e., mouse locations 150 pixels away from the answer options or question box) were excluded (remaining N = 5,269 trajectories). Third, only the mouse trajectories of the correctly answered problems were analysed, N = 3,906 trajectories (of N = 1,590 users). The reason for this was twofold: (1) The correctly answered problems are in principle not informative of errors when only analysing reaction time and/or accuracy at a behavioural level; (2) incorrectly answered problems contain actual mouse clicks on the incorrect answer options and would therefore bias attraction towards incorrect answer options.

The trajectories were first smoothed with a 10-point smoothing window using the Savitzky-Golay filter from the “trajr” package (McLean & Skowron Volponi, 2018) in R (Team, 2013), which is particularly suitable to preserve the shape of the trajectory while removing high-frequency squiggles.

To analyse a complex design with five incorrect options, I combined for each trajectory two measures of mouse attraction. The first method is similar to what was done in chapter 2. Here, a static method was used, where the number of points inside each answer box (radius of 40 pixels) was counted, in line with how eye movements are typically classified (Schulte-Mecklenbeck et al., 2019), see **Fig. 4.2 A** (see Appendix section A.2 for more details). Since it was shown in Chapter 2 that mouse movement rarely end inside an AOI, another method was added. This is a dynamic method, where the mouse directions were classified as a movement towards an answer box, from one mouse location to the other, visualised in **Fig. 4.2 B** (see Appendix section A.2). The dynamic method is inspired from animal movement analyses (see for example Michelot et al., 2019; Patterson et al., 2009), where the movement of an animal (in our case, the computer mouse) towards potential predators (in our case, the answer options) is analysed. The advantage of adding the dynamic method is that it detects movement towards an attractor, regardless of where the mouse is located at that point. For every trajectory, the findings of these methods were combined to ensure that the competition between the choices could be analysed from the start of the trajectory to the end. This was done by adding the number of movements from the dynamic method that were not inside an answer box, to the mouse locations obtained from the static method. The mouse locations and movements that could not be associated with any of the incorrect answer options were excluded from the analyses.

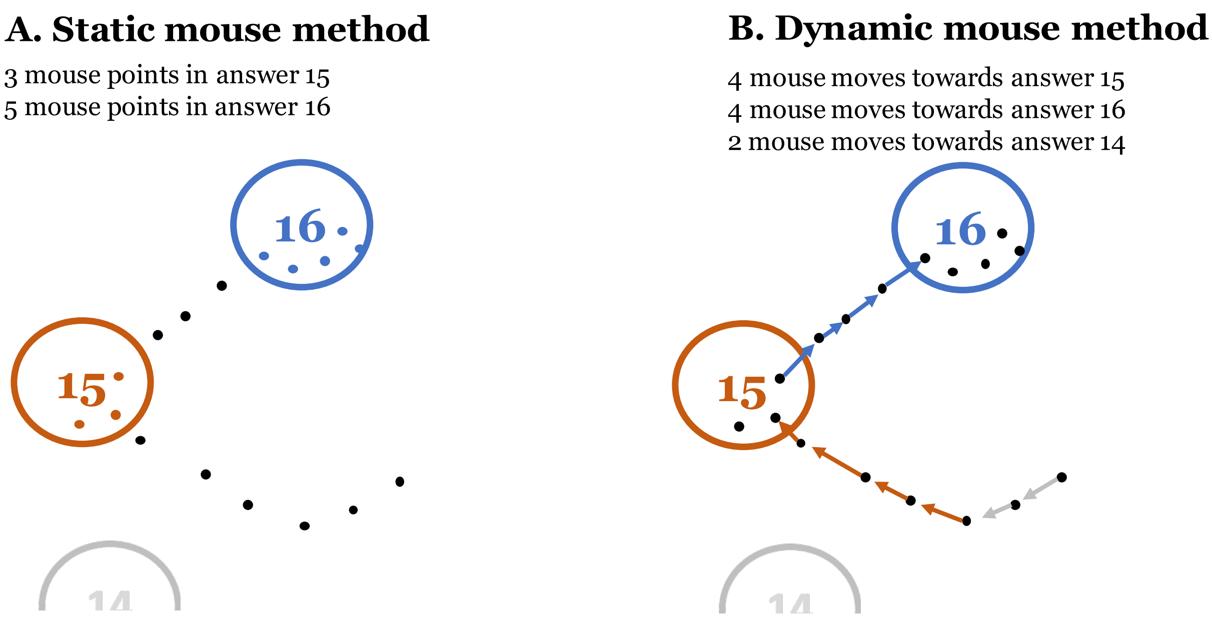


Figure 4.2. *Two measures of mouse attraction. In the static method **(A)** the mouse locations in the answer option boxes (40 pixels) are counted, e,g, three mouse locations in the answer box 15. The dynamic mouse method **(B)** counts the mouse movements towards the answer options, that is four mouse moves towards answer 15.*

4.2.6 *Analyses*

The main analysis focused on whether there were mouse movements towards incorrect answer options before the correct answer was chosen. More specifically, on a group level it was investigated whether the percentage of mouse movement towards errors correlated with the frequency of these error rates. Since these correlations were calculated per arithmetic problem, the next step was to calculate the correlation between the difficulty of the arithmetic problem with the size of the mouse mapping correlation. In addition, the correlation between the average age of the participants and the size of the correlation was calculated.

Apart from the group level, the individual mapping of mouse movements on false associations is of interest. Therefore, the correlation between mouse movement, averaged over all problems, and past error response rates of the individual was calculated.

4.3 RESULTS

4.3.1 *Response accuracy*

Of all responses, 74% were answered correctly. This meant that 26% of the trials and their mouse trajectories could not be examined, because only the trajectories with a final correct response were examined. The students answered correctly in 4.1 seconds on average (*SD* = 1.6 s), where the maximum time to respond is eight seconds.

4.3.2 *Mouse trajectories*

An average of 115.9 mouse points (*SD* = 107.6) were collected per mouse trajectory, at a frequency of 35 Hz. Of all mouse points, 22% were associated with the correct answer option, where 14% were inside the correct answer box and 7% were directed towards the correct answer. Of the mouse points, 47% could not be directly associated with any of the answer options and 31% were associated with one of the five incorrect answer options - 16% was inside and 15% towards the incorrect answer option boxes. Of this 31%, the proportion of mouse points associated with each incorrect answer option was averaged across the participants for each arithmetic problem. An example of three different types of trajectories from different children for the same problem can be found in **Fig. 4.1**: Movement straight towards the error response (green line); movement towards an incorrect option (blue line); a clockwise movement (red line).

4.3.3 *Mouse attraction to incorrect responses at the group level*

For the main analysis, we investigated whether the frequency of the response errors (in the past two years by all users of Learn) would also attract the greatest number of mouse movements in correctly answered trials collected in this study. A Pearson correlation between the response error rate and the number of mouse movements was calculated for each arithmetic problem (**Fig. 4.3 A**). These Pearson *r* correlations were transformed using a

Fisher Z transformation ($z = 0.5 \times \ln((1 + r) / (1 - r))$) for further analyses. Findings revealed that the correlations ranged across the arithmetic problems from .45 to .70 with an average of $M = .66$ ($\bar{r} = 0.57$), which was significantly different from zero, $t(35) = 42.33$, $p < 0.001$. These high correlations mean that there was more mouse movement towards the errors that are frequently made than towards less frequent errors, in line with what was hypothesized. A linear regression showed that these correlations varied as a function of the problem difficulty (based on the estimated Elo rating), where a higher correlation between the error rate and mouse movements was found when the problems were relatively easier, $\beta = -0.56$, $t(34) = 3.54$, $p = 0.001$ (**Fig. 4.3 C**). The average age of the users also associated positively with the size of the correlation, $\beta = -0.61$, $t(34) = 4.47$, $p < 0.001$ (**Fig. 4.3 B**). The problem difficulty and average age of the user also correlated highly with each other, $r = 0.94$, causing multicollinearity when analysed in the same model. This seems to suggest that mouse movements mapped better on the underlying arithmetic difficulties of younger children, who on average play somewhat easier problems.

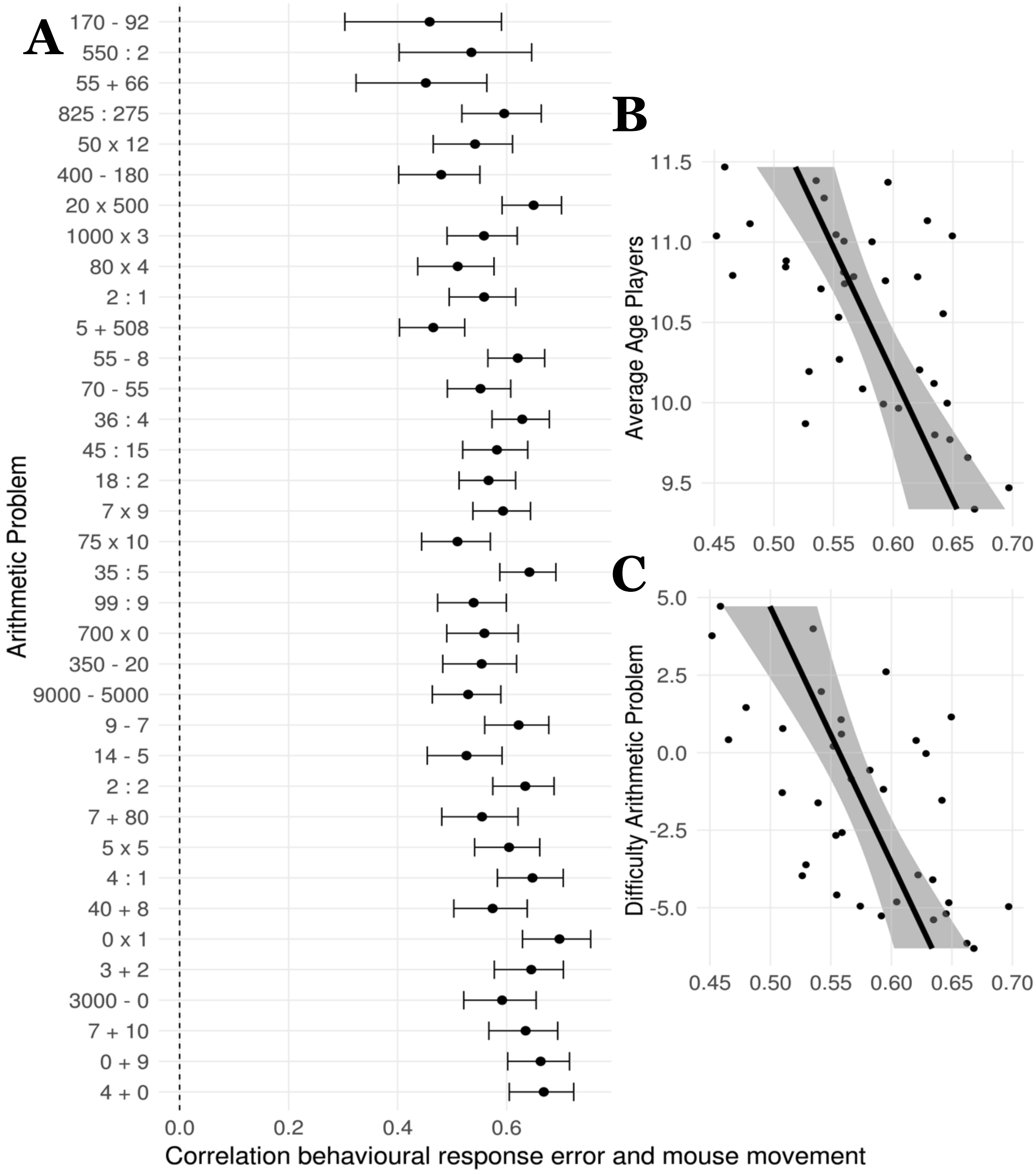


Figure 4.3. *Correlations between the proportion mouse movement towards the incorrect answer options in correctly answered trials and the corresponding error rate:* ***(A)*** *per arithmetic problem (ordered from high to low problem difficulty). The error bars show the 95% confidence interval of the correlation test.* ***(B)*** *Linear regression line showing that the average age of the players associated negatively with the size of the correlation;* ***(C)*** *Linear regression line showing that problem difficulty also associated negatively with the size of the correlation. The 95% confidence intervals are displayed as the shaded area around the mean for both lines.*

4.3.4 *Individual mapping of mouse movement on errors made in the past*

In the analyses above, error response rates were calculated based on all users that previously played the speed mix game, to ensure reliable error rate rankings. Next, only the error response rates of a selection of the participant pool in this study were calculated. The aim was to test whether the participants who made certain response errors in the past would also show mouse trajectories towards these specific incorrect responses, even when the final response was correct.

To investigate this, a Pearson correlation per participant, averaged over all problems, between their mouse movement and past error response rate was calculated. Correlations with missing values (either no history of previously made errors or mouse trajectories) were removed as well as participants that contained a correlation for only one arithmetic problem. This caused a big shrinkage in sample size for the next analyses, with a total of $N = 101$ participants. The Fisher Z transformed Pearson r correlations for these individuals were analysed with a t-test (**Fig. 4.4**). This test showed that the average correlation was significantly higher than zero, Fisher Z transformed $M = .42$ ($\bar{r} = 0.23$), $t\,(100) = 4.23$, $p < 0.001$. This means that participants showed more mouse movement towards their previously made response errors than to other incorrect options.

Furthermore, some checks were made to ensure that these findings were robust. Firstly, the average correlation per participant over all problems was weighted by how many errors the participant had made in their history of playing the speed mix game before data collection. It was assumed that problems with more error responses are more likely to reflect a consistent conceptual or procedural difficulty (e.g., a tendency to confuse multiplication and addition; see Appendix section A.3). The second check was to weigh the correlation by how many mouse trajectories were collected from the participant for a particular problem (Appendix section A.3). When a participant had not given an incorrect response to a particular problem within a year from when the mouse trajectory was registered, the data were excluded since the participant had presumably mastered the problem and would present no difficulty. The checks show stable correlations irrespective of how much data were collected and when the errors were made.

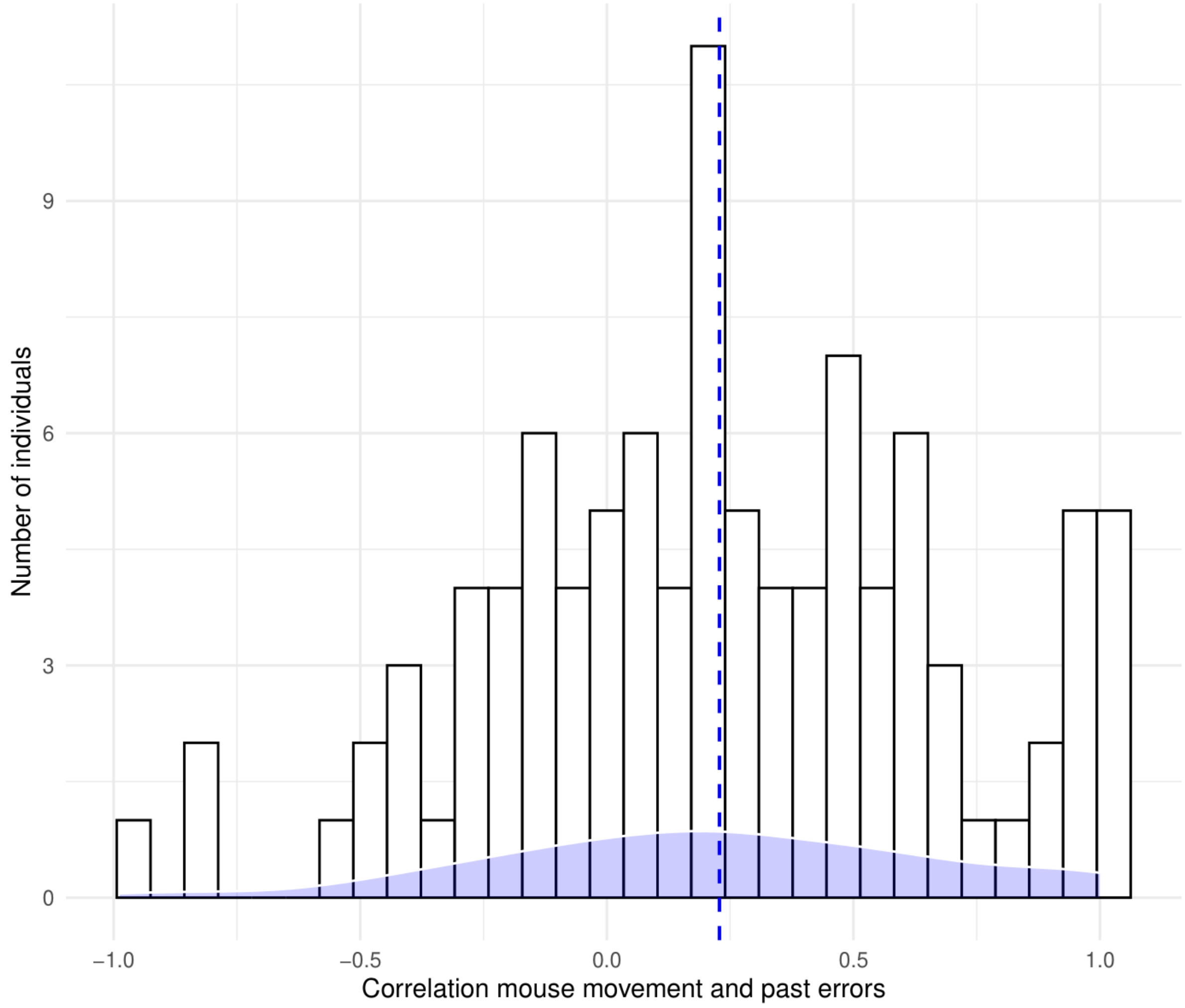


Figure 4.4. *Histogram showing the individual correlations between mouse movement and response errors made in the past. The blue dashed line shows the average correlation (r = .23). The blue shaded area represents the distribution of the correlations.*

4.4 DISCUSSION

The aim of this study was to investigate whether it is possible to detect difficulties children encounter when solving arithmetic problems, without relying on the errors themselves. This was done by tracking mouse movements, a method intended to measure the competition between responses during the process of problem-solving. This study is, to our knowledge, the first to implement mouse tracking outside the lab in an online adaptive learning system. We analysed the mouse movements towards multiple objects, which were in this case the incorrect response options of multiple-choice arithmetic problems. Our findings showed that, even when the final response was correct, the mouse trajectories revealed attraction towards the errors most frequently made by students playing in Learn. All ages and arithmetic problems showed this high attraction, but the mapping of the mouse was stronger for younger children and easy arithmetic problems. Furthermore, we found that individuals who had made certain errors in the past (past two years before data collection) still showed more mouse movement towards these errors than to other errors.

The results from this mouse tracking study show that children move towards wrong associations before the correct answer, which can be expected following the theory of retrieval of simple arithmetic facts, such as the network interference model (Campbell, 1995). The results also showed a stronger mapping of mouse movement to interfering answers for younger children and easy problems. Given that children are still developing their arithmetic network and facts, the associative strengths of the problem with wrong associations are thought to be stronger and more divided for younger children, who on average play the easier problems in the learning environment.

Thus, implementing mouse tracking in an online learning environment allows us to study whether students have systematic reasoning problems in solving arithmetic problems. The information on systematic difficulties can enable us to give targeted feedback and instruction on the learning process, after both an incorrect and correct response. To give an example: if the method detects that a student is consistently attracted to an answer corresponding to mistaking addition for multiplication, it will greatly benefit the learning process to present a reminder to carefully check the operand, and a series of problems where the operations are constantly mixed, to practise switching between operations.

The strength of investigating the use of mouse tracking in an online learning environment, instead of the lab study in Chapter 3, is that the sample size is both large and heterogeneous in terms of math abilities, age and SES background. Second, the students practice in their natural environment, either at school or at home. The limitless access to Learn for students in diverse circumstances ensures that our results are robust and would not only be reproducible in a lab-controlled experiment. This said, in a future implementation mouse tracking could be done daily for thousands of different maths problems at the same time, throughout the children's primary school period.

Looking for error patterns is not a new research line. Other studies have used the large scale Math Garden response database to categorise and detect systematic errors (Savi et al., 2018; Straatemeier, 2014), but they had to rely on the incorrect responses a child picked when they made an error. Since online adaptive environments require a high success rate for motivational reasons, errors are rare; in this study an incorrect response was given for 20% of the N = 6,443 problems administered. Our findings show that mouse movement can also reveal information about these errors in the other 80% of the responses, which would help with the diagnosis of systematic difficulties at an individual level.

There are also some limitations to this study. First, collecting data in such a naturalistic setting can cause the data to be noisy. There is no way to control the circumstances within which the child is performing the task. A general problem with doing a study remotely is that we can't know for sure who is manifesting the behaviour we are basing the behavioural metric on. For example for this study, we might have included the mouse trajectory of a parent or teacher that is demonstrating how to solve the problem. Other circumstances that could have impacted our data quality, is that children are distracted while playing the game or have a bad internet connection. Equipment is not standardized since the mouse used is different for every school and home. It is therefore necessary to collect a large dataset, such as in this study, to be very strict in terms of removing noisy mouse trajectories and unfinished game sessions. Second, many children practice their skills in online learning environments on tablets and touchscreens. Mouse tracking cannot be used for such devices. A third limitation is that mouse tracking studies are bound to a specific design and do not apply to all games and experimental tasks. Ideally, every trial needs to have the same starting point and multiple-choice options located at an equal distance from the starting point.

