



PhD-FLSHASE-2020-05

The Faculty of Language, Literature, Humanities, Arts and Education

## DISSERTATION

Defense held on 31/01/2020

To obtain the degree of

# DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG EN PSYCHOLOGIE

by

Max GREISEN

Born on February 3, 1986 in Luxembourg, Luxembourg

## TAKING LANGUAGE OUT OF THE EQUATION: THE ASSESSMENT OF BASIC MATH COMPETENCE WITHOUT LANGUAGE

### Dissertation defense committee

Dr. Christine Schiltz, dissertation supervisor  
*Professor, Université du Luxembourg*

Dr. Antoine Fischbach, chairman  
*Professor, Université du Luxembourg*

Dr. Caroline Hornung, vice-chairman  
*Université du Luxembourg*

Dr. Romain Martin  
*Premier Conseiller de Gouvernement, Ministère de l'Enseignement supérieur et de la Recherche*

Dr. Virginie Crollen  
*Professor, Université Catholique de Louvain*

*In memory of my late grandmother, Ada, who promised to wear a hat on the day I became a doctor.*

*Today would be the day.*

*Rest in peace.*

*“Those who have a natural talent for calculation are generally quick at every other kind of knowledge; and even the dull, if they have had an arithmetical training, although they may derive no other advantage from it, always become much quicker than they would otherwise have been.”*

(PLATO, 375BC)

# Abstract

Although numeracy, next to literacy, is an essential skill in many knowledge based societies of the 21<sup>st</sup> century, between 5 and 10 % of the population suffer from more or less severe mathematics learning disorders or dyscalculia. However, mathematical ability is not a pure construct. Instead, mathematical ability maintains a complex relationship with linguistic abilities. This relationship has significant implications for the assessment of a person's mathematical ability in multilingual contexts. The hereby presented research project addresses the consequences of that relationship in the context of Luxembourg, a highly multilingual country at the center of Europe. The aim of the project was to tackle psychometric issues that arise when the test taker does not master the language of the test sufficiently and to offer an alternative solution to available assessment batteries based on verbal instructions and tasks. In the first study, we demonstrate the role of reading comprehension in the language of instruction on third grader's performance in mathematics in Luxembourg and show that non-native speaker's underachievement in mathematics can be largely or entirely explained by their lacking reading comprehension in the language of assessment. In the next study we report on the two first pilot studies with NUMTEST, an assessment battery that aims to measure children's basic mathematical competence by replacing verbal instructions and task content with video instructions and animated tasks. The findings of these studies show that children's basic mathematical competence can indeed be reliably assessed using this new paradigm. Opportunities and limitations of the paradigm are discussed. The third and final study of this project addresses the psychometric characteristics of this newly developed assessment battery. Its findings show that the NUMTEST battery provides for good reliability and concurrent validity all while being language neutral. In summary, the presented project provides for an encouraging proof of concept for the video instruction method while offering preliminary evidence for its validity as an early screener for math learning difficulties.

# Table of contents

|  |     |
|--|-----|
| Table of contents  | 1   |
| Introduction   | 2   |
| Chapter 2: Mathematics and language: A complicated relationship  | 6   |
| Number-word structure  | 7   |
| Word problems  | 9   |
| Mathematical vocabulary  | 10  |
| Bilingual studies  | 11  |
| Reading comprehension, mathematics and the language of instruction   | 13  |
| Chapter 3: The (Mis)diagnosis of dyscalculia   | 16  |
| Chapter 4: Domain-specific predictors of mathematics achievement and NUMTEST task design   | 23  |
| Research questions   | 26  |
| Study report 1: The role of home language and comprehension of the instruction language in non-native speaker's relative underachievement in mathematics in a multilingual education system. | 28  |
| Introduction   | 29  |
| Methods  | 33  |
| Results  | 35  |
| Discussion   | 38  |
| Study report 2: Taking language out of the equation: The assessment of basic math competence without language  | 44  |
| Introduction   | 45  |
| Methods  | 50  |
| Results  | 59  |
| Discussion   | 66  |
| Study report 3: Assessing basic math competence without language: First steps towards psychometric validation.   | 79  |
| Introduction   | 80  |
| Methods  | 86  |
| Results  | 91  |
| Discussion   | 96  |
| Discussion   | 110 |
| Appendix: NUMTEST Item examples  | 133 |
| Acknowledgments  | 136 |