For further research, it would be interesting to also consider less known mouse metrics such as initiation time, to reflect difficulty with a certain problem. Initiation time stems from eye tracking studies but in the context of mouse tracking it reflects the time elapsed between clicking START and the first mouse movement (Freeman & Ambady, 2010; Incera & McLennan, 2016). In the context of error processing, it would be interesting to investigate whether mouse initiation strategies differ depending on how much individuals are thinking about alternative answer options. Another line of research is to investigate individual cases with both eye and mouse tracking in a laboratory setting to see how these methods complement each other with regard to signalling the attractiveness towards alternative options in the context of mathematical problem-solving. In Chapter 2, I showed a comparison between eye fixations and mouse points inside an AOI. With the newly developed method presented in this chapter, it would be interesting to repeat the comparison with gaze trajectories in the lab study from chapter 2 and see whether the attraction towards error options is more similar. Moreover, in the study presented in chapter 3 we saw that the visibility of time pressure affected the number of eye fixations at the response options. A time countdown would be a valuable manipulation to look into when comparing eye and mouse tracking

5 ADAPT TO IMPROVE: POST-ERROR SLOWING

ABSTRACT

The ability to monitor and adjust our performance is crucial for adaptive behaviour, a key component of human cognitive control. One widely studied metric of this behaviour is post-error slowing (PES), i.e., the finding that humans tend to slow down their performance after making an error (Rabbitt & Rodgers, 1977). This study aims to generalise the effect of PES to the online adaptive learning environment of Learn. Around eight million response patterns were collected from 150,000 users aged 5 to 13 years old over the course of six months, across 23 different learning activities related to mathematical and language skills. The findings show that PES could be observed in most learning activities, but the extent of PES varied as a function of several variables. At the task level, PES was greater when there was less time pressure and slower errors and in learning activities focusing on mathematical skills rather than language skills. At the individual level, students that chose the most difficult level for the game also showed the greatest PES. This effect varied as a function of age and ability: 7 to 9-year-old children, and more able children, showed a larger PES. This study is the first to show that PES not only occurs in simple repetitive reaction time tasks performed in a laboratory setting but also in educational tasks in an online learning environment.

The study of this chapter is in press:
Susanne M.M. de Mooij, Iroise Dumontheil, Natasha Z. Kirkham, Maartje E.J. Raijmakers, Han L.J. van der Maas (2021). Post-Error Slowing: Large scale study in an online learning environment for practicing mathematics and language, *Developmental Science,* https://osf.io/wjvms/

5.1 INTRODUCTION

Mistakes can be detrimental to the motivation of students. But, making mistakes is also part of the learning process, allowing a student to learn and improve (Moser et al., 2011). The capacity to detect errors through feedback was discussed in the previous chapter. But evaluating and adapting to these mistakes is equally important to the development of children and has been an active area of research in cognitive control literature (Regev & Meiran, 2014; Smulders, Soetens, & van der Molen, 2016; Tamnes, Walhovd, Torstveit, Sells, & Fjell, 2013; Ullsperger, Danielmeier, & Jocham, 2014, see section 1.7). One of the studied markers of adaptive behaviour is post-error slowing (PES). PES refers to the finding that humans slow down their performance after making an error, such that the RT after an error is greater than after a correct response (Laming, 1979; Rabbitt & Rodgers, 1977). Whilst prior studies have typically examined PES within experimental settings using executive function tasks, the present study examined PES outside the laboratory in a range of tasks in the online learning environment of Learn.

5.1.1 *Theoretical accounts of PES*

Two main theoretical accounts of the functionality of PES have been proposed in the literature. The functional account argues that PES serves the purpose of taking more time to plan an action to prevent future errors and increase performance. This is supported by the conflict monitoring theory (Botvinick et al., 2001), which states that slowing down reflects the regulation of cognitive control mechanisms evoked by a conflict, i.e., an error. People monitor their performance and consequently adjust their response thresholds, leading to slower but more accurate responses. This was also shown by Dutilh et al. (2012)'s drift-diffusion model, where participants increased their boundary separation, the model parameter for response caution, after making an erroneous decision. An alternative explanation supporting the functional account is considering PES as the product of automatic inhibition of the next response after an error (Gupta et al., 2009; Marco-Pallarés et al., 2008), where inhibitory mechanisms increase response cautiousness after making an error.

Other studies have suggested a non-functional account of PES, supposing that errors have a negative effect on subsequent performance. The orienting account argues that PES occurs especially with infrequent errors since the attention orients towards this surprise effect instead of the task (Houtman et al., 2012; Notebaert et al., 2009). This attention shift disrupts the information process, inducing slower response times and worse performance. Notably, Danielmeier and Ullsperger (2011) point out in their review that there is evidence for both the functional and non-functional account and that they are not necessarily mutually exclusive, as multiple mechanisms may contribute to the slowing down of performance.

Although PES effects have been presented as markers of adaptive behaviour, the majority of the studies are done with relatively simple and rapid tasks such as the flanker task (e.g., Schroder et al., 2019), or the Go-No/Go task (e.g., Jonker et al., 2013). These tasks are limited in terms of the adaptation strategies they allow and performance only be increased by paying more attention to the stimuli, in line with Notebaert's (2009) conflict monitoring theory. Some studies asked university students to perform more complex academic tasks, such as mental arithmetic tasks, where multiple strategies are available (Borght et al., 2016; Desmet et al., 2012; Lavro, Levin, et al., 2018; Núñez-peña et al., 2017). The findings of these studies indicate larger response times after an error than after a correct response. Interestingly, Borght et al. (2016) found that switching to a different strategy improved performance on subsequent trials in terms of accuracy and RT, therefore reducing PES.

5.1.2 *Development of PES*

Education is a plausible naturalistic setting for studying the richness of adaptive behaviour to errors. The development of error monitoring as measured by PES has mostly been examined in combination with neuroscientific methods, e.g., the ERN with electroencephalography (section 1.6.4), with children between 5 and 12 years of age (for reviews, see Ferdinand & Kray, 2014; Tamnes et al., 2013). Some behavioural studies such as Fairweather (1978) found that young children from the age of 5 slowed down after making errors during a two- to eight-choice RT task. Ever since, PES in children has been repeatedly examined and observed (Berwid et al., 2014; Fairweather, 1978; Gupta et al., 2009; Jones et al., 2003; Schachar et al., 2004; Smulders et al., 2016).

Following the orienting account where errors lead to interference, children are thought to be more prone to interference than young adults and therefore also exhibit more PES (Smulders et al., 2016; van der Molen, 2000). Conflict arising from interference is also related to the still-developing IC skills in children (see Chapter 3 and section 1.5.3). This indirect relation implies that improving inhibitory control skills would therefore lead to a reduction in PES magnitude. On the other hand, greater PES may also reflect a developmental increase in cognitive control, with greater performance monitoring and strategic adjustment of behaviour, in line with the functional account (Jones et al., 2003; Tamnes et al., 2013).

The directionality of changes in PES during development observed in the literature is heterogeneous. In recent research, Soetens and van der Molen (2016) used standard two-choice RT tasks and found that post-error response slowing was present from 5 years to adulthood, which was the whole of their examined age range. They reported PES stability with age rather than developmental differences. However, Jones, Rothbart and Posner (2003) reported increasing PES from 3- to 4-year olds using a Simple Simon task, interpreted as a developmental increase in cognitive control. The results of Gupta, Kar and Srinivasan (2009) showed a non-gradual development of PES in 6-11-year-old children performing two task-switching digit tasks, with a peak in the magnitude of PES at age 7. Schachar et al. (2004) also found a decreasing PES magnitude from 7 to 16 years of age in a stop-signal task. Due to the heterogeneity in these findings, the development of PES with age should be examined in more detail.

5.1.3 *Factors influencing the presence or magnitude of PES*

Prior research has also shown that PES varies as a function of the type of task as well as the difficulty of the task. Although PES has been found in both easy tasks (i.e., flanker task, Schroder et al., 2019) and more complex tasks (i.e., mental arithmetic task, Desmet et al., 2012), the non-functional account noted that response slowing after an error occurs mostly under conditions when errors are rather unexpected and infrequent events (Danielmeier & Ullsperger, 2011; Lavro, Ben-Shachar, et al., 2018; Notebaert et al., 2009).

Another factor influencing PES is the type of error that preceded it. Errors can be found to be systematically faster or slower than correct responses, with a range of underlying causes (Ratcliff & Rouder, 1998). Slow errors typically occur when accuracy is emphasized and the task is relatively difficult, whereas fast errors occur when responding is rushed. Damaso et al. (2020) categorised fast and slow responses as the 50% slowest and fastest responses of each participant. Results showed that PES mostly occurred after fast errors (Damaso et al., 2020). This finding explains that PES represents an increased response caution as an effective means to reduce speedy errors.

5.1.4 *The current study*

In the present study, we investigated students' post error performance in Learn. The first aim of this study was to assess the presence of PES in the various learning activities relating to different mathematical and language skills.

The second aim was to investigate predictors of the magnitude of PES. At the task level, we investigated whether the type of skill practised (mathematics vs. language) was associated with the magnitude of PES. Children can choose not only between different types of games but also the difficulty level of the games (hard, medium, easy; see section 2.1.6). This allowed us to test the following prediction, based on previous research (section 5.1.1; Danielmeier & Ullsperger, 2011; Lavro, Ben-Shachar, et al., 2018; Notebaert et al., 2009): Children who choose the easy level, where there is less chance of making errors, were more likely to show PES after an error than children playing the medium and hard level, as the errors are more unexpected with a high success rate.

Finally, we investigated the influence of the speed of responding on PES, based on Damaso's research (2020). We first assessed the effect of time pressure (i.e., how much time children were given to respond in each task) on PES magnitude. Based on Damaso et al's finding that PES mostly occurs with speed emphasized instruction, more PES was expected when there was more time pressure. Secondly, we distinguished fast and slow errors by categorising whether the RT was higher or lower than the median split of the correct trials before the error within a task, considering global fluctuations in skills and motivation of the participant. Based on Damaso's research it was predicted that a task with fast errors would lead to greater PES.

We also assessed individual differences between participants as potential predictors of PES. We first examined associations with age. As previous studies have shown regarding the development of PES, it was predicted that the PES magnitude would increase from the age of 5 to 7 but decrease with older children. Second, we investigated whether individual differences in PES may associate with children's ability to perform in the task, above and beyond age.

5.2 METHODS

5.2.1 *Participants*

The response data of 149,747 Dutch primary school children playing in the learning environment were collected. Their age was between 5 and 13 years old (M = 9.4 y, SD = 1.8 y, 48.6% female).

5.2.2 *Materials*

Data were collected in the learning environment of Math Garden and Language Sea (see section 2.1). In this study, 23 different learning activities were analysed (see Appendix A.1 for a description). Ten learning activities are focused on practising language skills in Dutch, such as reading, spelling and grammar. The other 13 learning activities relate to mathematical skills, such as counting, fractions and telling time. These are learning activities that are actively played in Learn and will therefore generate a lot of data to analyse. We chose to not focus on one particular learning activity as in the previous chapters but to use a variety of activities to investigate the robustness of PES. And this allowed us to investigate potential predictors of PES that are characteristic for the tasks.

5.2.3 *Procedure*

A total of 45 million trials were collected in a period of six months (June 2019 until November 2019) from 23 different learning activities. Students decide themselves which learning activity they participate in, and the minimum age required to practise these learning activities varies. As a consequence, the number of participants and the average age is different for each activity (see Appendix A.1 for details). Trials with an RT faster than 200 ms were regarded as guess responses and were therefore excluded. Trials, where no response was made within the time limit, were labelled as missing.

To ensure a reliable measure of post-error behaviour, RTs were selected from a specific pattern of response. In a game session, a sequence of four problems was selected when the accuracy pattern was 1-1-0-1 (1 = correct; 0 = incorrect), i.e., a minimum of two correct responses precedes an error and the trial after the error is correct. The additional correct pre-error response was added to the response sequences to ensure that a correct post-error response in a session could not at the same time be a correct pre-error response. In the collected 45 million trials, 8 million of such 1-1-0-1 sequences were found.

5.2.4 *Measures: Post error slowing quantification*

The majority of previous studies quantify the magnitude of PES as the difference in mean RT between trials following an error and trials following a correct response. As Dutilh et al. (2012) pointed out, this method can be confounded by the global fluctuations in ability and motivation during the task, since post-error responses are more likely to originate from the second half of a task where responses are inevitably slower due to motivation and tiredness. Dutilh's solution is to quantify PES as the average, across the selected sequences, of the difference between the RT after the error (RT_{E+1}) and the RT before the error (RT_{E-1}) (PES_{diff}):

$$\overline{PES_{diff}} = MRT_{E+1} - MRT_{E-1}$$

To ensure that our results do not fully depend on the choice of this absolute PES measure, we use two additional methods. First, we also report PES relative to the overall ability speed, calculated by dividing the RT difference with the average speed of the two trials, i.e., $\frac{1}{2}\ (MRT_{E+1} + MRT_{E-1})$. This ensures that individual differences in PES magnitude can be compared regardless of the overall speed/ability of the student. So the second method to quantify PES is as follows:

$$\widehat{PES_{rel}} = \frac{MRT_{E+1} - MRT_{E-1}}{\frac{1}{2}\ (MRT_{E+1} + MRT_{E-1})}$$

The third method is a robust way of measuring post-error behaviour overall, by quantifying PES as the number of sequences where the RT was larger after the error (RT_{E+1}) than before the error (RT_{E-1}), relative to the number of sequences (*N*). This is a measure that is not affected by the PES effect size in certain sequences, which can vary considerably in such a big dataset. The disadvantage of this robust method is its low power:

$$\widehat{PES_{robust}} = \frac{n(RT_{E+1} > RT_{E-1})}{N}$$

5.2.5 *Analyses*

The main analysis focused on whether PES could be found for different learning activities. To do this, a linear mixed model was performed in R (Team, 2013) using the package *lme4* (Bates et al., 2015) and *lmerTest* (Kuznetsova et al., 2017) to calculate p-values. Participant was treated as a random effect because the number of sequences collected per participant differed. In this main analysis, learning activity was treated as a fixed effect. When a general PES effect was found in the main analysis, we added a variety of predictors to the basic linear mixed model where both learning activity and participant were treated as random effects. A chi-square test was used to see if the added predictors explain PES better.

The first categorical predictor that was added was the type of learning activity, distinguishing between mathematical and language skills. The second task predictor was time pressure since tasks differed in the time limit given to participants to solve a problem, from 8 to 60 seconds (see Appendix A.1). The third categorical predictor, difficulty level (easy, medium, hard level), reflected the probability of answering correctly chosen by the child (see section 2.1.6). At the participant level, we also included age (5-13 years old) and child's ability level (a continuous scale from -10 to 10; see section 2.1.5) as predictors of PES magnitude. Both predictors were analysed in a linear and quadratic fashion since the study of Gupta et al. (2009) also found a non-linear development of PES with age. Lastly, the type of error was investigated at the trial level, making a comparison in the magnitude of PES between a fast and slow type of error. All errors were either categorised as a fast error when the RT was lower than the median RT of the correct items before the error as a slow error when the RT was higher than the median RT.

5.3 RESULTS

5.3.1 *Overall PES effect*

To measure whether there was PES in the learning activities, a linear mixed model was fitted to the data with learning activity as a fixed effect and participant as a random effect. Using the absolute PES_{diff} measure we found that in 20 of the 23 learning activities participants showed a significantly larger RT after an error than before, with a mean difference across learning activities of 225 ms (**Fig. 5.1** and **Table 5.1** for details on the statistics). The average age of participants differed considerably between the learning activities (8.2 - 10.8 years), but this did not account for the variation in PES magnitude between the learning activities, $\bar{r}$= -0.08, *p*= 0.70. For the PES_{rel} measure we found the same pattern, such that the learning activities *letterchaos* and *spelling* did not show PES but post error speeding (**Table 5.1**). The learning activity practising *grammar* was not significantly different from zero in PES_{diff}.

Since PES_{robust} is a proportion measure, we performed a separate proportion test for every learning activity. When there was a greater proportion than 0.5, it would mean that RT was in more than 50% of the time greater after an error than before. The results were in line with the linear mixed models (**Table 5.1**). The average proportion across learning activities was .52 [range .48 - .55], meaning that in 52% of the sequences the RT after an error was greater than before the error.

Table 5.1. *Statistical results of the linear mixed models for PESdiff and PESrel, Separate proportion tests were performed for PESrobust (prop > 0.50).*

	PES_{diff}				PES_{rel}				$PES_{robuust}$			
Learning activity	B	SE	*t*	*p*	B	SE	*t*	*p*	Proportion RTpost > RTpre	*N*	χ^2	*p*
Addition	419.8	15.0	28.0	<0.001	0.05	0.002	28.4	<0.001	0.54		687.1	<0.001
Counting	63.2	13.8	4.6	<0.001	0.01	0.001	5.0	<0.001	0.51		34.6	<0.001
Dictation	190.3	29.8	6.4	<0.001	0.03	0.003	9.1	<0.001	0.52		39.7	<0.001
Division	418.0	17.4	24.0	<0.001	0.07	0.002	38.6	<0.001	0.55		943.4	<0.001
Flowercode	150.8	26.6	5.7	<0.001	0.02	0.003	5.7	<0.001	0.51		24.0	<0.001
Fractions	461.9	55.6	8.3	<0.001	0.04	0.006	7.4	<0.001	0.53		33.0	<0.001
Grammar	**43.5**	**26.8**	**1.6**	**0.10**	0.01	0.003	2.7	0.008	**0.50**		**2.2**	**0.075**
LetterChaos	**-41.1**	**25.5**	**-1.6**	**0.11**	**-0.01**	**0.002**	**2.4**	**0.02**	**0.50**		**2.2**	**0.930**
Money	299.7	9.0	33.1	<0.001	0.03	0.001	27.5	<0.001	054		450.9	<0.001
Multiplication	385.6	18.7	20.6	<0.001	0.05	0.002	25.4	<0.001	0.54		451.0	<0.001
Numbers	115.3	33.2	3.4	0.001	0.01	0.004	2.6	0.01	0.51		4.0	0.022
Parsing words	81.5	29.8	2.7	0.006	0.01	0.003	3.3	<0.001	0.51		4.4	0.018
Proverbs	109.7	42.8	2.6	0.01	0.01	0.005	2.3	0.04	0.51		7.8	0.003
Reading	68.5	18.8	3.6	<0.001	0.01	0.002	4.8	<0.001	0.51		23.5	<0.001
Series	428.1	35.4	12.1	<0.001	0.05	0.004	12.9	<0.001	0.55		163.5	<0.001
Slow mix	611.9	21.2	28.9	<0.001	0.06	0.002	26.2	<0.001	0.55		403.4	<0.001
Speed mix	154.9	21.7	7.2	<0.001	0.04	0.002	16.8	<0.001	0.54		250.5	<0.001

Spelling	**-87.4**	**16.9**	**-5.2**	**<0.001**	**-0.01**	**0.002**	**8.0**	**<0.001**	**0.49**		**49.3**	**1.000**
Subtraction	438.6	16.8	26.2	<0.001	0.05	0.002	29.5	<0.001	0.54		628.9	<0.001
Telling time	150.4	20.7	7.3	<0.001	0.01	0.002	5.6	<0.001	0.51		16.9	<0.001
Verbs	232.6	23.7	9.8	<0.001	0.03	0.003	10.9	<0.001	0.52		76.7	<0.001
Vocabulary	175.4	18.0	9.8	<0.001	0.02	0.002	11.8	<0.001	0.52		92.4	<0.001
Word forms	320.5	22.6	14.2	<0.001	0.05	0.002	18.6	<0.001	0.53		210.0	<0.001

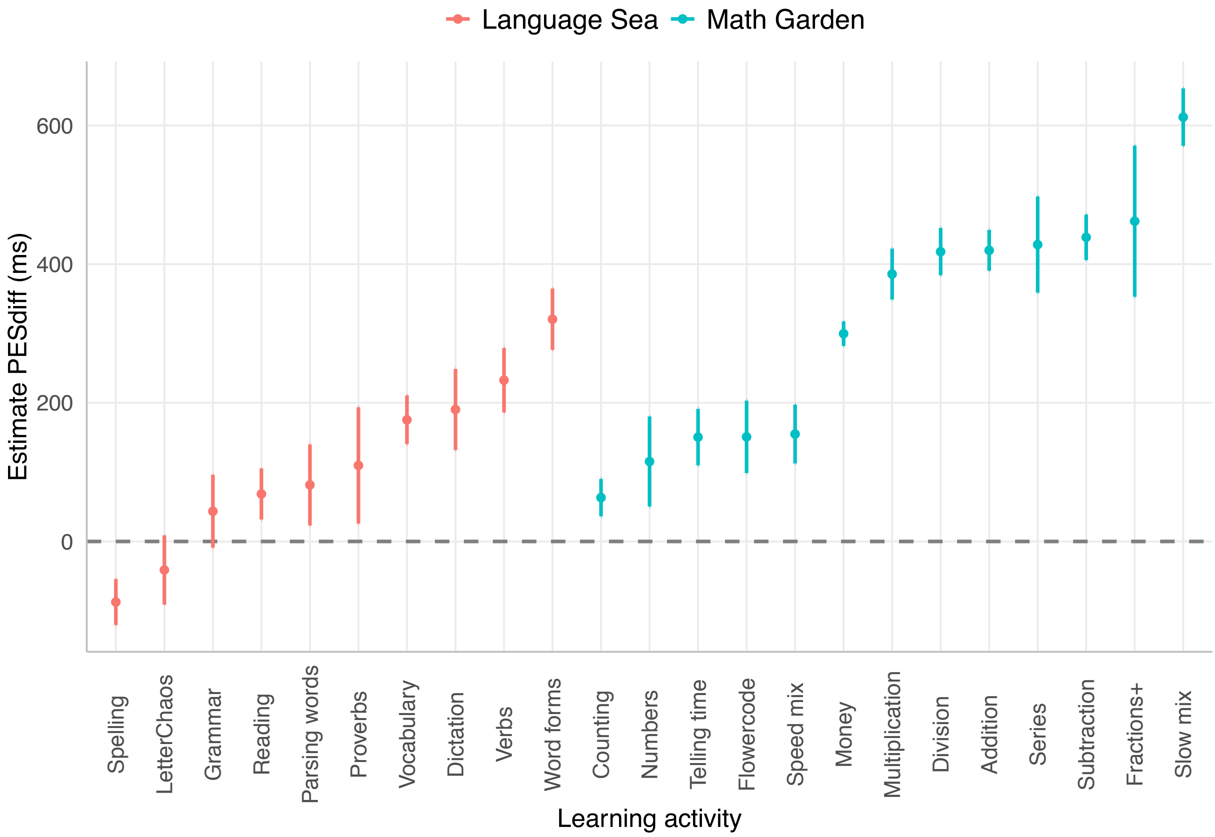


Figure 5.1. *Estimated absolute post-error slowing (PES_{diff}) in the 23 learning activities divided into practising language skills (red) and mathematical skills (blue). PES was observed for all activities, meaning coefficients above zero (dashed line), except for letterchaos and spelling (see Appendix A.1 for the game details). The points are the regression coefficients estimated from the linear mixed model; the vertical lines represent the 95% confidence intervals.*

5.3.2 *Type of learning activity and time pressure*

Next, a variety of predictors were added and compared to a basic model with participant and learning activity as random effects to see whether these predictors could explain variance in the magnitude of PES. Since the different PES calculations showed comparable outcomes for the following analyses, we only discuss PES_{diff}.