# Introduction

Numbers and mathematics are all around us and determine most aspects of modern life. At what time do I have to leave? How many days left until my birthday? How many servings should I buy for dinner? Can I afford to go to the restaurant tomorrow? Most aspects of life are quantifiable, the more so since the rise of digital computers and their panoptic role in many societies of the 21<sup>st</sup> century. With numbers playing such an important role throughout the strata of daily life, it comes as no surprise that, next to literacy, being able to handle numbers is one of the most if not the most important skill for predicting educational achievement (Chiswick, Lee, & Miller, 2003; Duncan et al., 2007; Watts, Duncan, Siegler, & Davis-Kean, 2014). Further than educational attainment, good numeracy skills during childhood have a significant positive impact on life outcomes such as wealth, good health and general quality of life (Gilmore, Göbel, & Inglis, 2018). However, many people struggle with learning and using numbers and mathematical reasoning as effectively as others. According to the 2015 PISA results, an international average of 23% of students do not attain level 2 proficiency in mathematics, which is considered the minimum skill requirement for being able to fully participate in a knowledge-based society (OECD, 2016). Hidden behind this average are very extreme cases on both ends, with 90% of students attaining this level in Hong Kong, but only fewer than 10% do in the Dominican Republic. In Luxembourg, where the hereby presented research project is located, students perform just below OECD average and thus 27% of 15-year-old students do not reach this minimum proficiency level. These results are corroborated by Luxembourg's very own school monitoring data from 2011 through 2013, which shows that already in third grade, between 26 and 30% of students do not reach the performance levels that are considered to be sufficient for progressing further into the curriculum (R. Martin, Ugen, &

Fischbach, 2013). Even further down the performance spectrum, between 5 and 10% of people suffer from dyscalculia (Kaufmann & Aster, 2012), a mathematical learning disorder that will be discussed in the third chapter of this thesis. Due to the hierarchical structure of mathematical knowledge (von Aster & Shalev, 2007), learning difficulties in mathematics and dyscalculia need to be discovered and addressed very early in the school curriculum as shaky foundations can only result in patchy numerical representations that are difficult to correct later in the school curriculum. Indeed, most people's difficulties in mathematics start at an early age and their skills as young children cast a long shadow over their abilities as they grow older. But what brain magic underlies performance in mathematics? What is it that makes it so hard for some people to grasp and use numbers? The mental representations and mechanisms of mathematical ability have gained increased attention in recent years and have been thoroughly studied in the field of numerical and mathematical cognition. While there is no definitive answer, the available literature is plenty and traditionally distinguishes between domain-specific and domain-general factors when trying to explain the cognitive foundations of mathematical thinking.

Concerning the domain-general brain mechanisms involved, available evidence shows that performing well on numerical reasoning is correlated to the performance of a multitude of brain modules. First, there is considerable evidence for a positive link between good executive functioning and mathematics achievement (Bull & Lee, 2014; Clements, Sarama, & Germeroth, 2016; Verdine, Irwin, Golinkoff, & Hirsh-Pasek, 2014). Of all executive components, working memory has received the most attention. Both visuo-spatial and verbal working memory performance (Baddeley & Hitch, 1974) have been found to be related to mathematics achievement (Alloway & Passolunghi, 2011; Friso-van den Bos, van der Ven, Kroesbergen, & van Luit, 2013; Hornung, Schiltz, Brunner, & Martin, 2014; Mou, Berteletti, & Hyde, 2018; Raghobar, Barnes, &

Hecht, 2010) and seem to be particularly relevant for subtraction (Caviola, Mammarella, Lucangeli, & Cornoldi, 2014). The exact role of working memory subcomponents in various mathematical tasks remains unclear. Some studies have found that visuo-spatial working memory performance was especially predictive of mathematics performance during the early stages of mathematics acquisition while verbal working memory was more predictive of mathematics performance in later grades (De Smedt et al., 2009; Van de Weijer-Bergsma, Kroesbergen, & Van Luit, 2015). However, other studies have suggested the inverse order of correlation (e.g. Alloway & Passolunghi, 2011). Beyond visuo-spatial working memory, the role of general visuo-spatial skills in mathematics performance has been investigated. Several longitudinal studies have been able to highlight the importance of early visuo-spatial ability in later mathematics achievement (Casey et al., 2015; Mix et al., 2016; Verdine et al., 2014), seemingly more so in girls than in boys (Laski et al., 2013). Reasoning abilities have also been shown to be related to early numerical skills (Hornung et al., 2014) and to predict later arithmetical reasoning (Stock, Desoete, & Roeyers, 2009). Especially fluid reasoning (Cormier, Bulut, McGrew, & Singh, 2017; Green, Bunge, Briones Chiongbian, Barrow, & Ferrer, 2017; Hornung et al., 2014) and processing speed (Cormier et al., 2017) have been identified as strong and reliable predictors of mathematics achievement throughout the school curriculum. All in all, the current state of research suggests that mathematical thinking and number competence seem to draw on a multitude of cognitive functions with considerable variety depending on the task. Finally, there is a growing research focus on the role of different language skills in mathematics acquisition and assessment. Gilmore and colleagues state that language skills play both a general and various specific roles in mathematics (Gilmore et al., 2018). The complex and, as I will show, often damaging relationship between language and mathematics will be discussed in chapter two. Chapter three will be dedicated to



dyscalculia, the problems that language brings about in the curricular and psychometric assessment of mathematical ability as well as the solution that I explored in the context of the hereby presented research project. The final introductory chapter will then address the domain-specific competencies that have been shown to have the strongest predictive power for later mathematics achievement and how the solution I proposed was designed around these fundamental numerical skills.

# Chapter 2

## Mathematics and language: A complicated relationship

While formal mathematics use their own symbolic language, it has been known for a long time that, nevertheless, learning and using mathematics taps into language skills, both at the general and at specific levels. References to the more general aspects can be found in educational research papers dating back to the 1970es. For example in 1979, Austin and Howson state that '*In the teaching and learning of mathematics, language plays a vitally important role*' (Austin & Howson, 1979). Indeed, it is difficult to imagine a classroom that does not use common language to instruct and communicate about mathematics or that mathematical concepts could be acquired without sufficient language competence. Beyond this general role, the more specific relationships between different aspects of language and mathematics have been studied. In 1996, Ellerton & Clarkson stated that '*Although language factors have long been recognized as having an important influence on mathematics learning, possible frameworks for researching the nature and extent of that influence have only been developed relatively recently.*' Over twenty years have passed since this statement, and there has been a vast body of research on the different ways in which language influences the acquisition and the use of mathematics. This chapter will focus on these aspects by providing an overview of the different angles under which these questions have been investigated.

## Number-word structure

A first aspect that has been studied for potential relationships between linguistics and mathematics is the structure of number words. In that context, one distinguishes between languages that use a transparent number system and those that use an intransparent, inverted number-naming system. In a transparent number word system, the structure of the number word follows the decimal structure of the number it represents. English for example follows a transparent number-word structure: in the word forty-two, the decade precedes the unit just as in the Arabic symbolic form 42. German on the other hand uses an intransparent number word structure: 42 is written as *zweiundvierzig*, which, if literally translated into English, would be said „two and forty“, thus inverting the decade-unit structure of the Arabic symbolic form. Other facets of language transparency in the context of mathematics include the use of irregular number words for some numbers. In French for example, the numbers 11, 12, 13...16 have proper, intransparent names: *onze, douze, treize...seize*. Starting from the number 17, a transparent system is used: *dix-sept, dix-huit, dix-neuf*, which translates into *ten-seven, ten-eight, ten-nine*. This system is then used up to and including 69, which is transparently named *soixante-neuf*. Arriving at 70 though, the French language switches to an intransparent naming system again: 70 is named *soixante-dix*, literally meaning sixty-ten and has no word for 70 per se. Following numbers are then named according to the system used for 11, 12, 13 etc.: 71 becomes *soixante-et-onze*, literally meaning sixty-and-eleven. This system is used up to and including 79. Starting from 80, another, new naming system is then used which refers to the vigesimal numeral system: 80 is called *quatre-vingt*, literally *four-twenty* and the following numbers are then named accordingly. 81 becomes *quatre-vingt-un*, translated as *four-twenty-one* and numbers up to and including 89 use this structure. But, there's more! As the French language doesn't have a specific word for ninety either, the structure used for