We first investigated task level predictors of PES, to understand why the learning activities differed in PES magnitude. We found that type of skill practised (mathematics vs language) significantly predicted PES magnitude. The Tukey comparison test showed that greater PES was shown in learning activities practising mathematical skills (M = 276 ms, SE = 12.4 ms) than language skills (M = 82.3 ms, SE = 11 ms), p = 0.02 (**Fig. 5.1**). As an addition to this model, we found that how much time participants were given to respond on each trial (i.e., time pressure) predicted PES, such that greater PES associated with more time to respond, β = 142.8, t (2942) = 8.27, p < 0.001 (**Fig. 5.2**). The time to respond did not differ between the learning activities practising language (M = 26.0 s) and mathematical skills (M = 29.2 s), p = 0.45. Time pressure also did not interact with the type of skill practised, p = 0.92. To ensure that the time pressure effect was not just driven for the reason being that PES magnitude is proportional to the time permitted, a follow-up analysis was done. In this follow-up analysis, both the time pressure metric was log-transformed as well as the pre-error RT and post error RT, before calculating PES_{diff}. Using these transformed variables in a linear mixed model, we replicated the finding that children showed greater PES when there was more time to respond, β = 0.026, t (2942) = 4.92, p < 0.001.

While all learning activities showed on average slower RT on error trials than on correct trials, there was an association between time to respond and error RT, such that longer time to respond associated with slower error RTs, β = 0.46, t (1519124) = 422.8, p < 0.001. Furthermore, when there was more time to respond, participants also exhibited slower errors compared to the RTs of the correct trials, β = 0.15, t (1519124) = 189.1, p < 0.001.

In addition, the trial level type of error (fast vs slow errors) was analysed and found to imply a better model to explain PES than the basic model, $\chi^2 = 380.96$, p < 0.001. Here, participants showed a greater PES effect after slow errors (M = 276 ms, SE = 11.6 ms) than after fast errors (M = 147 ms, SE = 12.4 ms), in line with what was found at the task level.

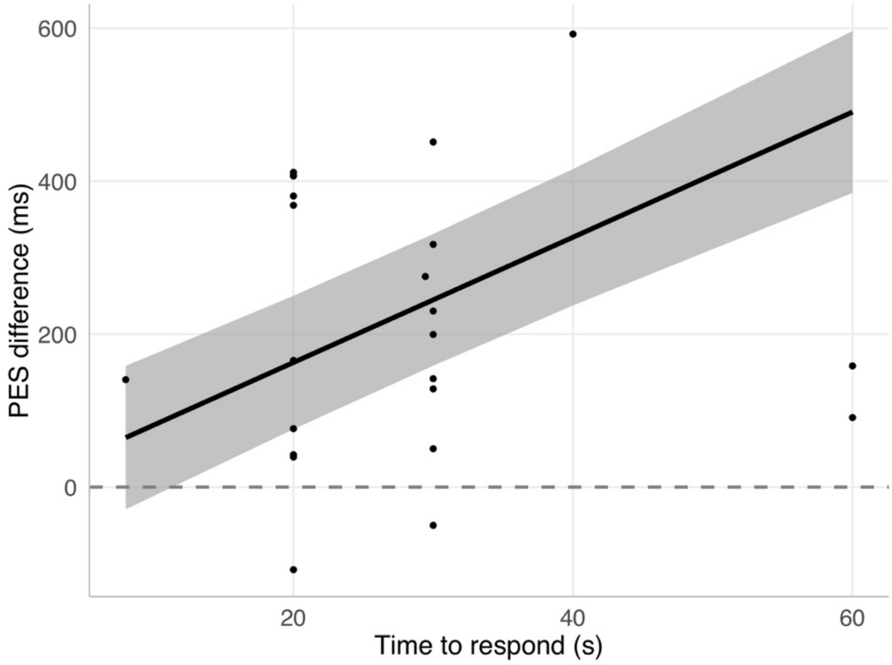


Figure 5.2. *Positive association between PES effect (ms) and how much time participants were given to respond (s) in the learning activities. The individual points represent the average in each of the 23 learning activities. The grey band around the linear function represents the 95% confidence interval.*

5.3.3 *Difficulty level (chosen by the participant)*

At the participant level, we examined whether including the difficulty level chosen by the student, which affects the error rate, would change the basic model (participant and learning activity as random effects) fit. As a check, there were no significant differences in age between the children choosing the difficulty levels. We found that adding difficulty level as a predictor significantly improved the model, $\chi^2 = 205.08$, $p < 0.001$. **Fig. 5.3** shows that participants choosing the most difficult level in learning activities also had the greatest PES effect (*M* = 270 ms, SE = 12.6 ms) compared to participants choosing the medium level (*M* = 208 ms, SE = 11.8 ms) and easy level (*M* = 129 ms, SE = 10.8 ms). Follow up comparisons using a Tukey test showed that participants choosing the hard level had a significantly greater PES than participants choosing the medium and the easy levels, *p's* < 0.001, and children choosing the medium level also showed significantly greater PES than children choosing the easy level, $p = 0.006$.

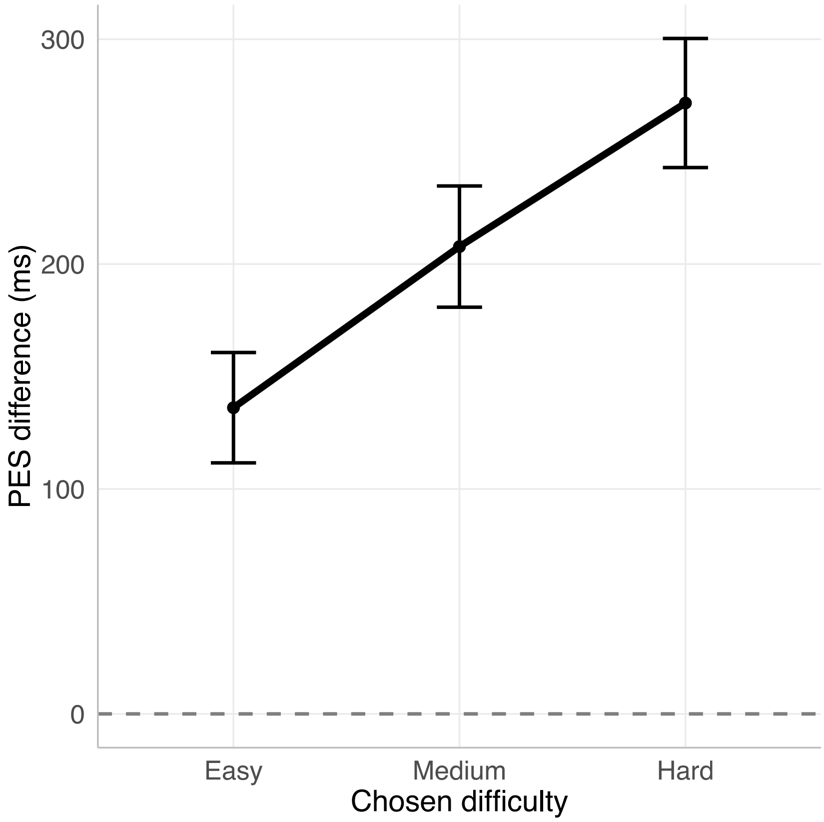


Figure 5.3. *PES magnitude as a function of the difficulty level chosen by the participants. The difficult level corresponds to the probability that participants will give a correct answer: 90% at the easy level, 75% at the medium level and 60% at the hard level (60%). The grey vertical lines represent the 95% confidence intervals.*

5.3.4 *Age and ability*

To investigate the development of PES, we examined the association of PES magnitude with children's age and ability level. We found that age predicted PES linearly (β = -50.6, p < 0.001) and quadratically (β = -26.5, p < 0.001), such that the magnitude of PES increased from 6 to 9 years old and decreased from 9 to 13 (**Fig. 5.4A**). In the same model, the children's ability level was found to predict the PES effect positively in a linear way (β = 156.6, p < 0.001), but not quadratically, p = 0.60 (**Fig. 5.4B).**

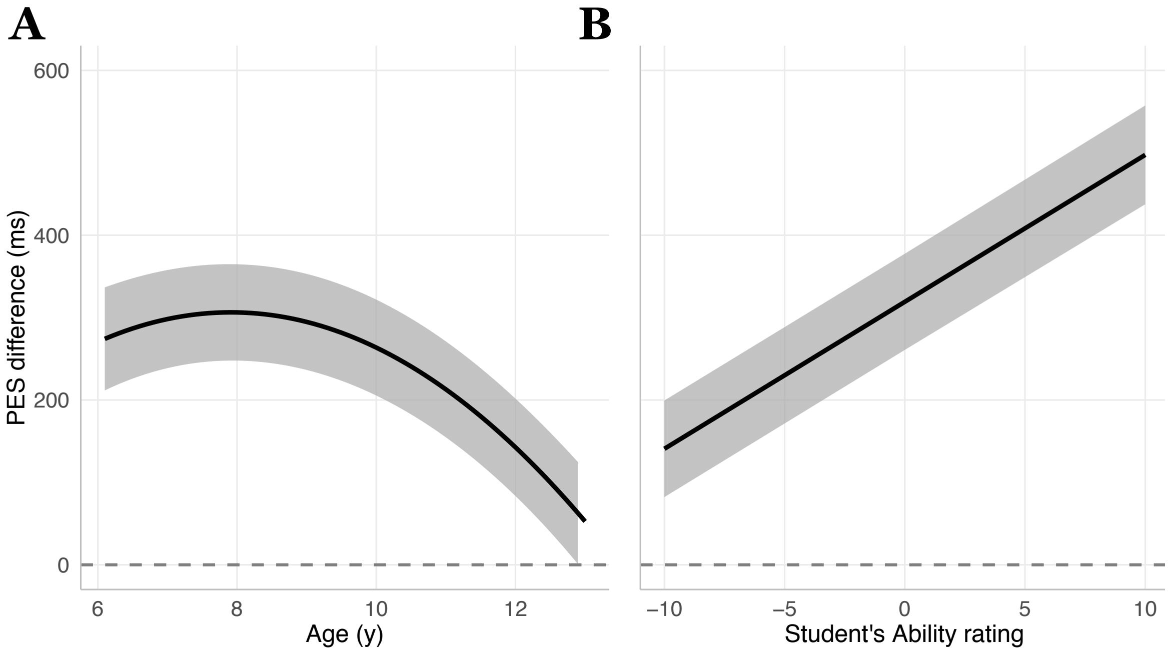


Figure 5.4. *PES difference predicted by **(A)** a combination of the linear and quadratic function of age and **(B)** a linear function of ability rating. The grey bands around the functions represent the 95% confidence interval.*

5.4 DISCUSSION

The purpose of this study was to investigate whether post error slowing, a marker of adaptive behaviour can also be found in an online learning environment for primary school children with complex tasks. In nearly all 23 mathematical and language-related learning activities we found post error slowing. To ensure a robust result, we used three different calculations of PES that take into account the impact of fatigue, general response speed and the magnitude of the RT difference. We found no difference in the main results between these calculations. Further analyses did reveal variability in the magnitude of PES as a function of task and individual differences. Learning activities that involved practising a mathematical skill and learning activities with longer time to respond showed greater PES effects. In line with this finding, we found that the PES effect was highest after making a slow error, in comparison to speeded errors. Looking at individual differences between children we found that children who chose the most difficult level had the greatest PES. With age, we found that from 6 to 9 years of age, PES effects increased and from 9 to 13 years of age declined and that greater ability on the learning activities independently (in addition to the age effect) predicted greater PES magnitude.

Overall, the study showed promising findings that PES is a behaviour that can be observed outside the laboratory. The findings relating to the mathematically focused learning activities are in line with prior research examining PES within the context of mental arithmetic tasks (Borght et al., 2016; Desmet et al., 2012; Lavro, Levin, et al., 2018; Núñez-peña et al., 2017). The current study supports the finding that children slow down their performance after an error in complex tasks in a range of mathematical tasks but more importantly, they also show this in an online practice system. No previous studies, to our knowledge, have addressed the PES effect in linguistic tasks. Although the effects were remarkably smaller, we did found PES in the linguistic domain. The substantial difference between the domains could be due to how the different learning platforms are designed to practice the particular skills. Further research is needed to support this finding using language-related tasks (such as spelling or reading), outside and inside the laboratory.

Studying PES in an online learning environment has both advantages and limitations, as was also discussed in Chapter 4 (section 4.4). A big advantage for this particular study is that PES could be investigated in cognitively demanding tasks for all ages and ability levels. But the drawbacks of collecting data in such a setting, where there is no control over the circumstances in which children can play, as well as what and when they play, was also a challenge in this study. The freedom in play leads to a different number of responses made per student and task, creating a lot of missing data. We attempted to accommodate these complex data by using learning activity and student as random effects in a linear mixed model.

Since we reliably observed post- error responses across tasks, the next question was why it occurs and in which conditions. Notebaert proposed that PES is the result of unexpected infrequent events, where the attention is shifted towards the errors, instead of performance (Notebaert et al., 2009). Our finding, that students choosing to practice in the highest probability level of making an error showed the greatest PES effect, is in contrast with Notebaert's argument that PES diminishes as error frequency increases. It is debatable whether the impact of committing an error in this study is the same as for some of the previous studies. The orienting account was originally described for tasks where the errors represent impulsive incorrect response selection due to stimulus ambiguity, where errors are rare. Notably, in the tasks investigated here, the errors are more focused on learning, rather than performance, and problems become gradually more difficult for everyone instead of being simply repeated. Arguably, the type of error monitoring in a learning environment impacts the strategy use of students more. When children are learning, they typically try less sophisticated strategies first, and after direct feedback (i.e., an error) more sophisticated strategies that might take longer to perform (Lemaire & Siegler, 1995; see section 1.7). Experimenting with strategies may lead to greater PES. Children choosing the most difficult levels are more challenged in terms of learning new skills, that require more sophisticated strategies, which may account for the larger PES magnitude they show. This is also in line with our finding that students who show greater PES and may therefore be more closely monitoring their errors and adjusting their strategies are the more able students who can solve more difficult problems. This argumentation line is more akin to the functional account of PES, where students will improve performance (become more able in the practised skills) by slowing down after an error (Botvinick et al., 2001; Dutilh, Ravenzwaaij, et al., 2012).

In contrast to Damaso et al. (2020), we found that children show greater PES after slow errors than after fast errors. We also found that learning activities with longer time to respond and therefore less pressure to answer quickly, with less chance of speedy errors, showed bigger PES magnitude. This could again be in line with the use of more sophisticated strategies after an error, such that tasks with less pressure invite students to try out different strategies resulting in longer RTs. What is different from previous studies such as Damaso et al. (2020), is that in this learning environment there is an emphasis on both speed and accuracy for all learning activities, therefore the influence of putting greater emphasis on speed, or accuracy, could not be investigated. In Chapter 3, I also discussed the differential impact of time pressure on academic performance depending on their cognitive abilities. The results showed that with no visible time pressure the children with better IC skills also showed better arithmetic performance. The enhanced performance could be the result of better attention to error monitoring, as well as using more sophisticated strategies after an error. So, less focus on time pressure might lead to more adaptive behaviour, depending on the child's cognitive abilities.

The age-related differences in PES, in this study, showed that from 6 to 13 years of age PES does occur, in line with the study of Smulders et al. (2016). But contrary to this study, where they found stability with age, we found a non-gradual development, with increasing PES until the age of 9 followed by decline. Gupta et al. (2009) also observed developmental changes, but with an earlier peak at the age of 7. However, between the ages of 7 and 8, this decrease was not uniform, and the biggest reduction occurred between 9 and 10. Their explanation for the curve of the development in error monitoring is that children, also go through a major development in executive function, especially task switching (see section 1.5.4), and error processing. Inhibition, which is required to withhold and delay a response, is thought to be the mechanism underlying error processing (Grammer et al., 2014; Gupta et al., 2009; Marco-Pallarés et al., 2008). Before the age of 7, children are still developing their ability to monitor and process errors and might therefore show large reaction time differences after an error. After the age of 7, children are more able and faster to recover from prior error trials with fewer switch costs, causing a decrease in PES. Similar to Chapter 3, where visible time pressure played a role, it could be that individual differences in the development of IC skills also differentially impact how children adapt to errors. For further research, it would therefore be interesting to study the development of error processing longitudinally as well as its association with executive function development, such as task switching and inhibitory control.

To conclude, we found that post-error slowing, a marker of adaptive behaviour, could be observed when children practised their skills in an online learning system. We also showed that the magnitude of post-error slowing depended on (1) the task, in terms of the amount of time pressure and the type of skill practised; (2) whether previous answers were given fast or slowly; (3) and characteristics of the learner, in terms of age, ability and their choice for difficulty level. The results of this study suggest that post-error slowing may be a good proxy for whether children monitor their errors and adapt their behaviour. In turn, this could be used to predict children's performance and progress in the skills practised.

6 DISCUSSION

“*To be the most convincing advocate of learning analytics, you also need to be its biggest critic*” *(Ian dolphin at the Learning & Student Analytics Conference 2018)*

6.1 Summary of main findings

In this dissertation, my aim was to find and develop possible markers that explain the large variance in children’s performance in online learning environments. Such markers could then help to optimise the adaptivity in such systems, which have the advantage to tailor to the learner’s needs. Previous cognitive research has provided insights into which markers such as individual differences in strategy use, executive function and error monitoring relate to learning, as reviewed in **Chapter 1**. To measure the contributions of these individual differences in an online learning environment, we need innovative methods, such as eye and mouse tracking. In **Chapter 2**, similarity and differences between eye and mouse data were assessed in a cognitive and arithmetic task. One aim was to investigate whether mouse tracking can give information about cognitive processing when eye tracking is not feasible. We found that both mouse and eye tracking can give a measure of one of the core abilities of executive function, namely IC. In the second task when solving arithmetic problems, children showed different patterns of strategies following their mouse and eye gaze movements and were not always in line with each other. Both mouse and eye trajectories did deviate towards errors before arriving at the correct answer. Eye and mouse tracking could therefore provide useful information on the cognitive process of the learner.

Chapter 3 described how individual differences in executive function in terms of verbal WM and IC predicted arithmetic performance in a lab study. In an arithmetic task, the cognitive load was manipulated by the visibility of time pressure, a key aspect of most gamified learning environments. Children’s WM positively predicted arithmetic performance, irrespective of the countdown visibility. For IC there was a positive association with performance only when time pressure was visible, and this was particularly true for younger children. Finally, the number of children's eye fixations on the question and errors differed depending on the visibility of time pressure. Overall, this study showed that visibility of time pressure may create cognitive overload and affect attention control during arithmetic problem-solving, depending on the cognitive profile of the child.

In this small lab study, mouse and eye movement provided useful information regarding the cognitive process during arithmetic problem-solving. In **Chapter 4**, mouse tracking was implemented in one of the games of Learn to do a larger scale study. This study aimed to expand on the results from **Chapter 2** and test whether mouse movements could reflect children's hesitations and incorrect associations (e.g., 7 × 8 = 52) during arithmetic problem-solving. We found that mouse movement during the thinking process correlated with systematic difficulties that children encounter, even when the final response was correct. These mouse movements mapped especially on frequently made errors for easy problems and younger children. Moreover, children that previously made certain errors still showed some attraction of mouse movement towards the same errors. Being able to detect such systematic difficulties allows the provision of adapted feedback and instruction, without relying on the children making errors.

At the same time, a child should monitor their errors. **Chapter 5** describes a large study in Learn investigating PES, a behavioural metric thought to reflect children's capacity to monitor and adapt to their errors (Rabbitt & Rodgers, 1977). The first aim of this study was to see whether PES could consistently be observed across tasks in a naturalistic setting. Results showed evidence of PES in most language and mathematical learning activities. However, there were also differences between the tasks. The PES effect was larger for mathematic tasks, as well as for tasks where children were given more time to respond. A second aim of the study was to investigate individual differences in PES. Individuals who chose more difficult problems showed larger PES effects; and more able children showed larger PES effects. These findings of larger PES were interpreted as reflecting more adaptive behaviour in maturing their strategies. Developmental differences in PES showed a non-linear pattern peaking at age 9. The age effect suggests increasing awareness of errors and the ability to change behaviour with age.

6.2 Contributions and considerations

My inspiration to start this PhD was to create a bridge between what we know from lab-based cognitive research to applied research and implementations carried out through an online educational tool used on a larger scale. Certain insights can be drawn from bridging both worlds.

1. The role of executive function for online adaptive learning.

The impact of individual differences in EF for learning is well known and studied extensively. Therefore, high hopes were set in research and education on training EF to improve other skills such as arithmetic as well, but this far transfer has been limited so far (Kassai et al., 2019). In the first study, we found that EF can be a limiting factor for arithmetic performance but that this also depends on the circumstances under which the task is conducted. Instead of focusing only on training EF to facilitate academic performance, we could also try to adapt learning systems to the EF limitations of the children. This would be consistent with the suggestion that in-class teachers should adapt their teaching to take into account the WM capacity limits of their pupils (J. G. Elliott et al., 2010; Gathercole et al., 2008). Online adaptive learning environments already adapt to the difficulty of the task to the learner's ability to foster learning by placing children in their zone of proximal development (Murray & Arroyo, 2002). This could be extended by presenting information in a way that is tailored to a child and adapting a cognitively efficient design. Adaptation can be in the form of reducing the distractions in the design of a task, such as the visibility of coins counting down that I investigated. For other children, having external markers of time may be a way to reduce cognitive load, as it saves them from having to estimate time in parallel to solving a problem. A concrete solution for an online learning environment would be for children or teachers to choose their design setting themselves, for example whether they keep time pressure visible or not. Another way would be to individually test the impact of changes in design on performance. This can be done with for example a bandit algorithm, an algorithm that recommends and tries out different settings at the right time for a user or a group of users to increase learning progress (Intayoad et al., 2020).

Dual-task environments can also arise in multimedia learning, for example when a child is given information out loud as well as in written text on the screen. The load on cognitive capacity can be reduced by presenting a single stream of presentation (e.g., narration) or by presenting the information at the same time in different formats (corresponding spoken and printed words) (Moreno & Mayer, 2003). In classrooms, a similar situation can occur where particularly young children, depending on their EF limitations, are cognitively overloaded by the format of instruction (Gathercole et al., 2008), or by too many visual distractions (Fisher et al., 2014) or noise, in the classroom (Massonnié et al., 2019). Too many distractions or lines of information should be avoided in all intelligent systems.

2. Eye and mouse tracking inform on the underlying cognitive process.

The studies described in **Chapters 2-4** showed that both eye and mouse tracking may be informative of the cognitive processes at play during arithmetic problem-solving. This information can be useful when inferring and reducing the cognitive load that a certain design bears upon a student. In the first study, children looked more at certain errors when the time pressure was visible, which could be an indication of cognitive overload. The findings also showed that the number of eye fixations on certain information during cognitive and arithmetic tasks were indicative of the performance of the child. Another useful implementation of eye and mouse tracking in educational tools is to infer the problem-solving strategies children use, to get an insight into their learning process and adapt the materials accordingly (see for example Causse et al., 2019). A combination of eye and mouse movements showed that children use their eyes to acquire information and can reveal whether children are mentally calculating a problem or are doubting between answer options for example. Notably, children did not always move their mouse to where their eyes went. But when there are attractive alternative options, children do tend to move their mouse towards the designated boxes.

Since the use of eye tracking remains difficult to implement on a larger scale such as in Learn, mouse tracking was investigated as an alternative tool to use online. The results showed promising findings, predominantly that mouse tracking may facilitate the identification of error patterns and potential misconceptions. Employing mouse tracking to this aim is useful because, while a lot of information is typically needed to find systematic errors, errors are rare in online learning environments. Mouse movement data on the correct trials can therefore provide additional information. Mouse tracking can be implemented in online learning environments as well as in lab experiments as an additional tool for inferring cognitive processes. Some issues that first need to be solved are (1) developing analysis pipelines, preferably open-source, and (2) storing such a large dataset, with a minimum number of 35 measures per second or more. For future educational research, sufficient webcam-based eye tracking or cheap smart glasses would also be a great benefit when available for larger-scale use.

More concretely, online learning environments can use mouse tracking or webcam-based eye tracking to map out children's network interference in arithmetic fact retrieval, without the need for them to make mistakes or to make it too challenging. Based on how differentiated the strength of attraction towards incorrect associations are, a particular problem as well as associated problems could be offered more often to make fact retrieval more automatic. The aim is that children build the correct associative strengths with the correct answers.

3. PES as a measure of the zone of proximal development for a child.

Error monitoring and adapting behaviour accordingly are key components of the development of human cognitive control. In **Chapter 5**, it was shown that PES could be observed in a range of tasks and was sensitive to task format and individual differences in children's ability and choice of difficulty level, suggesting that PES may be a useful marker of adaptive behaviour. Moreover, the PES measure showed a developmental trajectory with age and ability related differences. By tracking whether children are slowing down their performance after an error when practising their skills, we have an idea of which children are monitoring their errors and may be adapting their behaviour, whereas appropriate constructive feedback can be given to children who are not. In addition, we found that the development of PES is likely to depend on the development of executive function such as inhibition and task switching (e.g., Gupta et al., 2009). Therefore, it would be interesting to study individual differences in PES in conjunction with EF measures.

In the Learn environment, children can choose between three thresholds of success to play in, but the standard of .75 is set such that the level is engaging as well as motivating. The idea behind adaptivity is that children should play in their zone of proximal development (Vygotsky, 1978). The high accuracy rate is chosen to minimise children's anxiety and ensure they remain motivated to practise. At the same time, making errors allows reflecting and adapting behaviour. This is also shown in the PES study, where children in the most difficult level also showed the most PES. By monitoring errors and adapting their response, children develop more mature strategies. Therefore, some individuals might thrive under higher frequencies of errors to stay in their zone of proximal development.

Constantly tracking the PES metric as an indication of the user's error detection mechanism in an online learning environment can be useful for feedback and for selecting the next problems. Respectively, when children do not show a reaction to the feedback on errors, i.e., no PES, the algorithm can be programmed such that easier problems are selected. When repeating these particular problems, it might be good to provide clearer feedback (through videos or tips for other strategies) so that children are more equipped to learn from their mistakes.

Related to the discussion of this challenge versus motivation trade-off is the mindset framework, introduced by Dweck (2008). According to Dweck, there are two mindsets an individual can have, namely a fixed mindset or a growth mindset. Children with a fixed mindset have the belief that intelligence is predetermined and cannot be changed: you are either smart or not. Individuals with a fixed mindset tend to avoid challenge, due to the risk of failing. These students avoid the risk of making errors because they regard mistakes as indicators of their low ability (Boaler, 2013). On the other hand, students with a growth mindset believe that intelligence is malleable, where anything can be learned through effort. Mistakes are seen as a predictable outcome on the way to mastery. These students are shown to work and learn more effectively with a desire for a challenge (Dweck, 2008). Following this finding, numerous studies have examined the impact of growth mindset interventions on student's achievement. A meta-analysis of Sarrasin et al. (2018) shows that inducing a growth mindset in participants from the age of 7 to adulthood has a positive impact on their motivation and achievement, especially mathematics achievement. Moreover, some studies have also linked a growth mindset to post-error performance (Moser et al., 2011; Sarrasin et al., 2018; Schroder et al., 2017). They found that children with a growth mindset allocated more attention towards the mistakes and showed great post error accuracy. These authors demonstrated that having a growth mindset leads to greater use of optimal strategies after mistakes, enhancing performance. A growth mindset might enhance adaptive behaviour and performance, and therefore mindset interventions should be encouraged. But also, when adapting the probability of making errors, the mindset of the child should be considered: a child with a fixed mindset is likely to give up more easily with a high frequency of errors, whereas a child with a growth mindset is more likely to thrive in the process of making mistakes.

6.3 Limitations and future directions

The studies that I conducted were both in the lab with a small sample and in the more natural setting of an online learning environment on a larger scale. Both settings have some limitations that should be mentioned.

6.3.1 *Lab experiments versus online environment*

One of the difficulties for cognitive science in general is finding participants, even more so when you want the participant pool to be somewhat heterogeneous in terms of demographics, intelligence, and cognitive ability. If children need to come to the lab, you need parents willing to bring them in, outside of school hours. Even if the study can be done in the classroom, it is still very difficult to find schools that want to participate because teachers have a high workload. Although not all platforms allow the use of data for research, there are some collaborations with science, such as in Learn, Knewton, and ALEKS. The big advantage of researching in such an environment is that there are enough participating schools and students and that it does not require extra time from the teachers, parents or students. Online learning environments also provide a unique way to measure children's performance daily, longitudinally, which is not possible for an experimenter in the lab. The risk of having such a dataset, however, is that big data analyses are often exploratory and spurious associations are easy to find. In addition, the researcher often has to make many, and sometimes arbitrary, choices for the data selection and the handling of missing data. Another downside of this approach is that the assessment is much less controlled than when an experimenter is present because children can usually engage in online learning activities anywhere, at any time. Educational toolmakers and companies are also limited in terms of what can be measured, which is less of a problem in the lab: for example, a lab study can easily include additional questionnaires for parents or children or ask for sensitive information, such as any developmental diagnosis of the child or the parents' socio-economic status. The combination of using both controlled experiments and large-scale data collection in online learning environments is probably the best way to expand our knowledge of children's learning.

6.3.2 *Testing*

The studies conducted in this PhD measured children's performance at a current time point or for short periods, either six weeks or six months. Measuring performance is not the same as measuring learning, which is characterised by changes in performance over time. The idea of a learning environment is that repeated practice will lead to learning. Predictors of performance at a given time point are important proxies for this learning process. To measure the impact of the predictors for the learning and developmental trajectories for education, the predictors of performance measures should be repeated over the schoolyears during primary school. Longitudinal studies and student assessment surveys such as PISA (http://www.oecd.org/pisa/data/) and TIMSS (https://timssandpirls.bc.edu) do this by testing children's cognitive or academic performance once per school Year. As explained in

Chapter 1, it is difficult to capture learning characteristics without regular longitudinal testing. In light of this trade-off, I would suggest a combination of high-frequency testing/tracking of the performance in short periods and repeating this over the school Years, which is possible in daily used learning environments. Instead of just characterising children's thinking at certain time points we should investigate the changes in strategy use over time within individuals to understand variations in learning trajectories.

At the same time, testing in such an environment as Learn is very different because it is a typical example of low stakes testing. How does this technology relate to exams or other tests that have consequences (i.e., high stakes testing)? Usually in the educational field, there is a clear distinction between practice and testing. In Learn, practice and testing are combined into one monitoring program. The disadvantage of low-stakes testing is that there is not a carefully executed standardisation that the child can be compared to. However, the advantage of low stakes testing is that children experience less performance anxiety. From previous research, we know that students with a low level of test anxiety tend to perform better on tests in comparison to high-level anxious children (Elliot & Dweck, 2013; Ferla et al., 2010). Second, a bad day or sickness does not matter for the profiling of performance over a long period. A combination of both a high frequency, low stakes monitoring system and infrequent high stakes testing seems like the best way to assess children's learning. For the Learn environment in general, a comparison was made between the national maths test, administered twice a year throughout primary school, and the ability scores measured in Learn. The correlations between these two measures ranged from .78 to .8. Online low stakes practice tools should be linked more often to high stakes testing.

For example, low stakes testing lends itself well to study the individual differences in learning through measuring repeated practice. During low stakes testing, students feel free to practice their skills and to experiment with certain strategies and online platforms are perfect for researchers to measure this progress. These practice stages are also a great platform to test different adaptive versions of the learning environment to optimise the adaptivity for the student. In a more formal high stakes test this progress can be standardised to see whether the optimisations in the learning environment leads to better performance.

6.3.3 *Limited executive function measures*

From previous research and the study in Chapter 3, it can be concluded that it is important to have a picture of the cognitive profile of a child to optimise adaptivity. At least verbal WM and IC were marked as important predictors of arithmetic performance. In this study, we chose a particular task to measure WM and IC, but there are many more to consider and to add if one wanted to build a more complete cognitive profile. In the data described in Chapter 3, there was also no association between the IC interference effect measured by a simon task and verbal WM measured by a backward digit span task, which you would expect given that both are underlying common EF. A combination of other measures of EF in these studies could have resulted in different findings. Multiple measures would then ideally be incorporated in a latent variable approach to remove task-specific variance (Huizinga et al., 2006; Miyake et al., 2000). When designing a study, the time needed to collect multiple measures of each EF component needs to be weighed as well. In a lab study, you can't hold the child engaged for too long. In an online learning environment, it is also difficult to motivate children for tedious EF tasks. The advantage of online learning environments is that it is not limited to 1 or 2 testing sessions such as in lab-based or school-based testing. In the future, it might be a possibility to alternate a range of short EF tasks weekly, to build the EF profile of a child. This EF profile could help to predict the performance and learning progress.

EF is a broad term and commonly includes shifting as a core component of EF (A. Baddeley & Della Sala, 1996; Miyake et al., 2000). Especially when practising arithmetic skills with games such as speed mix, a child needs to shift between applying various rules, for example from addition to multiplication and vice versa. As mentioned in section 1.5.1, the building blocks underlying EF can change with age, where some are more important at certain critical stages of development than others (e.g., Bardikoff & Sabbagh, 2017). Age is an essential moderator to consider when investigating the association between EF and academic abilities, preferably in longitudinal studies. For example, in the study of Chapter 3, we did not find that age (8-11 years) moderated the association between verbal WM and arithmetic performance, but that age did interact with the association between IC interference effect and arithmetic performance. To optimise the adaptivity in online learning environments, the algorithms should incorporate several EF measures such as working memory, inhibition and the ability to shift between mental states in these critical developmental stages.

6.3.4 *Trade-off between fun and serious learning*

Adaptive learning is intended to captivate and engage children in learning new skills and knowledge with online games. The literature has shown that this approach has positive effects on motivation and learning (Cheung & Slavin, 2013; Wouters et al., 2013). In the Learn environment, children enjoy the games sufficiently for them to be motivated to play at home outside of school hours. The data shows that around 20% of the problems are solved at home and 80% at school. The motivation stems from presenting the child with problems at the appropriate difficulty level, keeping the child in the zone of proximal development. Although the 'ideal' difficulty level to keep a child in the flow seems to depend on the system: For example, in Learn this is 75% and for Knewton (2012) this is set at 50%, in accordance with the IRT theorem.

A big part of the appeal of these environments is the use of external motivations to keep children engaged. In Learn, the coins that children receive for the correct answer is a big external motivation. This is strengthened by digital trophies that can be bought with the coins, similar to the Knewton platform. However, one consequence of this design is that some children tend to play already mastered games to receive more reward, which is not desirable behaviour. With a variety of games at hand, children also choose the more enjoyable and interesting games instead of practising basic skills such as multiplication tables. New studies in the Learn environment are working on 'governing' games so that children are supervised in which games are important for them to practice (Brinkhuis et al., 2020).

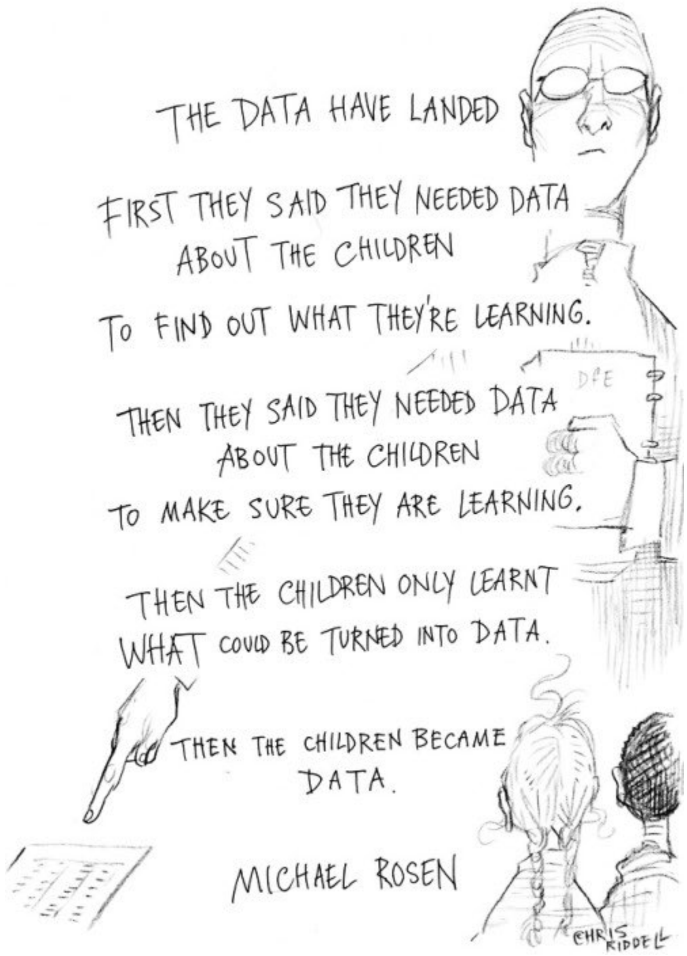
The challenge for educational toolmakers is to have rewards that also lead to intrinsic motivation for learning and to retain this for the long run. For some time, psychologists thought that the presentation of rewards reduces subsequent intrinsic motivation for the task, the alleged *undermining effect* (for a meta-analysis, see: Cameron et al., 2001). But a more recent meta-analysis (Cerasoli et al., 2014) showed that both intrinsic motivation and extrinsic incentives are critical for performance: intrinsic motivation matters more for the quality of performance (i.e., creativity, focus, being autonomous) and extrinsic incentives explain more of the variance in the quantity (i.e., productivity, drive, number of repetitions) of performance. They also positively interact with each other, such that the intrinsic motivation-performance link becomes stronger in the presence of incentives. These findings suggest that although it is always beneficial for people to be intrinsically motivated for the given task, rewards also influence performance.

6.3.5 *Concerns with embracing educational technology*

Researchers, educators and parents have been debating the place technology should have in schools and at home (Facer, 2011; Selwyn, 2011). Digital technology defines and shapes modern childhood, and it is difficult to entangle which aspects are beneficial to them and which ones are not. Almost every school in western countries have interactive technology in the classrooms such as whiteboards. The use of personal laptops or tablets for every student is increasing significantly as well: A school survey in the EU in 2013 showed that between three and seven EU students share one computer on average. And around 50% of the students in Year 7 use a device weekly during lessons at school (European Schoolnet, 2013). In the Netherlands, teachers spend around 8 hours a week on online lessons (European Schoolnet & Liege, 2013).

One of the concerns is that the increase in screen time leaves less time for physical activities such as face-to-face connection, playing games at school and outdoor play. There is limited hard evidence to date on the impact of digital technology on children's health. But so far, there has not been clear evidence of detrimental effects of technology use on children's behaviour, health, or learning when there is no excessive use (Howard-Jones, 2011; OECD, 2020). These reports show that the impact on children's well-being seems to be u-shaped, where no use or extensive use of technology can have a small negative impact, and moderate use a small positive impact (Kardefelt-Winther, 2017). Second, some studies have tried to link obesity to screen time, due to reduced physical activity and mindless eating, but there is no evidence for a causal relationship to suggest this (OECD, 2020). Overall, the literature argues for the 'goldilocks principle', meaning that moderate use of technology is the most positive outcome for children's development (Przybylski & Weinstein, 2017). The challenge for educators is to guide children in developing awareness of technology, and how to use devices to gain knowledge, but also when to put devices away (Plowman & McPake, 2013).

As pointed out by Michael Rosen's poem below, some people have the misconception that adaptive learning algorithms will replace teachers (Cavanagh et al., 2020). But these learning algorithms would never be able to replace all instructional context. A simple argument for this is that it is difficult to incorporate the social context of the learner's interactions with other learners and learning objectives (Selwyn, 2011). We know that from an early age the social environment is key to developing skills and knowledge (Frith & Frith, 2012; Leblanc & Ramirez, 2020). Educational tools therefore only work with good teachers that provide the structure and guidance to use online adaptive systems (Essa & Laster, 2017). Adaptive technology can facilitate student's learning process by providing teachers with individualised materials and insights into student's achievement, creating more time to focus on teaching.



A related limitation of online learning platforms is that they don't allow children to manipulate physical objects, which are thought to improve the understanding of abstract concepts. For example, to understand the concept of the number six teachers can use a range of concrete manipulatives, such as a dice, a finger pattern, six o'clock on an analogue clock, or six dinosaur toys in a row. In a meta-analysis (Carbonneau et al., 2013), studies were compared between teaching mathematics using physical manipulatives and teaching with abstract mathematical symbols. Small to moderate effect sizes were found in favour of teaching with manipulatives compared to instruction that only use abstract symbols. Additionally, moderate to large effects were found on the outcome of retention. Students and teachers also find enjoyment in using manipulatives (Moyer, 2001).

Very recently, the outbreak of COIVD-19 shed new light on another debate, namely inequality in learning opportunities due to a child's SES background and ethnicity. E-learning is dependent on having access to a personal device and strong internet connection, which is not available for all families and schools. In the UK, Andrew (2020) found that three times as many of the poorest students had no device to access schoolwork than the richest students. In addition, pupils of higher-income parents spent much more time on home learning during lockdown; had more access to individualised resources and felt they were better supported (Andrew et al., 2020). Similar results were found in the Netherlands (Bol, 2020). The inequality in access to learning tools such as adaptive learning tools will inevitably widen the education gap even more. Research is already suggesting that the lockdown will have far-reaching educational consequences (Andrew et al., 2020; Bayrakdar & Guvely, 2020; Bol, 2020). To minimise these inequalities, and if e-learning is to play a major role in education, governments will need to ensure that all children have the appropriate resources for e-learning.

6.3.6 *The end goal*

In a collaborative effort between education and research, adaptive learning could become available to all, with the aim to individually tailor the learning process on a massive scale. In this dissertation, several markers and measures were identified that may allow the optimisation of such an environment. The promise is that these data will help in observing patterns across children and over time, providing tools to enhance learning and allow children to reach their potential.

REFERENCES

Ackerman, P. L. (1990). A correlational analysis of skill specifity: learning abilities, and individual differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(5), 883–901.

Adkins, S. S. (2020). *The 2019 Global Learning Technology Investment Patterns : Another Record Shattering Year*. https://seriousplayconf.com/wp-content/uploads/2020/01/Metaari-2019-Global-Learning-Technology-Investment-Patterns.pdf

Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. *Child Neuropsychology*, *8*(2), 71–82. https://doi.org/10.1076/chin.8.2.71.8724

Anderson, S. E., Farmer, T. A., Goldstein, M., Schwade, J., & Spivey, M. J. (2011). Individual differences in measures of linguistic experience account for variability in the sentence processing skill of five-year-olds. In *Experience, variation, and generalization: Learning a first language* (pp. 203–221).