the seventies is now used again for numbers from 90 to 99, but with a twist! Instead of being based on the decimal system used for 70 to 79, the naming structure for the nineties now follows the vigesimal system used for the preceding eighties: 90 becomes *quatre-vingt-dix*, 91 becomes *quatre-vingt-onze* etc<sup>1</sup>. Similar but less extensive irregularities are found in the English language for the numbers eleven and twelve which, if the system used transparent number naming would be called *oneteen* and *twoteen*, just like their successors thirteen, fourteen etc. While there are historical and etymological reasons behind convoluted naming systems like the French one, a reader unfamiliar with the French language will have little trouble understanding the problems this naming system can produce in the context of learning numbers and mathematics. In fact, the cognitive effects of intransparent number naming systems on performance in mathematics have been investigated by numerous studies. Some have suggested that the use of an intransparent number-word system requires additional working-memory resources when compared to a transparent system (Zuber, Pixner, Moeller, & Nuerk, 2009). The same authors also report a study conducted in a Czech sample. The Czech language is the perfect candidate for studying the effects of transparency of the numberword structure as it features both a transparent and a nontransparent system. They found that the intransparent numberword structure lead to a majority of inversion related errors which were practically absent in the transparent variant (Pixner, Zuber, et al., 2011). Similar effects have been found by Imbo and colleagues when comparing French (transparent) with Dutch (intransparent) (Imbo, Vanden Bulcke, De Brauwer, & Fias, 2014). Krinzinger and colleagues found that intransparent number-word structure had negative effects on writing Arabic numbers from dictation, but found no such effect on number recognition (Krinzinger et al., 2011).

---

<sup>1</sup> Some variations of the French language use a somewhat more transparent naming system when it comes to decades. Belgian French uses *septante* instead of *soixante-dix* and *nonante* instead of *quatre-vingt-dix*. Curiously, eighty is still named intransparently: *quatre-vingt*. No trace of the vigesimal number system remains in Swiss French however, which uses *octante* for eighty.

However, others have found that an intransparent number word structure had detrimental effects on place-value processing, even with nonverbal Arabic symbolic digits (Moeller, Shaki, Göbel, & Nuerk, 2015; Pixner, Moeller, Hermanova, Nuerk, & Kaufmann, 2011). Taken together, currently available research seems to suggest that transparent number word systems provide an easier setting for children who are learning transcoding skills. The observed detrimental effects of an intransparent number-word system can be substantial and as such, the use of a completely transparent number-word structure as found in mandarin Chinese and many other Asian languages has been proposed as the source of superior performance in mathematics by Asian children: Using a transparent number-word system could indeed give them a running start in manipulating basic numerical concepts (Siegler & Mu, 2008).

### **Word problems**

The most intuitively obvious offenders when it comes to links between language and mathematics are traditional word problems. Failure in solving a word problem can be due to two main reasons. On one hand we have a potential failure to understand the verbal components of the problem, leading inevitably to difficulties in deriving the latent mathematical problem. On the other hand, we have a failure to successfully solve the mathematical problem itself. Research suggests that the former is more often the case than not. In a study conducted in the Philippines on Filipino-English dual language learners, researchers have found that students performed better in word problems when they were formulated in their first languages and that when student's performance improved, the improvement was mostly due to an improvement in their text comprehension (Bernardo, 1999). In a similar vein, Vilenius-Tuohimaa and colleagues investigated the relationship between word problem performance and reading comprehension in a sample of 225 fourth grade children (Vilenius-Tuohimaa, Aunola, & Nurmi, 2008). They found that performance on word problems

was strongly related to reading comprehension: the more fluent a child's reading skills, the better it performed on mathematical word problems. Kempert and colleagues came to similar conclusions after studying the effects of reading comprehension in a sample of third graders that included both German monolinguals as well as Turkish-German bilinguals. (Kempert, Saalbach, & Hardy, 2011). They found that the more proficient a student was in the language of instruction and assessment, the better he was at solving word problems. Another study by Sepeng and Madzorera, aimed at identifying the sources of difficulty in solving word problems in a sample of grade 11 students in South Africa, found that the prevalent source of difficulty was related to deriving algebraic terms from the textual source and comprehension of the instructional vocabulary (Sepeng & Madzorera, 2014). Wang and colleagues tell the same story after investigating a group of 701 second graders in the United states. When investigating the best predictors for solving word problems, they found that after initial arithmetic and word-problem solving skills, language competence and verbal working memory were significant and meaningful predictors of word problem performance (Wang, Fuchs, & Fuchs, 2016).