Anderson, V., Jacobs, R., & Anderson, P. (2011). Executive functions and the frontal lobes: A lifespan perspective. In *Executive Functions and the Frontal Lobes: A Lifespan Perspective* (pp. 1–541). USA: Taylor & Francis. https://doi.org/10.4324/9780203837863

Andrew, A., Costa-dias, M., & Costa-dias, M. (2020). *Learning during the lockdown : real-time data on children ' s experiences during home learning*. https://www.ifs.org.uk/publications/14848

Anwyl-irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst : An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388–407. https://doi.org/10.1101/438242

Artemenko, C., Soltanlou, M., Bieck, S. M., Ehlis, A. C., Dresler, T., & Nuerk, H. C. (2019). Individual differences in math ability determine neurocognitive processing of arithmetic complexity: A combined fNIRS-EEG study. *Frontiers in Human Neuroscience*, *13*(227), 1–13. https://doi.org/10.3389/fnhum.2019.00227

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

2297(82)90012-0
Ashcraft, M. H., & Battaglia, J. (1978). Cognitive arithmetic: Evidence for retrieval and decision processes in mental addition. *Journal of Experimental Psychology: Human Learning & Memory*, *4*(5), 527–538. https://doi.org/10.1037//0278-7393.4.5.527
Ashcraft, M. H., & Kirk, E. P. (2001). The relationship among working memory, math anxiety, and performance. *Journal of Experimental Psychology: General*, *130*(2), 224–237. https://doi.org/10.1037//0096-3445.130.2.224
Ashcraft, M. H., & Krause, J. A. (2007). Working memory, math performance, and math anxiety. *Psychonomic Bulletin & Review*, *14*(2), 243–248. https://doi.org/10.3758/BF03194059
Asteriadis, S., Tzouveli, P., Karpouzis, K., & Kollias, S. (2009). Estimation of behavioral user state based on eye gaze and head pose-application in an e-learning environment. *Multimedia Tools and Applications*, *41*(3), 469–493. https://doi.org/10.1007/s11042-008-0240-1
Avgerinou, V. A., & Tolmie, A. (2020). Inhibition and cognitive load in fractions and decimals. *British Journal of Educational Psychology*, *90*, 240–256. https://doi.org/10.1111/bjep.12321
Baddeley, A. D. (1992). Working Memory. *Science*, *255*(5044), 556–559. https://doi.org/10.1016/S0079-7421(08)60452-1
Baddeley, A., & Della Sala, S. (1996). Working memory and executive control. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*(1346), 1397–1404.
Bakker, M., Van Den Heuvel-Panhuizen, M., & Robitzsch, A. (2015). Longitudinal data on the effectiveness of mathematics mini-games in primary education. *British Journal of Educational Technology*, *46*(5), 999–1004. https://doi.org/10.1111/bjet.12320
Bardikoff, N., & Sabbagh, M. (2017). The differentiation of executive functioning across development: Insights from developmental cognitive neuroscience. In *New perspectives on human development* (pp. 47–66). https://doi.org/https://doi.org/10.1017/cbo9781316282755.005
Barkatsas, A., Kasimatis, K., & Gialamas, V. (2009). Learning secondary mathematics with technology: Exploring the complex interrelationship between students' attitudes, engagement, gender and achievement. *Computers and Education*,

52(3), 562–570. https://doi.org/10.1016/j.compedu.2008.11.001
Barrios, V. M. G., Gütl, C., Preis, A. M., Andrews, K., Pivec, M., Mödritscher, F., & Trummer, C. (2004). AdELE: A Framework for Adaptive E-Learning through Eye Tracking. *Iknow*, *4*(October), 1–8.
Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and Cognitive Load in Working Memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, *33*(3), 570–585. https://doi.org/10.1037/0278-7393.33.3.570
Barrouillet, P., & Lépine, R. (2005). Working memory and children's use of retrieval to solve addition problems. *Journal of Experimental Child Psychology*, *91*(3), 183–204. https://doi.org/10.1016/j.jecp.2005.03.002
Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01
Bauerlein, M. (2008). *The dumbest generation—how the digital age stupefies young americans and jeopardizes our future*. Penguin Group. https://doi.org/10.1061/(ASCE)1532-6748(2009)9
Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology*, *46*(3), 610–620. https://doi.org/10.1037/0022-3514.46.3.610
Bayrakdar, S., & Guvely, A. (2020). *Inequalities in home learning and schools ' provision of distance teaching during school closure of COVID-19 lockdown in the UK* (ISER Working Paper Series).
Beilock, S. L., & Carr, T. H. (2005). When High-Powered People Fail: Working Memory and "Choking under Pressure" in Math. *Psychological Science*, *16*(2), 101–105. https://doi.org/10.1111/j.0956-7976.2005.00789.x
Beilock, S. L., & DeCaro, M. S. (2007). From Poor Performance to Success Under Stress: Working Memory, Strategy Selection, and Mathematical Problem Solving Under Pressure. *Journal of Experimental Psychology: Learning Memory and Cognition*, *33*(6), 983–998. https://doi.org/10.1037/0278-7393.33.6.983
Ben-Zeev, T. (1998). Rational errors and the mathematical mind. *Review of General Psychology*, *2*(4), 366–383. https://doi.org/10.1037/1089-2680.2.4.366
Berwid, O. G., Halperin, J. M., Jonhson, R. J., & Marks, D. J. (2014). Preliminary Evidence for Reduced Post-Error Reaction Time Slowing in

Hyperactive/Inattentive Preschool Children. *Child Neuropsychology*, *20*(2), 196–209. https://doi.org/10.1080/09297049.2012.762760

Best, J. R., & Miller, P. H. (2010). A Developmental Perspective on Executive Function. *Child Development*, *81*(6), 1641–1660. https://doi.org/https://doi.org/10.1111/j.1467-8624.2010.01499.x

Best, J. R., Miller, P. H., & Naglieri, J. A. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large , representative national sample. *Learning and Individual Differences*, *21*(4), 327–336. https://doi.org/10.1016/j.lindif.2011.01.007

Bjorklund, D. F., & Causey, K. B. (2017). *Children's thinking: cognitive development and individual differences*. Sage Publications.

Blair, C., Zelazo, P. D., & Greenberg, M. T. (2010). The Measurement of Executive Function in Early Childhood. *Developmental Neuropsychology*, *28*(2), 561–571. https://doi.org/10.1207/s15326942dn2802

Block, R. A., & Grondin, S. (2014). Timing and time perception : A selective review and commentary on recent reviews. *Frontiers in Psychology*, *5*, 648. https://doi.org/10.1121/1.1918172

Block, R. A., & Gruber, R. P. (2014). Time perception , attention , and memory : A selective review. *Acta Psychologica*, *149*, 129–133. https://doi.org/10.1016/j.actpsy.2013.11.003

Block, R. A., Hancock, P. A., & Zakay, D. (2010). How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychologica*, *134*(3), 330–343. https://doi.org/10.1016/j.actpsy.2010.03.006

Bloom, B. S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, *13*(6), 4–16. https://doi.org/10.3102/0013189x013006004

Boaler, J. O. (2013). Ability and Mathematics: the mindset revolution that is reshaping education. *Forum*, *55*(1), 143–152. https://doi.org/10.2304/forum.2013.55.1.143

Bobadilla-suarez, S., & Love, B. C. (2018). Fast or Frugal , but Not Both : Decision Heuristics Under Time Pressure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(1), 24–33. https://doi.org/10.1037/xlm0000419

Bol, T. (2020). *Inequality in homeschooling during the Corona crisis in the Netherlands. First results from the LISS Panel.*

https://doi.org/10.31235/osf.io/hf32q
Bolden, D., Barmby, P., Raine, S., & Gardner, M. (2015). How Young Children View Mathematical Representations: A Study Using Eye-Tracking Technology. *Educational Research*, *57*(1), 59–79.
Borght, L. Van Der, Desmet, C., & Notebaert, W. (2016). Strategy Changes After Errors Improve Performance. *Frontiers in Psychology*, *6*(2051), 1–6. https://doi.org/10.3389/fpsyg.2015.02051
Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. https://doi.org/10.1037/0033-295x.108.3.624
Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357
Bray, B., & Mcclaskey, K. (2010). *Personalization vs Differentiation vs Individualization.* http://www.my-ecoach.com/online/resources/925/PersonalizationvsDifferentiationvsIndividualization.pdf
Brinkhuis, M., Cordes, W., & Hofman, A. (2020). Governing games: Adaptive game selection in the Math Garden. *ITM Web of Conferences*, *03003*, 1–10. https://doi.org/https://doi.org/10.1051/itmconf /20203303003
Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science*, *2*(2), 155–192. https://doi.org/10.1016/S0364-0213(78)80004-4
Brown, S. W., Collier, S. A., & Night, J. C. (2013). Timing and executive resources: Dual-task interference patterns between temporal production and shifting, updating, and inhibition tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(4), 947–963. https://doi.org/10.1037/a0030484
Brown, S. W., & Perreault, S. T. (2017). Relation between temporal perception and inhibitory control in the Go/No-Go task. *Acta Psychologica*, *173*, 87–93. https://doi.org/10.1016/j.actpsy.2016.12.004
Bruin, A. De, & Sala, S. Della. (2018). Effects of age on inhibitory control are affected by task-specific features. *Quarterly Journal of Experimental Psychology*, *71*(5), 11219–112233. https://doi.org/10.1080/17470218.2017.1311352
Bruning, R., Schraw, G., Norby, M., & Ronning, R. (2004). *Cognitive psychology and*

instruction. Columbus.

Brydges, C. R., Fox, A. M., Reid, C. L., & Anderson, M. (2014). The differentiation of executive functions in middle and late childhood : A longitudinal latent-variable analysis. *Intelligence*, *47*, 34–43. https://doi.org/10.1016/j.intell.2014.08.010

Bull, R., Johnston, R. S., & Roy, J. A. (1999). Exploring the roles of the visual-spatial sketch pad and central executive in children's arithmetical skills: Views from cognition and developmental neuropsychology. *Developmental Neuropsychology*, *15*(3), 421–442. https://doi.org/10.1080/87565649909540759

Bull, R., & Lee, K. (2014). Executive functioning and mathematics achievement. *Child Development Perspectives*, *8*(1), 36–41. https://doi.org/10.1111/cdep.12059

Bull, R., & Scerif, G. (2001). Executive Functioning as a Predictor of Children ' s Mathematics Ability : Inhibition , Switching , and Working Memory. *Developmental Neuropsychology*, *19*(3), 273–293. https://doi.org/10.1207/s15326942dn1903_3

Buscher, G., White, R. W., Dumais, S. T., & Huang, J. (2012). Large-scale analysis of individual and task differences in search result page examination strategies. *WSDM 2012 - Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, 373–382. https://doi.org/10.1145/2124295.2124341

Buwalda, T. A., Borst, J. P., van der Maas, H. L. J., & Taatgen, N. A. (2016). Explaining mistakes in single digit multiplication: A cognitive model. *Proceedings of the 14th International Conference on Cognitive Modeling*, *1*, 131–136.

Cameron, J., Banko, K. M., & Pierce, W. D. (2001). Pervasive Negative Effects of Rewards on Intrinsic Motivation: The Myth Continues. *Behavior Analyst*, *24*(1), 1–44. https://doi.org/10.1007/BF03392017

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

Campbell, J. (1995). Mechanisms of simple addition and multiplication: A modified network-interference theory and simulation. *Mathematical Cognition*, *1*(2), 121–164.

Canfield, W. (2001). ALEKS: A Web-based intelligent tutoring system. *Mathematics and Computer Education*, *35*(2), 152.

Carbonneau, K. J., Marley, S. C., & Selig, J. P. (2013). A meta-analysis of the efficacy of teaching mathematics with concrete manipulatives. *Journal of Educational Psychology*, *105*(2), 380–400. https://doi.org/10.1037/a0031084

Carpenter, T. P., & Moser, J. M. (1984). The Acquisition of Addition and Subtraction Concepts in Grades One through Three. *Journal for Research in Mathematics Education*, *15*(3), 179–202. https://doi.org/10.2307/748348

Case, R. (2013). *The mind's staircase: Exploring the conceptual underpinnings of children's thought and knowledge*. Psychology Press.

Castellar, E. N., All, A., de Marez, L., & van Looy, J. (2015). Cognitive abilities , digital games and arithmetic performance enhancement : A study comparing the effects of a math game and paper exercises. *Computers & Education*, *85*, 123–133. https://doi.org/10.1016/j.compedu.2014.12.021

Causse, M., Lancelot, F., Maillant, J., Behrend, J., Cousy, M., & Schneider, N. (2019). Encoding decisions and expertise in the operator's eyes: Using eye-tracking as input for system adaptation. *International Journal of Human Computer Studies*, *125*, 55–65. https://doi.org/10.1016/j.ijhcs.2018.12.010

Cavanagh, T., Chen, B., Ait, R., Lahcen, M., & James, R. (2020). Constructing a Design Framework and Pedagogical Approach for Adaptive Learning in Higher Education : A Practitioner ' s Perspective. *International Review of Research in Open and Distributed Learning*, *21*(1), 173–197.

Caviola, S., Carey, E., Mammarella, I. C., & Szucs, D. (2017). Stress, time pressure, strategy selection and math anxiety in mathematics: A review of the literature. *Frontiers in Psychology*, *8*, 1–13. https://doi.org/10.3389/fpsyg.2017.01488

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin*, *140*(4), 980–1008. https://doi.org/10.1037/a0035661

Cheadle, J. E. (2008). Educational Investment, Family Context, and Children's Math and Reading Growth from Kindergarten Through the Third Grade. *Sociology of Education*, *81*, 1–31.

Chen, M.-C., Anderson, J. R., & Sohn, M.-H. (2001). What can a mouse cursor tell us more? Correlation of eye/mouse movements on web browsing. *CHI'01 Extended Abstracts on Human Factors in Computing Systems*, 281–282.

Cheung, A. C. K., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms : A meta-analysis. *Educational Research Review*, *9*, 88–113.

https://doi.org/10.1016/j.edurev.2013.01.001

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-Student Matching and the Assessment of Teacher Effectiveness. *The Journal of Human Resources*.

Cloutier, J., Freeman, J. B., & Ambady, N. (2014). Investigating the early stages of person perception: The asymmetry of social categorization by sex vs. age. *PLoS ONE*, *9*(1). https://doi.org/10.1371/journal.pone.0084677

Cockcroft, W. H. (1982). *Mathematics counts*. HM Stationery Office.

Cognolato, M., Atzori, M., & Mu, H. (2018). *Head-mounted eye gaze tracking devices : An overview of modern devices and recent advances*. *5*, 1–13. https://doi.org/10.1177/2055668318773991

Cohen, J. (1988). *Statistical power analysis for the behaviors science*. (2nd ed.). New Jersey: Laurence Erlbaum Associates.

Collette, F., Linden, M. Van Der, Laureys, S., Delfiore, G., Degueldre, C., Luxen, A., & Salmon, E. (2005). Exploring the Unity and Diversity of the Neural Substrates of Executive Functioning. *Human Brain Mapping*, *25*, 409–423. https://doi.org/10.1002/hbm.20118

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. https://doi.org/10.1021/cen-v020n021.p1399

Cooke, L. (2006). *Is the Mouse a " Poor Man ' s Eye Tracker "?* 252–255.

Cormier, S. M., & Hagman, J. D. (1987). *Transfer of learning: contemporary research and applications*. Academic Press.

Corno, L., & Snow, R. E. (1986). Adapting teaching to individual differences among learners. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 605–629). Macmillan.

Cox, A. L., & Silva, M. M. (2006). The Role of Mouse in Interactive Search. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *28*(28). https://escholarship.org/content/qt4ch1s1bv/qt4ch1s1bv.pdf

Cragg, L., & Gilmore, C. (2014). Skills underlying mathematics: The role of executive function in the development of mathematics proficiency. *Trends in Neuroscience and Education*, *3*(2), 63–68. https://doi.org/10.1016/j.tine.2013.12.001

Cragg, L., & Nation, K. (2009). Shifting Development in Mid-Childhood : The Influence of Between-Task Interference. *Developmental Psychology*, *45*(5), 1465–1479. https://doi.org/10.1037/a0015360

Crone, E. A., Wendelken, C., Donohue, S., Leijenhorst, L. Van, & Bunge, S. A. (2006). *Neurocognitive development of the ability to manipulate information in working memory*. *103*(24). https://doi.org/10.1073/pnas.0510088103

Csikszentmihalyi, M. (1990). *Flow: the psychology of optimal experience*. Harper and Row.

Curriculum Associates. (2020). *Get to Know i-Ready Assessment Providing educators the most accurate and actionable data to help guide students toward continued growth*. https://www.curriculumassociates.com/-/media/mainsite/files/i-ready/iready-assessment-flyer-2020.pdf

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001

Dajani, D. R., & Uddin, L. Q. (2015). Demystifying cognitive flexibility : Implications for clinical and developmental neuroscience. *Trends in Neurosciences*, *38*(9), 571–578. https://doi.org/10.1016/j.tins.2015.07.003

Dale, R., Kehoe, C., & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory & Cognition*, *35*(1), 15–28.

Damaso, K., Williams, P., & Heathcote, A. (2020). Evidence for different types of errors being associated with different types of post-error changes. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-019-01675-w

Danielmeier, C., & Ullsperger, M. (2011). Post-error adjustments. *Frontiers in Psychology*, *2*(233), 1–10. https://doi.org/10.3389/fpsyg.2011.00233

David, C. V. (2012). Working memory deficits in Math learning difficulties: A meta-analysis. *British Journal of Developmental Disabilities*, *58*(2), 67–84. https://doi.org/10.1179/2047387711Y.0000000007

de Smedt, B., Janssen, R., Bouwens, K., Verschaffel, L., Boets, B., & Ghesquière, P. (2009). Working memory and individual differences in mathematics achievement : A longitudinal study from first grade to second grade. *Journal of Experimental Child Psychology*, *103*(2), 186–201. https://doi.org/10.1016/j.jecp.2009.01.004

Demir-lira, Ö. E., Prado, J., Booth, J. R., Price, G., & Evans, T. M. (2016). Neural Correlates of Math Gains Vary Depending on Parental Socioeconomic Status (SES). *Frontiers in Psychology*, *7*, 1–12.

https://doi.org/10.3389/fpsyg.2016.00892

Demšar, U., & Çöltekin, A. (2017). Quantifying gaze and mouse interactions on spatial visual interfaces with a new movement analytics methodology. *PLoS ONE*, *12*(8), 1–36. https://doi.org/10.1371/journal.pone.0181818

Department for Education. (2016). *Trends in Maths and Science Study (TIMSS): National Report for England*. https://www.gov.uk/government/publications/timss-2015-national-report-for-england

Desmet, C., Imbo, I., Brauwer, J. De, Brass, M., Fias, W., Notebaert, W., Desmet, C., Imbo, I., Brauwer, J. De, Brass, M., & Fias, W. (2012). Error adaptation in mental arithmetic. *The Quarterly Journal of Experimental Psychology*, *65*(5), 1059–1067. https://doi.org/10.1080/17470218.2011.648943

Diamond, A. (1989). Developmental Time Course in Human Infants and Infant Monkeys , and the Neural Bases of, Inhibitory Control in Reaching. *Annals of the New York Academy of Sciences*, *608*(1), 637–676. https://doi.org/10.1111/j.1749-6632.1990.tb48913.x

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, *64*(1), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Diamond, A., & Goldman-Rakic, P. S. (1989). Comparison of human infants and rhesus monkeys on Piaget's AB task: evidence for dependence on dorsolateral prefrontal cortex. *Experimental Brain Research*, *74*, 24–40. https://doi.org/10.1007/bf00248277

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 1–17. https://doi.org/10.3389/fpsyg.2014.00781

Dodge, S., Weibel, R., & Forootan, E. (2009). Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. *Computers, Environment and Urban Systems*, *33*(6), 419–434. https://doi.org/10.1016/j.compenvurbsys.2009.07.008

Doebel, S. (2020). Rethinking Executive Function and its Development. *Perspectives on Psychological Science*, *15*(4), 942–956. https://doi.org/10.1177/1745691620904771

Domahs, F., Delazer, M., & Nuerk, H. C. (2006). What makes multiplication facts difficult: Problem size or neighborhood consistency? *Experimental Psychology*, *53*(4), 275–282. https://doi.org/10.1027/1618-3169.53.4.275

Donders, F. C. (1969). on the speed of mental processes. *Acta Psychologica*, 412–431.

Dotan, D., & Dehaene, S. (2013). How do we convert a number into a finger trajectory? *Cognition*, *129*(3), 512–529. https://doi.org/10.1016/j.cognition.2013.07.007

Dowker, A. (2005). *Individual Differences in Arithmetic*. Psychology Press.

Droll, J. A., & Hayhoe, M. M. (2007). Trade-offs Between Gaze and Working Memory Use. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1352–1365. https://doi.org/10.1037/0096-1523.33.6.1352

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*(3), 259–268. https://doi.org/10.1016/j.jrp.2011.02.004

Dumontheil, I., & Klingberg, T. (2012). Brain activity during a visuospatial working memory task predicts arithmetical performance 2 years later. *Cerebral Cortex*, *22*(5), 1078–1085. https://doi.org/10.1093/cercor/bhr175

Duncan, G., & Magnuson, K. (2011). The Nature and Impact of Early Achievement Skills, Attention Skills, and Behavior Problems. In *Whither Opportunity: Rising Inequality, Schools, and Children's Life Chances* (Issue 0322356, pp. 47–69). Russell Sage.

Duncan, Johnson, R., Swales, M., & Freer, C. (1997). Frontal Lobe Deficits after Head Injury : Unity and Diversity of Function Frontal Lobe. *Cognitive Neuropsychology*, *14*(5), 713–741. https://doi.org/10.1080/026432997381420

Duprez, J., Houvenaghel, J.-F., Naudet, F., Dondaine, T., Auffret, M., Robert, G., Drapier, D., Argaud, S., Vérin, M., & Sauleau, P. (2016). Evaluating Cognitive Action Control Using Eye-Movement Analysis: An Oculomotor Adaptation of the Simon Task. *Frontiers in Human Neuroscience*, *10*, 1–8. https://doi.org/10.3389/fnhum.2016.00084

Durston, S., Thomas, K. M., Yang, Y., Ulu, A. M., Zimmerman, R. D., & Casey, B. J. (2002). A neural basis for the development of inhibitory control. *Developmental Science*, *5*(4), 9–16. https://doi.org/10.1111/1467-7687.00235

Dutilh, G., Ravenzwaaij, D. Van, Nieuwenhuis, S., van der Maas, H. L. J., Forstmann, B. U., & Wagenmakers, E. (2012). How to measure post-error slowing : A confound and a simple solution. *Journal of Mathematical Psychology*, *56*(3), 208–216. https://doi.org/10.1016/j.jmp.2012.04.001

Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E. (2012). Testing theories of post-error slowing. *Attention,* *Perception, and Psychophysics*, *74*, 454–465. https://doi.org/10.3758/s13414-011-0243-2

Dweck, C. S. (2008). *Mindset the New Psychology of Success*. Random House Digital, Inc.

Eggen, T. J. H. M., & Verschoor, A. J. (2006). Optimal Testing With Easy or Difficult Items in Computerized Adaptive Testing. *Applied Psychological Measurement*, *30*(5), 379–393. https://doi.org/10.1177/0146621606288890

Elliot, A. J., & Dweck, C. S. (2013). *Handbook of competence and motivation*. Guilford Publications.

Elliott, J. G., Gathercole, S. E., Alloway, T. P., Holmes, J., & Kirkwood, H. (2010). An Evaluation of a Classroom-Based Intervention to Help Overcome Working Memory Difficulties and Improve Long-Term Academic Achievement. *Journal of Cognitive Education and Psychology*, *9*(3), 227–250. https://doi.org/10.1891/1945-8959.9.3.227

Elliott, L., & Bachman, H. J. (2018). SES disparities in early math abilities : The contributions of parents ' math cognitions , practices to support math , and math talk. *Developmental Review*, *49*, 1–15. https://doi.org/10.1016/j.dr.2018.08.001

Enge, S., Behnke, A., Fleischhauer, M., Küttler, L., Kliegel, M., & Strobel, A. (2014). No Evidence for True Training and Transfer Effects After Inhibitory Control Training in Young Healthy Adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 987–1001. https://doi.org/10.1037/a0036165

Erb, C. D. (2018). The developing mind in action: measuring manual dynamics in childhood. *Journal of Cognition and Development*, *19*(3), 233–247. https://doi.org/10.1080/15248372.2018.1454449

Eriksen, B., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception and Psychophysics*, *16*(1), 143–149. https://doi.org/10.3758/bf03203267

Espy, K. A., McDiarmid, M. M., Cwik, M. F., Meade Stalers, M., Hamby, A., & Senn, T. E. (2004). The Contribution of Executive Functions to Emergent Mathematic Skills in Preschool Children. *Developmental Neuropsychology*, *26*(1), 465–486. https://doi.org/10.1207/s15326942dn2601

Essa, A., & Laster, S. (2017). Bloom's 2 sigma problem and data-driven approaches for improving student success. In *The first year of college: Research, theory, and* *practice on improving the student experience and increasing retention* (pp. 212–246).

European Schoolnet, & Liege, U. of. (2013). *Survey of schools: ICT in education. Country profile: Netherlands*.

European Schoolnet, U. of L. (2013). *Survey of Schools: ICT in Education, Benchmarking Access, Use and Attitudes to Technology in Europe's Schools, Final Study Report*. https://doi.org/10.2759/94499

Eylon, B., & Linn, M. C. (1988). *Learning and Instruction : An Examination of Four Research Perspectives in Science*. *58*(3), 251–301.

Facer, K. (2011). Learning futures: Education, technology and social change. In *Learning futures: Education, technology and social change*. https://doi.org/10.4324/9780203817308

Fairweather, H. (1978). Choice Reaction Times in Children : Error and Post-error Responses , and the Repetition Effect. *Journal of Experimental Child Psychology*, *26*, 407–418. https://doi.org/10.1016/0022-0965(78)90121-2

Faulkenberry, T. J. (2014). Hand Movements Reflect Competitive Processing in Numerical Cognition. *Canadian Journal of Experimental Psychology*, *3*, 147–151. https://doi.org/10.1037/cep0000021

Faulkenberry, T. J., Cruise, A., Lavro, D., & Shaki, S. (2016). Response trajectories capture the continuous dynamics of the size congruity effect. *Acta Psychologica*, *163*, 114–123. https://doi.org/10.1016/j.actpsy.2015.11.010

Faulkenberry, T. J., Witte, M., & Hartmann, M. (2018). Tracking the Continuous Dynamics of Numerical Processing : A Brief Review and Editorial. *Journal of Numerical Cognition*, *4*(2), 271–285. https://doi.org/10.5964/jnc.v4i2.179

Federico, P. A. (2000). Learning styles and student attitudes toward various aspects of network-based instruction. *Computers in Human Behavior*, *16*(4), 359–379. https://doi.org/10.1016/S0747-5632(00)00021-2

Ferdinand, N. K., & Kray, J. (2014). Developmental changes in performance monitoring : How electrophysiological data can enhance our understanding of error and feedback processing in childhood and adolescence. *Behavioural Brain Research*, *263*, 122–132. https://doi.org/10.1016/j.bbr.2014.01.029

Ferla, J., Valcke, M., & Schuyten, G. (2010). Judgments of self-perceived academic competence and their differential impact on students ' achievement motivation , learning approach , and academic performance. *European Journal of*

Psychological Education, *25*, 519–536. https://doi.org/10.1007/s10212-010-0030-9

Ficarella, S. C., Rochet, N., & Burle, B. (2019). Becoming aware of subliminal responses : An EEG / EMG study on partial error detection and correction in humans. *Cortex*, *120*, 443–456. https://doi.org/10.1016/j.cortex.2019.07.007

Fischer, M. H., & Hartmann, M. (2014). Pushing forward in embodied cognition: May we mouse the mathematical mind? *Frontiers in Psychology*, *5*, 1–4. https://doi.org/10.3389/fpsyg.2014.01315

Fisher, A. V., Godwin, K. E., & Seltman, H. (2014). Visual Environment, Attention Allocation, and Learning in Young Children: When Too Much of a Good Thing May Be Bad. *Psychological Science*, *25*(7), 1362–1370. https://doi.org/10.1177/0956797614533801

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2007). *Learning disabilities: from Identification to Intervention*. Guilford Publications.

Franco-Watkins, A. M., & Johnson, J. G. (2011). Applying the decision moving window to risky choice: Comparison of eye-tracking and mouse-tracing methods. *Judgment and Decision Making*, *6*(8), 740–749. https://doi.org/http://dx.doi.org/10.1016/B978-012369442-3/50198-7

Freedman, J. L., & Edwards, D. R. (1988). Time pressure, task performance, and enjoyment. In J. E. McGrath (Ed.), *The social psychology of time* (pp. 113–133). Sage Publications.

Freeman, J. B. (2018). Doing psychological science by hand. *Current Directions in Psychological Science*, *27*(5), 315–323. https://doi.org/10.1177/0963721417746793

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226–241. https://doi.org/10.3758/BRM.42.1.226

Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*, 1–6. https://doi.org/10.3389/fpsyg.2011.00059

Friedman, N. P., & Miyake, A. (2017). Review Unity and diversity of executive functions : Individual differences as a window on cognitive structure. *Cortex*, *86*, 186–204. https://doi.org/10.1016/j.cortex.2016.04.023

Friso-van den Bos, I., van der Ven, S. H. G., Kroesbergen, E. H., & van Luit, J. E. H.

(2013). Working memory and mathematics in primary school children : A meta-analysis. *Educational Research Review*, *10*, 29–44. https://doi.org/10.1016/j.edurev.2013.05.003

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology*, *63*, 287–313. https://doi.org/10.1146/annurev-psych-120710-100449

Garon, N., Bryson, S. E., & Smith, I. M. (2008). Executive Function in Preschoolers : A Review Using an Integrative Framework. *Psychological Bulletin*, *134*(1), 31–60. https://doi.org/10.1037/0033-2909.134.1.31

Gärtner, A., & Strobel, A. (2019). *Individual differences in inhibitory control : A latent variable analysis*. https://psyarxiv.com/gnhmt

Gathercole, S. E., Lamont, E. M. I. L. Y., & Alloway, T. P. (2008). Working memory in the classroom. In *Working memory and learning: A practical guide for teachers* (pp. 219–240). Sage. https://doi.org/https://doi.org/10.1016/B978-012554465-8/50010-7

Gathercole, S. E., & Pickering, S. J. (2000). Working memory deficits in children with low achievements in the national curriculum at 7 years of age. *British Journal of Educational Psychology*, *70*, 177–194. https://doi.org/10.1037//0022-0663.92.2.377

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The Structure of Working Memory From 4 to 15 Years of Age. *Developmental Psychology*, *40*(2), 177–190. https://doi.org/10.1037/0012-1649.40.2.177

Geary, D. C. (2004). Mathematics and Learning Disabilities. *Journal of Learning Disabilities*, *37*(1), 4–15.

Geary, D. C., & Brown, S. C. (1991). Cognitive Addition: Strategy Choice and Speed-of-Processing Differences in Gifted, Normal, and Mathematically Disabled Children. *Developmental Psychology*, *27*(3), 398–406. https://doi.org/10.1037/0012-1649.27.3.398

Geary, D. C., Frensch, P. A., & Wiley, J. G. (1993). Simple and Complex Mental Subtraction: Strategy Choice and Speed-of-Processing Differences in Younger and Older Adults. *Psychology and Aging*, *8*(2), 242–256. https://doi.org/10.1037/0882-7974.8.2.242

Gierasimczuk, N., van der Maas, H. L. J., & Raijmakers, M. E. J. (2013). An Analytic Tableaux Model for Deductive Mastermind Empirically Tested with a Massively

Used Online Learning System. *Journal of Logic, Language and Information*, *22*(3). https://doi.org/10.1007/s10849-013-9177-5

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*, 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gomez, A., Piazza, M., Jobert, A., Dehaene-Lambertz, G., & Huron, C. (2017). Numerical abilities of school-age children with Developmental Coordination Disorder (DCD): A behavioral and eye-tracking study. *Human Movement Science*, *55*, 315–326. https://doi.org/10.1016/J.HUMOV.2016.08.008

Gopalan, V., Aida, J., Bakar, A., Nasir, A., Alwi, A., & Mat, R. C. (2017). A review of the motivation theories in learning. *AIP Conference Proceedings*, *1891*(1).

Grammer, J. K., Carrasco, M., Gehring, W. J., & Morrison, F. J. (2014). Age-related changes in error processing in young children : A school-based investigation. *Developmental Cognitive Neuroscience*, *9*, 93–105. https://doi.org/10.1016/j.dcn.2014.02.001

Grondin, S. (2010). Timing and time perception: A review of recent behavioral and neuroscience findings and theoretical directions. *Attention, Perception, & Psychophysics*, *72*(3), 561–582. https://doi.org/10.3758/APP

Grützmann, R., Endrass, T., Klawohn, J., & Kathmann, N. (2014). Response accuracy rating modulates ERN and Pe amplitudes. *Biological Psychology*, *96*, 1–7. https://doi.org/10.1016/j.biopsycho.2013.10.007

Guerra-Carrillo, B. C., & Bunge, S. A. (2018). Eye gaze patterns reveal how reasoning skills improve with experience. *Npj Science of Learning*, *3*(19), 1–9. https://doi.org/10.1038/s41539-018-0035-8

Guo, Q., & Agichtein, E. (2010). Towards predicting web searcher gaze position from mouse movements. *Conference on Human Factors in Computing Systems - Proceedings*, 3601–3606. https://doi.org/10.1145/1753846.1754025

Gupta, R., Kar, B. R., & Srinivasan, N. (2009). Development of task switching and post-error-slowing in children. *Behavioral and Brain Functions*, *5*(1), 38. https://doi.org/10.1186/1744-9081-5-38

Hailikari, T., Katajavuori, N., & Lindblom-Ylanne, S. (2008). The Relevance of Prior Knowledge in Learning and Instructional Design. *American Journal of Pharmaceutical Education*, *72*(5), 1–8.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox : Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*, 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hehman, E., Stolier, R. M., & Freeman, J. B. (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes and Intergroup Relations*, *18*(3), 384–401. https://doi.org/10.1177/1368430214538325

Henderson, J. M., & Smith, T. J. (2009). How are eye fixation durations controlled during scene viewing? Further evidence from a scene onset delay paradigm. *Visual Cognition*, *17*(6–7), 1055–1082. https://doi.org/10.1080/13506280802685552

Hinault, T., & Lemaire, P. (2016). What does EEG tell us about arithmetic strategies? A review. *International Journal of Psychophysiology*, *106*, 115–126. https://doi.org/10.1016/j.ijpsycho.2016.05.006

Hofman, A. D., Brinkhuis, M. J. S., Bolsinova, M., Klaiber, J., Maris, G. K. J., & van der Maas, H. L. J. (2020). Tracking with (Un)Certainty. *Journal of Intelligence*, *8*(10), 1–15. https://doi.org/10.3390/jintelligence8010010

Hofman, A. D., Jansen, B. R. J., de Mooij, S. M. M., Stevenson, C. E., & van der Maas, H. L. J. (2018). A solution to the measurement problem in the idiographic approach using computer adaptive practicing. *Journal of Intelligence*, *6*(1), 14. https://doi.org/10.3390/jintelligence6010014

Hofman, A. D., Kievit, R. A., Stevenson, C. E., Molenaar, D., & van der Maas, H. L. J. (2018). *The dynamics of the development of mathematics skills : A comparison of theories of developing intelligence*. https://doi.org/10.31219/osf.io/xa2ft

Hofman, A. D., Visser, I., Jansen, B. R. J., Marsman, M., & van der Maas, H. L. J. (2018). Fast and slow strategies in multiplication. *Learning and Individual Differences*, *68*(406), 30–40. https://doi.org/10.1016/j.lindif.2018.09.007

Holmqvist, K., & Nyström, M. (2011). *Eye tracking: a comprehensive guide to methods and measures*. Oxford University Press.

Houtman, F., Castellar, E. N., & Notebaert, W. (2012). Orienting to errors with and without immediate feedback. *Journal of Cognitive Psychology*, *24*(3), 278–285. https://doi.org/10.1080/20445911.2011.617301

Howard-Jones. (2011). *The impact of digital technologies on human wellbeing: Evidence from the sciences of mind and brain*. Nominet Trust.

Huang, H., & Mercer, V. S. (2001). Dual-Task Methodology : Applications in Studies of Cognitive and Motor Performance in Adults and Children. *Pediatric Physical Therapy*, 133–140.

Huang, J., & White, R. (2012). User see, user point: gaze and cursor alignment in web search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1341–1350. https://doi.org/10.1145/2207676.2208591

Huang, W. H. (2011). Evaluating learners' motivational and cognitive processing in an online game-based learning environment. *Computers in Human Behavior*, *27*(2), 694–704. https://doi.org/10.1016/j.chb.2010.07.021

Huebner, M. G., & LeFevre, J. A. (2018). Selection of procedures in mental subtraction: Use of eye movements as a window on arithmetic processing. *Canadian Journal of Experimental Psychology*, *72*(3), 171–182. https://doi.org/10.1037/cep0000127

Hui, K., & Lee, K. (2009). Inhibiting interference from prior knowledge : Arithmetic intrusions in algebra word problem solving. *Learning and Individual Differences*, *19*(2), 262–268. https://doi.org/10.1016/j.lindif.2009.01.004

Huizinga, M., Dolan, C. V., & van der Molen, M. W. (2006). Age-related change in executive function: Developmental trends and a latent variable analysis. *Neuropsychologia*, *44*(11), 2017–2036. https://doi.org/10.1016/j.neuropsychologia.2006.01.010

Huizinga, M., & van der Molen, M. W. . (2011). Task switching and shifting between stopping and going : Developmental change in between-trial control adjustments. *Journal of Experimental Child Psychology*, *108*, 484–503. https://doi.org/10.1016/j.jecp.2010.10.004

Hyönä, J. (2010). The use of eye movements in the study of multimedia learning. *Learning and Instruction*, *20*(2), 172–176. https://doi.org/10.1016/j.learninstruc.2009.02.013

Imbo, I., & Vandierendonck, A. (2007). The development of strategy use in elementary school children : Working memory and individual differences. *Journal of Experimental Child Psychology*, *96*, 284–309. https://doi.org/10.1016/j.jecp.2006.09.001

Incera, S., & McLennan, C. T. (2016). Mouse tracking reveals that bilinguals behave like experts. *Bilingualism*, *19*(3), 610–620. https://doi.org/10.1017/S1366728915000218

Intayoad, W., Kamyod, C., & Temdee, P. (2020). Reinforcement Learning Based on

Contextual Bandits for Personalized Online Learning Recommendation Systems. *Wireless Personal Communications*, *115*(4), 2917–2932. https://doi.org/10.1007/s11277-020-07199-0

Jansen, B. R. J., Hofman, A. D., Straatemeier, M., van Bers, B. M. C. W., Raijmakers, M. E. J., & van der Maas, H. L. J. (2014). The role of pattern recognition in children's exact enumeration of small numbers. *British Journal of Developmental Psychology*, *32*(2), 178–194. https://doi.org/10.1111/bjdp.12032

Jansen, B. R. J., Louwerse, J., Straatemeier, M., van der Ven, S. H. G., Klinkenberg, S., & van der Maas, H. L. J. (2013). The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and Individual Differences*, *24*, 190–197. https://doi.org/10.1016/j.lindif.2012.12.014

JASP Team. (2019). *JASP (Version 0.8. 6)[computer software]*. (0.10).

Jones, L. B., Rothbart, M. K., & Posner, M. I. (2003). Development of executive attention in preschool children. *Developmental Science*, *6*(5), 498–504. https://doi.org/https://doi.org/10.1111/1467-7687.00307

Jonker, T. R., Seli, P., Cheyne, J. A., & Smilek, D. (2013). Performance reactivity in a continuous-performance task: Implications for understanding post-error behavior. *Consciousness and Cognition*, *22*(4), 1468–1476. https://doi.org/10.1016/j.concog.2013.10.005

Joy, S., & Kolb, D. A. (2009). Are there cultural differences in learning style? *International Journal of Intercultural Relations*, *33*, 69–85. https://doi.org/10.1016/j.ijintrel.2008.11.002

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*(4), 441–480.

Kardefelt-Winther, D. (2017). *How does the time children spend using digital technology impact their mental well-being, social relationships and physical activity?* https://doi.org/10.1080/0300443951060119

Karr, J. E., Areshenkoff, C. N., Rast, P., Hofer, S. M., Iverson, G. L., & Garcia-Barrera, M. A. (2019). The Unity and Diversity of Executive Functions: A Systematic Review and Re-Analysis of Latent Variable Studies. *Psychological Bulletin*, *144*(11), 1147–1185. https://doi.org/10.1037/bul0000160.The

Kasprowski, P., & Harezlak, K. (2018). Biometric identification using gaze and mouse dynamics during game playing. *Communications in Computer and Information*

Science, *928*, 494–504. https://doi.org/10.1007/978-3-319-99987-6_38

Kassai, R., Futo, J., & Demetrovics, Z. (2019). A Meta-Analysis of the Experimental Evidence on the Near- and Far-Transfer Effects Among Children ' s Executive Function Skills. *Psychological Bulletin*, *145*(2), 165–188.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology*, *62*(4), 498–525.

Kebritchi, M., Hirumi, A., & Bai, H. (2010). The effects of modern mathematics computer games on mathematics achievement and class motivation. *Computers and Education*, *55*(2), 427–443. https://doi.org/10.1016/j.compedu.2010.02.007

Kellogg, J. S., Hopko, D. R., & Ashcraft, M. H. (1999). The effects of time pressure on arithmetic performance. *Journal of Anxiety Disorders*, *13*(6), 591–600. https://doi.org/10.1016/S0887-6185(99)00025-0

Kieslich, P. J., & Henninger, F. (2017). Mousetrap: An integrated, open-source mouse-tracking package. *Behavior Research Methods*, *49*(5), 1652–1667. https://doi.org/10.3758/s13428-017-0900-z

Kiili, K. (2005). Digital game-based learning: Towards an experiential gaming model. *Internet and Higher Education*, *8*(1), 13–24. https://doi.org/10.1016/j.iheduc.2004.12.001

Kirkham, N. Z., Cruess, L., & Diamond, A. (2003). Helping children apply their knowledge to their behavior on a dimension-switching task. *Developmental Science*, *6*(5), 449–467. https://doi.org/10.1111/1467-7687.00300

Klem, A. M., & Connell, J. P. (2004). Relationships Matter : Linking Teacher Support to Student Engagement and Achievement. *Journal of School Health*, *74*(7), 27–34.

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers and Education*, *57*(2), 1813–1824. https://doi.org/10.1016/j.compedu.2011.02.003

Knewton. (2012). *Knewton Adaptive Learning*. www.knewton.com

Koedinger, K. R., & Corbett, A. T. (2012). The Knowledge-Learning-Instruction Framework : Bridging the Science-Practice Chasm to Enhance Robust Student Learning. *Cognitive Science*, *36*, 757–798. https://doi.org/10.1111/j.1551-6709.2012.01245.x

Koop, G. J., & Johnson, J. G. (2013). The response dynamics of preferential choice. *Cognitive Psychology*, *67*(4), 151–185. https://doi.org/10.1016/j.cogpsych.2013.09.001

Krassanakis, V., & Kesidis, A. L. (2020). MatMouse : A Mouse Movements Tracking and Analysis Toolbox for Visual Search Experiments. *Multimodal Technologies and Interaction*, *4*(83). https://doi.org/10.3390/mti4040083

Kray, J., Eber, J., & Lindenberger, U. (2004). Age differences in executive functioning across the lifespan : The role of verbalization in task preparation. *Acta Psychologica*, *115*, 143–165. https://doi.org/10.1016/j.actpsy.2003.12.001

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package : tests in linear mixed effects. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lai, K., & Hong, K. (2015). Technology use and learning characteristics of students in higher education: Do generational differences exist? *British Journal of Educational Technology*, *46*(4), 725–738. https://doi.org/10.1111/bjet.12161

Lai, M., Tsai, M. J., Yang, F. Y., Hsu, C. Y., Liu, T. C., Lee, S. W. Y., Lee, M. H., Chiou, G. L., Liang, J. C., & Tsai, C. C. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review*, *10*(88), 90–115. https://doi.org/10.1016/j.edurev.2013.10.001

Laming, D. (1979). Choice reaction performance following an error. *Acta Psychologica*, *43*, 199–224. https://doi.org/10.1016/0001-6918(79)90026-x

Lappin, J. S., & Eriksen, C. W. (1966). Use of a delayed signal to stop a visual reaction-time response. *Journal of Experimental Psychology*, *72*(6), 805–811. https://doi.org/10.1037/h0021266

Larson, R. W., & Rusk, N. (2011). Instrinsic motivation and positive development. In *Positive Youth Development* (pp. 89–130). Elsevier Inc. https://doi.org/10.1016/B978-0-12-386492-5.00005-1

Lavro, D., Ben-Shachar, M. S., Saville, C. W. N., Klein, C., & Berger, A. (2018). Testing the bottleneck account for post-error slowing beyond the post-error response. *Biological Psychology*, *138*, 81–90. https://doi.org/10.1016/j.biopsycho.2018.08.010

Lavro, D., Levin, D., Klein, C., & Berger, A. (2018). Response time distribution parameters show posterror behavioral adjustment in mental arithmetic. *Acta Psychologica*, *186*, 8–17. https://doi.org/10.1016/j.actpsy.2018.04.001

Leblanc, H., & Ramirez, S. (2020). Linking social cognition to learning and memory. *Journal of Neuroscience*, *40*(46), 8782–8798. https://doi.org/10.1523/JNEUROSCI.1280-20.2020

Lefevre, J., Sadesky, G. S., & Bisanz, J. (1996). Selection of Procedures in Mental Addition : Reassessing the Problem Size Effect in Adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 216–230.

Lemaire, P. (2010). Executive Functions and Strategic Aspects of Arithmetic Performance: The Case of Adults' and Children's Arithmetic Patrick Lemaire. *Psychologica Belgica*, *50*, 335–352.

Lemaire, P., & Siegler, R. S. (1995). Four Aspects of Strategic Change: Contributions to Children's Learning of Multiplication. *Journal of Experimental Psychology: General*, *124*(1), 83–97. https://doi.org/10.1037/0096-3445.124.1.83

Lewis, B. P., & Linder, D. E. (1997). Thinking about choking? Attentional processes and paradoxical performance. *Personality and Social Psychology Bulletin*, *23*(9), 937–944.

Lim, D. H. (2004). Cross Cultural Differences in Online Learning Motivation. *Educational Media International*, *41*(2), 163–175. https://doi.org/10.1080/09523980410001685784

Lyons, K. E., & Zelazo, P. D. (2011). Monitoring, metacognition, and executive function: elucidating the role of self-reflection in the development of self-regulation. In J. B. Benson (Ed.), *Advances in Child Development and Behavior* (1st ed., Vol. 40). Elsevier Inc. https://doi.org/10.1016/B978-0-12-386491-8.00010-4

MacLeod, C. M. (1992). The stroop task: the gold standard of attentional measures. *Journal of Experimental Psychology*, *121*(1), 12–14.

Mandich, A., Buckolz, E., & Polatajko, H. (2003). Children with developmental coordination disorder (DCD) and their ability to disengage ongoing attentional focus : More on inhibitory function. *Brain and Cognition*, *51*, 346–356. https://doi.org/10.1016/S0278-2626(03)00039-3

Manor, O., & Gross-tsur, R. S. S. V. (2000). Developmental dyscalculia : prevalence and prognosis. *European Child & Adolescent Psychiatry*, 58–64.

Marco-Pallarés, J., Camara, E., Münte, T. F., & Rodríguez-Fornells, A. (2008). Neural mechanisms underlying adaptive actions after slips. *Journal of Cognitive Neuroscience*, *20*(9), 1595–1610. https://doi.org/10.1162/jocn.2008.20117

Marghetis, T., Núñez, R., & Bergen, B. K. (2014). Doing arithmetic by hand: Hand movements during exact arithmetic reveal systematic, dynamic spatial processing. *Quarterly Journal of Experimental Psychology*, *67*(8), 1579–1596. https://doi.org/10.1080/17470218.2014.897359

Maris, G., & van der Maas, H. L. J. (2012). Speed-accuracy response models: scoring rules based on response time and accuracy. *Psychometrika*, *77*(4), 615–633. https://doi.org/10.1007/s11336-012-9288-y

Massonnié, J., Rogers, C. J., Mareschal, D., & Kirkham, N. Z. (2019). Is classroom noise always bad for children? The contribution of age and selective attention to creative performance in noise. *Frontiers in Psychology*, *10*(FEB), 1–12. https://doi.org/10.3389/fpsyg.2019.00381

Matthews, W. J., & Meck, W. H. (2016). Temporal Cognition : Connecting Subjective Time to Perception , Attention , and Memory. *Psychological Bulletin*, *142*(8), 865–907.

Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning. *Journal of Educational Psychology*, *90*(2), 312–320.

Mazzocco, M. M. M., & Kover, S. T. (2007). A Longitudinal Assessment of Executive Function Skills and Their Association with Math Performance. *Child Neuropsychology ISSN:*, *13*(1), 18–45. https://doi.org/10.1080/09297040600611346

Mcguire, J. T., & Botvinick, M. M. (2010). Prefrontal cortex , cognitive control , and the registration of decision costs. *PNAS*, *107*(17), 7922–7926. https://doi.org/10.1073/pnas.0910662107

McLean, D. J., & Skowron Volponi, M. A. (2018). trajr: An R package for characterisation of animal trajectories. *Ethology*, *124*(6), 440–448. https://doi.org/10.1111/eth.12739

Meaux, J. B., & Chelonis, J. J. (2005). The Relationship between Behavioral Inhibition and Time Perception in Children. *Journal of Child and Adolescent Psychiatric Nursing*, *18*(42), 148–160. https://doi.org/10.1057/978-1-137-40883-9_7

Mendl, M. (1999). Performing under pressure: Stress and cognitive function. *Applied Animal Behaviour Science*, *65*(3), 221–244. https://doi.org/10.1016/S0168-1591(99)00088-X

Michelot, T., Langrock, R., & Patterson, T. (2019). moveHMM An R package for the analysis of animal movement data. *ArXiv Preprint*, 1–24.

https://doi.org/https://doi.org/10.1111/2041-210x.12578

Milisavljevic, A., Hamard, K., Petermann, C., Gosselin, B., & Dor, K. (2018). Eye and Mouse Coordination During Task : From Behaviour to Prediction. *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, *2*, 86–93.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Moeller, K., Fischer, U., Cress, U., & Nuerk, H. (2012). Diagnostics and Intervention in Developmental Dyscalculia: Current Issues and Novel Perspectives. In *Reading, Writing, Mathematics and the developing brain: listening to many voices* (pp. 233–277). Springer.

Molenaar, P. C. M. (2004). A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Montez, D. F., Calabro, F. J., & Luna, B. (2017). The expression of established cognitive brain states stabilizes with working memory development. *ELife*, *6*, 1–26. https://doi.org/10.7554/eLife.25606

Moreno, R., & Mayer, R. E. (2003). Nine Ways to Reduce Cognitive Load in Multimedia Learning Richard. *Educational Psychologist*, *38*(1), 43–52. https://doi.org/10.1207/S15326985EP3801

Moser, J. S., Schroder, H. S., Heeter, C., Moran, T. P., & Lee, Y. (2011). Mind Your Errors : Evidence for a Neural Mechanism Linking Growth Mind-Set to Adaptive Posterror Adjustments. *Psychological Science*, *22*(12), 484–489. https://doi.org/10.1177/0956797611419520

Moyer, P. S. (2001). Are We Having Fun Yet? How Teachers Use Manipulatives to Teach Mathematics. *Educational Studies in Mathematics*, *47*(2), 175–197.

Müller, U., Liebermann, D., Frye, D., & Zelazo, P. D. (2008). Executive function, school readiness, and school achievement. In S. K. Thurman & C. A. Fiorello (Eds.), *Applied Cognitive Research in K-3 Classrooms* (pp. 41–84). Routledge.

Murray, T., & Arroyo, I. (2002). Toward measuring and maintaining the zone of proximal development in adaptive instructional systems. *International*

Conference on Intelligent Tutoring Systems. https://doi.org/10.1007/3-540-47987-2_75

Niendam, T. A., Laird, A. R., Ray, K. L., Dean, Y. M., Glahn, D. C., & Carter, C. S. (2012). *Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions*. 241–268. https://doi.org/10.3758/s13415-011-0083-5

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J. O. S., Band, G. P. H., & Kok, A. (2001). *Error-related brain potentials are differentially related to awareness of response errors : Evidence from an antisaccade task*. 752–760.

Nisbett, R. (2004). *The geography of thought: How Asians and Westerners think differently... and why*. Simon and Schuster.

Norman, D. A. (1981). Categorization of Action Slips. *Psychological Review*, *88*(1), 1–15.

Notebaert, W., Houtman, F., Opstal, F. Van, Gevers, W., Fias, W., & Verguts, T. (2009). Post-error slowing: An orienting account. *Cognition*, *111*(2), 275–279. https://doi.org/10.1016/j.cognition.2009.02.002

Núñez-peña, M. I., Tubau, E., & Suárez-pellicioni, M. (2017). Post-error response inhibition in high math-anxious individuals: Evidence from a multi-digit addition task. *Acta Psychologica*, *177*, 17–22. https://doi.org/10.1016/j.actpsy.2017.04.002

Núñez Castellar, E., van Looy, J., Szmalec, A., & de Marez, L. (2014). Improving arithmetic skills through gameplay : Assessment of the effectiveness of an educational game in terms of cognitive and affective learning outcomes. *Information Sciences*, *264*, 19–31. https://doi.org/10.1016/j.ins.2013.09.030

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257.

OECD. (2020). Education in the Digital Age. In *Educational Research and Innovation* (Vol. 1, Issue 3). https://www.oecd-ilibrary.org/education/education-in-the-digital-age_1209166a-en

Ofcom. (2020). *Children and Parents: Media Use and Attitudes report 2019*. http://stakeholders.ofcom.org.uk/binaries/research/media-literacy/october-2013/research07Oct2013.pdf

Ormrod, J. (2014). *Human learning*. Pearson.

Orquin, J. L., & Loose, S. M. (2013). Attention and choice: A review on eye movements in decision making. *Acta Psychologica*, *144*(1), 190–206. https://doi.org/10.1016/j.actpsy.2013.06.003

Ou, L., Hofman, A. D., Simmering, V. R., Bechger, T., Maris, G., & van der Maas, H. L. J. (2019). Modeling person-specific development of math skills in continuous time: New evidence for mutualism. *The 12th International Conference on Educational Data Mining*, 372–377.

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology*, *66*(2), 232–258. https://doi.org/10.1016/j.cogpsych.2012.12.002

Palfrey, J., Gasser, U., Simun, M., & Barnes, R. F. (2009). Youth, Creativity, and Copyright in the Digital Age. *International Journal of Learning & Media*, *1*(2), 79–97. https://doi.org/10.1162/ijlm.2009.0022

Papoutsaki, A., & Laskey, J. (2016). WebGazer : Scalable Webcam Eye Tracking Using User Interactions. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence-IJCAI 2016*.

Park, O., & Lee, J. (2003). Adaptive instructional systems. *Educational Technology Research and Development*, *1911*, 651–684.

Patterson, T. A., Basson, M., Bravington, M. V, & Gunn, J. S. (2009). Classifying movement behaviour in relation to environmental conditions using hidden Markov models. *Journal of Animal Ecology*, *78*, 1113–1123. https://doi.org/10.1111/j.1365-2656.2009.01583.x

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3).

Peirce, J. W. (2007). PsychoPy — Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Peng, A., Kirkham, N. Z., & Mareschal, D. (2018). Information processes of task-switching and modality-shifting across development. *PLoS ONE*, *13*(6), 1–18. https://doi.org/10.1371/journal.pone.0198870 June

Peng, Namkung, J., Barnes, M., & Sun, C. (2015). A Meta-Analysis of Mathematics and Working Memory: Moderating Effects of Working Memory Domain, Type of Mathematics Skill, and Sample Characteristics. *Journal of Educational Psychology*, *108*(4). https://doi.org/10.1037/edu0000079.supp

Peters, L., & de Smedt, B. (2018). Arithmetic in the developing brain : A review of brain imaging studies. *Developmental Cognitive Neuroscience*, *30*, 265–279. https://doi.org/10.1016/j.dcn.2017.05.002

Piaget, J. (1964). Cognitive development in children: development and learning. *Journal of Research in Science Teaching*, *2*, 176–186.

Plowman, L., & McPake, J. (2013). Seven Myths About Young Children and Technology. *Childhood Education*, *89*(1), 27–33. https://doi.org/10.1080/00094056.2013.757490

Prensky, M. (2001). Digital Natives , Digital Immigrants. *On the Horizon*, *9*(5), 1–6.

Przybylski, A. K., & Weinstein, N. (2017). A Large-Scale Test of the Goldilocks Hypothesis: Quantifying the Relations Between Digital-Screen Use and the Mental Well-Being of Adolescents. *Psychological Science*, *28*(2), 204–215. https://doi.org/10.1177/0956797616678438

Rabbitt, P., & Rodgers, B. (1977). What does a man do after he makes an error ? an analysis of response programming. *The Quarterly Journal of Experimental Psychology*, *29*(4), 727–743. https://doi.org/10.1080/14640747708400645

Raghubar, K. P., Barnes, M. A., & Hecht, S. A. (2010). Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches. *Learning and Individual Differences*, *20*(2), 110–122. https://doi.org/10.1016/j.lindif.2009.10.005

Ramsay, J., & Terras, M. M. (2015). The pendulum swing of user instruction and interaction: The resurrection of 'how to use' technology to learn in the 21st century. *E-Learning and Digital Media*, *12*(3–4), 372–390. https://doi.org/10.1177/2042753015571827

Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, *9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067

Regev, S., & Meiran, N. (2014). Post-error slowing is influenced by cognitive control demand. *Acta Psychologica*, *152*, 10–18. https://doi.org/10.1016/j.actpsy.2014.07.006

Reimers, S., & Maylor, E. A. (2005). Task Switching Across the Life Span : Effects of Age on General and Specific Switch Costs. *Developmental Psychology*, *41*(4), 661–671. https://doi.org/10.1037/0012-1649.41.4.661

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The Role of the Medial Frontal Cortex in Cognitive Control. *Cognition and Behavior*, *306*, 443–448.

Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure : How opportunity costs affect strategy selection. *Acta Psychologica*, *127*, 258–276. https://doi.org/10.1016/j.actpsy.2007.05.004

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement : Evidence from Panel data. *The American Economic Review*, *94*(2), 247–252.

Rodden, K., View, M., Fu, X., View, M., Aula, A., View, M., & Spiro, I. (2008). Eye-Mouse Coordination Patterns on Web Search Results Pages. *CHI 2008 Proceedings*, 2997–3002.

Rose, J., & Awad, A. (2017). Biometric Authentication using Mouse and Eye Movement Data. *IEEE Symposium on Security and Privacy Workshops Biometric*, 47–55. https://doi.org/10.1109/SPW.2017.18

Rosen, B. L. D. (2010). *Understanding the iGeneration and the way they learn*. St. Martin's Press.

Rourke, B. P. (1993). Arithmetic disabilities, specific and otherwise: a neuropsychological perspective. *Journal of Learning Disabilities*, *26*(4), 214–226. https://doi.org/10.1080/13552600410001470973

Roy, P., Rutt, S., Easton, C., & Sims, D. (2019). *Stop and Think : Learning Counterintuitive Concepts*.

Saboundji, R. R. (2020). Predicting Human Errors from Gaze and Cursor Movements. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Santens, S., Goossens, S., & Verguts, T. (2011). Distance in motion: Response trajectories reveal the dynamics of number comparison. *PLoS ONE*, *6*(9). https://doi.org/10.1371/journal.pone.0025429

Sarrasin, J. B., Nenciovici, L., Foisy, L. M. B., Allaire-Duquette, G., Riopel, M., & Masson, S. (2018). Effects of teaching the concept of neuroplasticity to induce a growth mindset on motivation, achievement, and brain activity: A meta-analysis. *Trends in Neuroscience and Education*, *12*, 22–31. https://doi.org/10.1016/j.tine.2018.07.003

Sattizahn, J. R., Moser, J. S., & Beilock, S. L. (2016). A closer look at who "chokes under pressure." *Journal of Applied Research in Memory and Cognition*, *5*(4), 470–477. https://doi.org/10.1016/j.jarmac.2016.11.004

Savi, A. O., Deonovic, B. E., Bolsinova, M., van der Maas, H. L. J., & Maris, G. K. J.

(2018). *Automated Diagnosis of Misconceptions in Single Digit Multiplication*. https://doi.org/https://doi.org/10.31234/osf.io/53muj

Savi, A. O., Williams, J. J., & Maris, G. K. J. (2017). *The Role of A/B Tests in the Study of Large-Scale Online Learning*. https://doi.org/10.17605/OSF. IO/83JSG, 128

Schachar, R. J., Chen, S., Logan, G. D., Ornstein, T. J., Crosbie, J., Ickowicz, A., & Pakulak, A. (2004). Evidence for an Error Monitoring Deficit in Attention Deficit Hyperactivity Disorder. *Journal of Abnormal Child Psychology*, *32*(3), 285–293. https://doi.org/10.1023/b:jacp.0000026142.11217.f2

Scherbaum, S., & Kieslich, P. J. (2018). Stuck at the starting line: How the starting procedure influences mouse-tracking data. *Behavior Research Methods*, *50*(5), 2097–2110. https://doi.org/10.3758/s13428-017-0977-4

Schleifer, P., & Landerl, K. (2011). Subitizing and counting in typical and atypical development. *Developmental Science*, *14*(2), 280–291. https://doi.org/10.1111/j.1467-7687.2010.00976.x

Schmidt-weigand, F., Kohnert, A., & Glowalla, U. (2010). A closer look at split visual attention in system- and self-paced instruction in multimedia learning. *Learning and Instruction*, *20*(2), 100–110. https://doi.org/10.1016/j.learninstruc.2009.02.011

Schroder, H. S., Fisher, M. E., Lin, Y., Lo, S. L., Danovitch, J. H., & Moser, J. S. (2017). Neural evidence for enhanced attention to mistakes among school-aged children with a growth mindset. *Developmental Cognitive Neuroscience*, *24*, 42–50. https://doi.org/10.1016/j.dcn.2017.01.004

Schroder, H. S., Nickels, S., Cardenas, E., Breiger, M., Perlo, S., & Pizzagalli, D. A. (2019). Optimizing assessments of post-error slowing: A neurobehavioral investigation of a flanker task. *Psychophysiology*, 1–17. https://doi.org/10.1111/psyp.13473

Schulte-Mecklenbeck, M., Kuehberger, A., & Johnson, J. G. (2019). *A handbook of process tracing methods* (2nd ed.). Routledge. https://doi.org/https://doi.org/10.4324/9781315160559

Schunk, D. H. (2012). *Learning theories: an educational perspective*. Pearson.

Selwyn, N. (2011). Education and Technlogy: Key Issues and Debates. In *Media Education Research Journal*. Bloomsbury Publishing.

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science : A first look. *Behavior Research Methods*, *50*, 451–465.

https://doi.org/10.3758/s13428-017-0913-7
Shalev, R. S., Manor, O., Auerbach, J., & Gross-tsur, V. (1998). persistence of developmental dyscalculia: What counts? *The Journal of Pediatrics*, *133*(3), 358–362.
Shallice, T., & Burgess, P. (1996). The domain of supervisory processes and temporal organization of behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *351*(1346), 1405–1411.
Sieber, J. E., O'Neil, H. F., & Tobias, J. S. (1977). *Anxiety, Learning, and Instruction*. Routledge.
Siegler, R. S. (1988). Strategy choice procedures and the development of multiplication skill. *Journal of Experimental Psychology: General*, *117*(3), 258–275. https://doi.org/https://doi.org/10.1037/0096-3445.117.3.258
Siegler, R. S. (1989). Hazards of Mental Chronometry : An Example From Children's Subtraction. *Journal of Educational Psychology*, *81*(4), 497–506. https://doi.org/10.1037/0022-0663.81.4.497
Siegler, R. S. (1996). *Emerging minds: the process of change in children's thinking*. Oxford University Press.
Siegler, R. S., & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. *American Psychologist*, *46*(6), 606–620.
Siegler, R. S., & Lemaire, P. (1997). Older and younger adults' strategy choices in multiplication: testing predictions of ASCM using the choice/no-choice method. *Journal of Experimental Psychology: General*, *126*(1), 71–92.
Siegler, R. S., & Shrager, J. (1984). Strategy choices in addition and subtraction: How do children know what to do? *Origins of Cognitive Skills*, *January 1984*, 229–293.
Simon, J. R. (1969). Reactions toward the source of stimulation. *Journal of Experimental Psychology*, *81*(1), 174–176. https://doi.org/10.1037/h0027448
Skinner, E. A., Belmont, M. J., Lynch, M., Mellor-crummey, C., Miserandino, M., Patrick, B., Regan, C., Wellborn, J., & Usinger, P. (1993). Motivation in the Classroom : Reciprocal Effects of Teacher Behavior and Student Engagement Across the School Year. *Journal of Educational Psychology*, *85*(4), 571–581.
Skinner, E. A., Furrer, C., Marchand, G., Kindermann, T., Sherwood, H., & Usinger, P. (2008). *Engagement and Disaffection in the Classroom : Part of a Larger Motivational Dynamic ? 100*(4), 765–781. https://doi.org/10.1037/a0012840

Smulders, S. F. A., Soetens, E., & van der Molen, M. W. (2016). What happens when children encounter an error? *Brain and Cognition*, *104*, 34–47. https://doi.org/10.1016/j.bandc.2016.02.004

Snow, R. E., & Lohman, D. F. (1984). Toward a Theory of Cognitive Aptitude for Learning From Instruction '. *Journal of Educational Psychology*, *76*(3), 347–376.

Soares, N., Evans, T., & Patel, D. R. (2018). Specific learning disability in mathematics : a comprehensive review. *Translational Pediatrics*, *7*(1), 48–62. https://doi.org/10.21037/tp.2017.08.03

Song, J. H., & Nakayama, K. (2009). Hidden cognitive states revealed in choice reaching tasks. *Trends in Cognitive Sciences*, *13*(8), 360–366. https://doi.org/10.1016/j.tics.2009.04.009

Sorvo, R., Koponen, T., Viholainen, H., Aro, T., Räikkönen, E., Peura, P., Tolvanen, A., & Aro, M. (2019). Development of math anxiety and its longitudinal relationships with arithmetic achievement among primary school children. *Learning and Individual Differences*, *69*, 173–181. https://doi.org/10.1016/j.lindif.2018.12.005

Spivey, M. J. (2007). *The continuity of the mind*. Oxford University Press. https://doi.org/10.1093/acprof

Spivey, M. J., & Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, *15*(5), 207–211. https://doi.org/10.1111/j.1467-8721.2006.00437.x

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*(29), 10393–10398. https://doi.org/10.1073/pnas.0503903102

Spüler, M., Walter, C., Rosenstiel, W., Gerjets, P., Moeller, K., & Klein, E. (2016). EEG-based prediction of cognitive workload induced by arithmetic: a step towards online adaptation in numerical learning. *ZDM Mathematics Education*, *48*, 267–278. https://doi.org/10.1007/s11858-015-0754-8

St Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *Quarterly Journal of Experimental Psychology*, *59*(4), 745–759. https://doi.org/10.1080/17470210500162854

Starcke, K., & Brand, M. (2012). Decision making under stress: A selective review. *Neuroscience and Biobehavioral Reviews*, *36*(4), 1228–1248.

https://doi.org/10.1016/j.neubiorev.2012.02.003
Sternberg, R. J., & Kaufman, S. B. (2013). *The Cambridge handbook of Intelligence*. Cambridge university press.
Stillman, P. E., Shen, X., & Ferguson, M. J. (2018). How mouse-tracking can advance social cognitive theory. *Trends in Cognitive Sciences*, *22*(6), 531–543. https://doi.org/10.1016/j.tics.2018.03.012
Stojić, H., Orquin, Jacob, L., Dayan, P., Dolan, R. J., & Speekenbrink, M. (2020). Uncertainty in learning, choice, and visual fixation. *PNAS*, *117*(6), 3291–3300. https://doi.org/10.1073/pnas.1911348117
Straatemeier, M. (2014). *Math Garden: A new educational and scientific instrument*. http://hdl.handle.net/ 11245/1.417091
Susi, T., Johannesson, M., & Backlund, P. (2007). *Serious Games – An Overview*.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science: A Multidisciplinary Journal*, *12*(2), 257–285.
Tamnes, C. K., Walhovd, K. B., Torstveit, M., Sells, V. T., & Fjell, A. M. (2013). Performance monitoring in children and adolescents: A review of developmental changes in the error-related negativity and brain maturation. *Developmental Cognitive Neuroscience*, *6*, 1–13. https://doi.org/10.1016/j.dcn.2013.05.001
Team, R. C. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.r-project.org/
Terras, M. M., & Ramsay, J. (2012). The five central psychological challenges facing effective mobile learning. *British Journal of Educational Technology*, *43*(5), 820–832. https://doi.org/10.1111/j.1467-8535.2012.01362.x
Thompson, P. (2013). The digital natives as learners : Technology use patterns and approaches to learning. *Computers & Education*, *65*, 12–33. https://doi.org/10.1016/j.compedu.2012.12.022
Tobias, S. (1994). Interest, prior knowledge and learning. *Review of Educational Research*, *64*(1), 37–54.
Tsai, C., Yu, Y., Chen, Y., & Wu, S. (2009). Brain and Cognition Inhibitory response capacities of bilateral lower and upper extremities in children with developmental coordination disorder in endogenous and exogenous orienting modes. *Brain and Cognition*, *69*(2), 236–244. https://doi.org/10.1016/j.bandc.2008.07.012
Ullsperger, M., Danielmeier, C., & Jocham, G. (2014). Neurophysiology of performance monitoring and adaptive behavior. *Physiological Reviews*, *94*(1),

35–79. https://doi.org/10.1152/physrev.00041.2012

Valenti, R., Staiano, J., Sebe, N., & Gevers, T. (2009). Webcam-Based Visual Gaze Estimation. *International Conference on Image Analysis and Processing*, *1*, 662–671.

van 't Noordende, J. E., van Hoogmoed, A. H., Schot, W. D., & Kroesbergen, E. H. (2016). Number line estimation strategies in children with mathematical learning difficulties measured by eye tracking. *Psychological Research*, *80*(3), 368–378. https://doi.org/10.1007/s00426-015-0736-z

van der Linden, W. J., & Glas, C. A. . (2000). *Computerized adaptive testing: theory and practice*. Kluwer Academic Publishers.

van der Maas, H. L. J., Dolan, C. V, Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A Dynamical Model of General Intelligence : The Positive Manifold of Intelligence by Mutualism. *Psychological Review*, *113*(4), 842–861. https://doi.org/10.1037/0033-295X.113.4.842

van der Maas, H. L. J., Kan, K. J., Hofman, A., & Raijmakers, M. E. J. (2014). Dynamics of development: A complex systems approach. In *Handbook of developmental systems theory & methodology*.

van der Maas, H. L. J., & Wagenmakers, E. J. (2005). A psychometric analysis of chess expertise. *American Journal of Psychology*, *118*(1), 29–60. https://doi.org/10.2307/30039042

van der Molen, M. W. (2000). Developmental changes in inhibitory processing : evidence from psychophysiological measures. *Biological Psychology*, *54*, 207–239. https://doi.org/10.1016/s0301-0511(00)00057-0

van der Ven, S. H. G., Boom, J., Kroesbergen, E. H., & Leseman, P. P. M. (2012). Microgenetic patterns of children ' s multiplication learning : Confirming the overlapping waves model by latent growth modeling. *Journal of Experimental Child Psychology*, *113*(1), 1–19. https://doi.org/10.1016/j.jecp.2012.02.001

van der Ven, S. H. G., Kroesbergen, E. H., Boom, J., & Leseman, P. P. M. (2012). The development of executive functions and early mathematics : A dynamic relationship. *British Journal of Educational Psychology*, *82*, 100–119. https://doi.org/10.1111/j.2044-8279.2011.02035.x

van der Ven, S. H. G., Straatemeier, M., Jansen, B. R. J., Klinkenberg, S., & van der Maas, H. L. J. (2015). Learning multiplication: An integrated analysis of the multiplication ability of primary school children and the difficulty of single digit and multidigit multiplication problems. *Learning and Individual Differences*, *43*, 48–62. https://doi.org/10.1016/j.lindif.2015.08.013

van der Ven, S. H. G., van der Maas, H. L. J., Straatemeier, M., & Jansen, B. R. J. (2013). Visuospatial working memory and mathematical ability at different ages throughout primary school. *Learning and Individual Differences*, *27*, 182–192. https://doi.org/10.1016/j.lindif.2013.09.003

van Gog, T., & Scheiter, K. (2010). Eye tracking as a tool to study and enhance multimedia learning. *Learning and Instruction*, *20*(2), 95–99. https://doi.org/10.1016/j.learninstruc.2009.02.009

van Renswoude, D. R., Raijmakers, M. E. J., Koornneef, A., Johnson, S. P., Hunnius, S., & Visser, I. (2017). Gazepath: An eye-tracking analysis tool that accounts for individual differences and data quality. *Behavior Research Methods*. https://doi.org/10.3758/s13428-017-0909-3

Vygotsky, L. . (1978). *Mind in society: the development of higher psychological processes*. Harvard University Press.

Wagenmakers, E., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology . Part I : Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wang, J., Huang, M. X., Ngai, G., & Leong, H. V. (2017). Are you stresser? Your eyes and the mouse can tell. *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 222–228.

Wang, Z., & Shah, P. (2014). The effect of pressure on high- and low-working-memory students: An elaboration of the choking under pressure hypothesis. *British Journal of Educational Psychology*, *84*(2), 226–238. https://doi.org/10.1111/bjep.12027

Welsh, M., Friedman, S. L., & Spieker, S. J. (2006). Executive Functions in Developing Children: Current Conceptualizations and Questions for the Future. In *Blackwell Handbook of Early Childhood Development* (pp. 167–187). Blackwell Publishing.

Wiebe, S. A., Sheffield, T., Mize, J., Clark, C. A. C., Chevalier, N., & Andrews, K. (2011). The structure of executive function in 3-year-olds. *Journal of Experimental Child Psychology*, *108*(3), 436–452. https://doi.org/10.1016/j.jecp.2010.08.008

Wijnen, J. G., & Ridderinkhof, K. R. (2007). Response inhibition in motor and oculomotor conflict tasks: Different mechanisms, different dynamics? *Brain and Cognition*, *63*(3), 260–270. https://doi.org/10.1016/j.bandc.2006.09.003

Wilkinson, H. R., Smid, C., Morris, S., Farran, E. K., Dumontheil, I., Mayer, S., Tolmie, A., Bell, D., Porayska-pomsta, K., Holmes, W., Mareschal, D., Thomas, M. S. C., & Team, T. U. (2020). Domain-Specific Inhibitory Control Training to Improve Children ' s Learning of Counterintuitive Concepts in Mathematics and Science. *Journal of Cognitive Enhancement (2020)*, *4*, 296–314. https://doi.org/10.1007/s41465-019-00161-4

Williams, B. R., Ponesse, J. S., Schachar, R. J., Logan, G. D., & Tannock, R. (1999). Development of Inhibitory Control Across the Life Span. *Developmental Psychology*, *35*(1), 205–213.

Wilson, K., & Nichols, Z. (2015). *The Knewton platform: a general-purpose adaptive learning infrastructure*. https://www.profijt.nu/wp-content/uploads/2015/09/20150902-White-paper-The-Knewton-Platform.pdf

Wouters, P., Nimwegen, C. Van, & Spek, E. D. Van Der. (2013). *A Meta-Analysis of the Cognitive and Motivational Effects of Serious Games*. *105*(2), 249–265. https://doi.org/10.1037/a0031311

Wu, S. S., Meyer, M. L., Maeda, U., Salimpoor, V., Tomiyama, S., Geary, D. C., & Menon, V. (2008). Standardized assessment of strategy use and working memory in early mental arithmetic performance. *Developmental Neuropsychology*, *33*(3), 365–393. https://doi.org/10.1080/87565640801982445

Yeniad, N., Malda, M., Mesman, J., Ijzendoorn, M. H. Van, & Pieper, S. (2013). Shifting ability predicts math and reading performance in children : A meta-analytical study. *Learning and Individual Differences*, *23*(October 2017), 1–9. https://doi.org/10.1016/j.lindif.2012.10.004

Yilmaz, B. (2017). *Effects of adaptive learning technologies on math achievement: a quantitative study of aleks math software*. University of Missouri-Kansas city.

Zakay, D. (1993). The Impact of Time Perception Processes on Decision Making under Time Stress. In *Svenson O., Maule A.J. (eds) Time Pressure and Stress in Human Judgment and Decision Making* (pp. 59–72). Springer.

APPENDIX

Table A.1. Description of Games

	Name of the game	Number of participants	Number of accuracy response sequences	Average age (y) (SD)	Description	Application	Response mode asked for question	Max time to respond (s)
1	Addition	44949	104778	9.1 (1.9)	Sums from 1 + 1 = ... to 26900 + 4400 = ...	Mathematics	Mixed	20
2	Chaos of letters	13540	35278	9.4 (1.7)	Putting the letters of a word in the right order.	Language	Open answer	30
3	Counting	32395	67782	8.2 (1.9)	Counting the number of fishes on the screen [2-100 fishes].	Mathematics	Mixed	20
4	Dictation	11949	25077	9.8 (1.7)	Listening to a sentence, where a word is repeated afterwards. The repeated word needs to be written correctly.	Language	Open answer	30
5	Division	29109	78670	10.1 (1.5)	Divisions from 4 : 2 = ... to 4601 : 1000 = ...	Mathematics	Open answer	20
6	Flowercode	12912	32770	9.4 (1.7)	Logical reasoning game, similar to Mastermind, where the code needs to be cracked	Mathematics	Open answer	60

					with limited but sufficient information.			
7	Fractions	4250	7136	10.9 (1.4)	Variety of exercises of fractions, percentages and proportions. For example, which of the five fractions is the smallest. Or there is a 96% chance of rain today. What is the chance that it remains dry?	Mathematics	Mixed	30
8	Grammar	11892	31995	10.5 (1.4)	Naming the word class of a word in a sentence. For example, "I am giving the dog food. Is "the" a verb, article, noun or numeral?"	Language	Multiple Choice	20
9	Money	46421	124599	9.6 (1.6)	Practice with coins and banknotes to estimate the price of a product. For example, what is the combined cost of a €1.50 ice cream and €2 fries?	Mathematics	Mixed	20/30/40
10	Multiplication	30789	65168	9.9 (1.5)	Multiplications from 1 x 0 = ... to 6000 x 803 = ...	Mathematics	Open answer	20

11	Numbers	7087	21292	9.6 (1.7)	Obtain a target number by using a set of provided numbers and operations. For example, obtain 1 by using the numbers 2, 4 and 9 and the operations x and – (solution is 2 x 4 = 8, 9-8=1).	Mathematics	Open answer	60
12	Parsing words	9643	25549	10.9 (1.1)	Parsing the function of a word in the sentence. For example, “I will become a pilot”. What is “I”? Direct/indirect object, the subject or the finite verb?	Language	Mixed	20
13	Proverbs	5007	12186	10.6 (1.6)	The right meaning of a proverb needs to be chosen out of 5 options.	Language	Multiple Choice	30
14	Reading	20547	70307	9.6 (1.7)	Reading a text and clicking on the nonsense words.	Language	Open answer	30
15	Series	9848	17490	9.0 (2)	Exercises where multiple operations are combined, e.g. 3 x 5 + 2 = … or very difficult, (8-2) x 10 : 5 x 4.	Mathematics	Mixed	20

16	Slow mix	29623	49447	9.2 (1.8)	Mix of arithmetical operations sums, subtractions, multiplications and divisions) at a slow pace.	Mathematics	Open answer	40
17	Speed mix	18176	50898	10.1 (1.7)	Mix of arithmetical operations (sums, subtractions, multiplications and divisions) at a fast pace.	Mathematics	Multiple Choice	8
18	Spelling	27194	85865	10.0 (1.6)	Six different spellings of a word are presented. Five of them are spelled incorrectly and 1 is spelled correctly. The right spelling needs to be clicked.	Language	Multiple Choice	20
19	Subtraction	38590	82293	9.1 (1.9)	Subtractions from 8 – 8 = ... to 85200 – 8870 = ...	Mathematics	Mixed	20
20	Telling time	23471	52975	9.4 (1.7)	Telling time with analogue and digital clocks.	Mathematics	Open answer	30
21	Verbs	14888	41235	10.7 (1.3)	Conjugating verbs in different tenses (present/past tense etc). For example, "He [want]...an ice cream (present tense)."	Language	Open answer	30

22	Vocabulary	25718	74749	10.1 (1.5)	The right meaning of a word needs to be chosen out of 5 options. For example, courage = caring, bravery, cowardice, timid or honest.	Language	Multiple Choice	20
23	Word forms	18818	44609	10.3 (1.5)	Practicing setting words in the right singular/plural form, such as: "One belt. Five ...."	Language	Open answer	30

A.2 Static and dynamic mouse method

The classification of the attraction of mouse points and movement between two consecutive mouse points towards the answer options (A1:A6) was calculated by two methods. First, the static method is visualised as (1) in the **Fig. A.2** below. The mouse locations were classified as one of A1:A6 when they were inside one of the green area circles (answer option circles of 40 pixels). Mouse locations within the red area in the middle (30 pixels), corresponding to the starting button were excluded from the analyses.

The dynamic method considered whether there was movement, i.e., attraction, towards an answer box. This method was used for the mouse locations that were not inside the box. This attraction was calculated from one time point to the next (see **Fig. A.2 2-6** for the processing steps). The vector between consecutive mouse locations (n and n+1) was first computed (r_m; **step 2**); then the vectors from the mouse location n to all the answer options (A1:A6) were calculated ($r_{n\_A1:A6}$; **step 3**). Thirdly, the angles ($\theta_{m\_A1:A6}$) between the mouse locations vector (r_m) and the answer option vectors ($r_{n\_A1:A6}$) were computed (**step 4)**. The smallest of these angles between the mouse locations vector and the answer options was classified as the answer option where the movement is drawn towards (i.e., towards answer option A1, in the case of the first vector), only when there was a small enough angle towards one of the answer options (i.e. smaller than 25 degrees, chosen based on pilot testing with similar trajectories; **step 5**). This pre-processing procedure was done for all consecutive mouse locations (**step 6**) for every trajectory individually and eventually combined per trajectory.

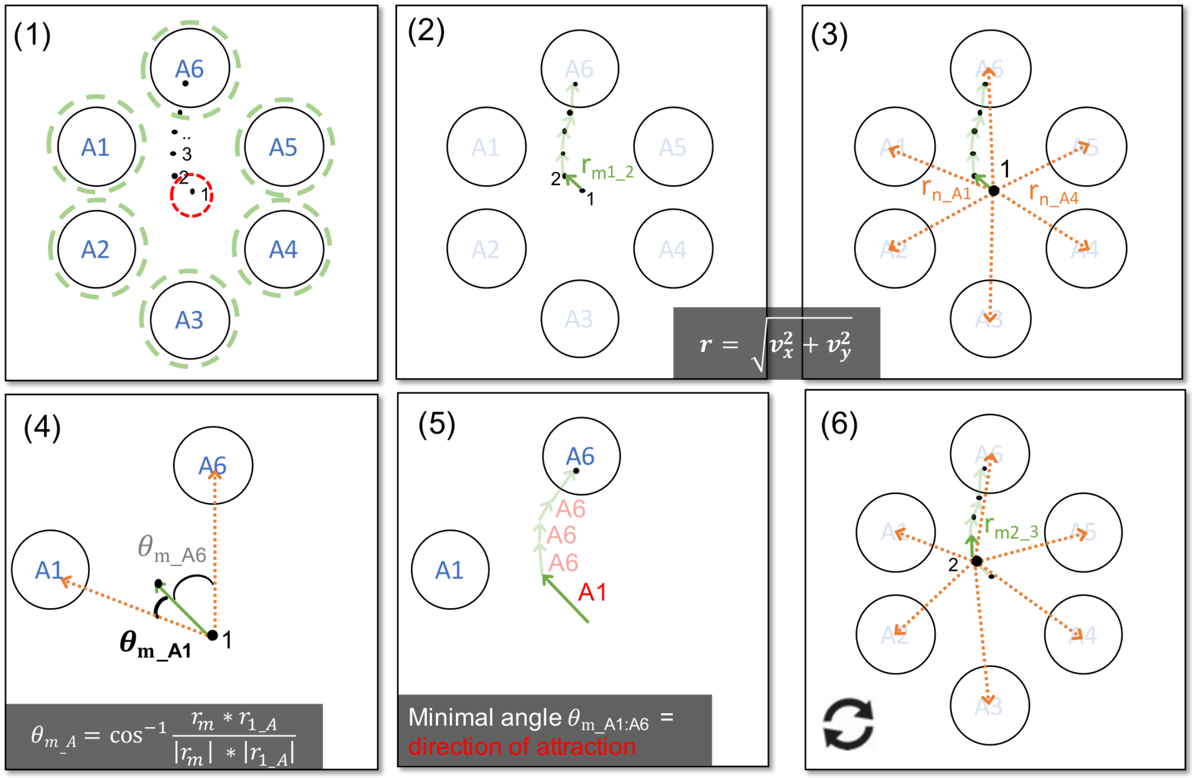


Figure A.2. *Classification of the attraction of mouse points and movement between two consecutive mouse points towards the answer options (A1:A6). (1) Static method: The mouse locations were classified as one of A1:A6 when they were inside one of the green area circles (answer option circles of 40 px). Mouse locations within the red area in the middle (30 pixels), corresponding to the starting button were excluded from the analyses (2) Dynamic method: Vectors between two mouse locations (green arrow; rm) were computed, as well as (3) vectors towards each answer option (green arrows r_{n_A1}-r_{n_A6}). (4) The angles (θ) between rm and $r_{n\_A1\text{-}A6}$ were computed. (5) The smallest angle, if <25°, leads to a categorisation of the mouse's direction of attraction towards the corresponding answer option (here A1). (6) This process was repeated for every two consecutive mouse points in the trajectory.*

A.3 Individual mapping of mouse movement to errors

Fig. A.3 shows the weighted checks for individual correlations between mouse movements and errors made in the past. The first check was weighing the correlations with a combination of the number of mouse trajectories and response errors made in the past. **Fig A.3 A** shows that the correlations were variable when not many mouse trajectories or errors were found for that individual but become more stable and remain above zero on average when more data points were found. The second check was weighing the correlations with how long ago the response error was made. **Fig A.3 B** shows that the correlations with mouse movement did not depend on how long ago the response error was made.

Figure A.3*.** ***(A) *average correlation between individual response errors and mouse movements towards those errors plotted against the weighted sum of the number of mouse trajectories collected and errors made.* ***(B)*** *Average correlation plotted against how many days had passed since the last response error was registered before the individual made the first mouse trajectory with a correct response. The 95% confidence intervals are displayed as the shaded area around the mean for both lines.*