### **Mathematical vocabulary**

More recently, other researchers have investigated the dependency on vocabulary for solving word problems more closely with the aim of disentangling the role of general reading comprehension from that of specific mathematical vocabulary knowledge in mathematics performance. Mathematical vocabulary refers to quantitative words and concepts such as *more, many, less, few, fewest* etc. (Purpura, Napoli, & King, 2019). Indeed, Mou and colleagues for example have shown that already in preschool, general word and letter knowledge is a meaningful predictor of performance in mathematics (Mou et al., 2018). Other studies further showed that above general vocabulary, preschoolers performance in basic numerical tasks is specifically related to knowledge

of mathematical vocabulary, above and beyond general vocabulary knowledge (Hornburg, Schmitt, & Purpura, 2018). However, the relationship was only significant for some tasks, while there was no such relationship for purely numerical tasks such as subitizing or formal addition (see also (Peng & Lin, 2019)). Corroborating these findings, Purpura and colleagues have found not only similar results in a sample of 4 year old preschoolers, but they also found that children whose parents had less than a college level education knew significantly less mathematical words than their peers (Purpura & Reid, 2016). These results indicate that already at a very early stage, before or during the preschool years, many children have already acquired a significant amount of mathematical vocabulary which in return allows them faster access to numerical concepts during preschool and early primary education. Based on these observations, it has been suggested that training mathematical vocabulary specifically during these early years is essential for the successful development of mathematical skills later on (Riccomini, Smith, Hughes, & Fries, 2015).

### **Bilingual studies**

Another angle under which the relationship between mathematics and language have been investigated are studies on bilinguals (see (Poncin, Van Rinsveld, & Schiltz, 2018) for a review). Bilinguals are an interesting population to study the effects of language on knowledge retrieval as their behavior and performance can be compared between languages in the same person. Language-dependent memory effects have been shown for general fact retrieval (see e.g. (Marian & Fausey, 2006)) but also for mathematical knowledge. In that sense, a line of studies has been able to show that mathematical representations are not stored in a language independent format. Saalbach and colleagues report a study in which bilingual high-school students were trained on subtraction and multiplication problems over four days in one language (German or French) and then assessed in the other one (Saalbach, Eckstein, Andri, Hobi, & Grabner, 2013). They found

significant cognitive costs due to language switching, indicating that the newly learned material was stored in the language of instruction and that translation to the language of assessment comes with additional effort (see also (Kempert et al., 2011) for similar findings in a population of Turkish-German students in Germany). Another study by Spelke and colleagues found that Russian/English bilingual college students that were trained on items containing both exact and approximate numerical information in both languages retrieved information more accurately when the language of acquisition matched the language of retrieval, but only for exact numbers (Spelke & Tsivkin, 2001). No such effect was found for approximate numerical information, suggesting that exact, large number representations are stored in a language-specific format. In the same line of studies, Van Rinsveld and colleagues have been able to show that in Luxembourgish dual-language learners, providing linguistic context improved their performance in solving arithmetical problems, but only when the problems had to be solved in their second language, which is also the language of instruction (Van Rinsveld, Schiltz, Brunner, Landerl, & Ugen, 2016). These findings suggest that the bilingual brain defaults to the language in which knowledge has been acquired during retrieval and that retrieval in another language constitutes a bigger effort that can be facilitated by providing linguistic context in this second language. Similarly to the previous study, participants were faster when solving arithmetical problems in their first language of instruction (German) than in their second (French) (Van Rinsveld, Dricot, Guillaume, Rossion, & Schiltz, 2017). Behavioral findings such as these have been further corroborated by neuro-imagery studies. Van Rinsveld and colleagues for example have found that in highly proficient German/French bilinguals, fMRI activation patterns resulting from solving complex arithmetical problems differed between their two languages (Van Rinsveld et al., 2017). Lin and colleagues however found that activation patterns between languages in a population of Chinese/English bilinguals were largely



identical, but found higher activation levels when problems were solved in the participants second language, again suggesting that when language of instruction and language of retrieval differ, switching comes at a cost.

### **Reading comprehension, mathematics and the language of instruction**

The literature presented so far points to the conclusion that sufficient language competence is necessary for succeeding in many but not all mathematical tasks. While numerical representations are largely independent of language (but see (Salillas & Carreiras, 2014) for findings that suggest otherwise), it is in the construction and retrieval of these representations that language plays a significant role (Gelman & Butterworth, 2005). It is thus no surprise that reading comprehension and mathematics achievement have been shown to share considerable covariance at different levels of the school curriculum. Gjicali and colleagues have shown that linguistic competence measured at a very young age (1,5-3,5 years) in language minority and low income pupils predicts arithmetical competence at preschool age (4,5-6,5 years) (Gjicali, Astuto, & Lipnevich, 2019). Similar observations were made by Zhang based on a study conducted on preschoolers in Hong Kong (Zhang, 2016). The author concludes that written language is an essential building block for children's acquisition of number concepts at an early age. Beyond the early stages of acquisition, other studies have shown that reading comprehension and mathematics achievement are consistently related throughout the curriculum. For example, Korpipää and colleagues have found that reading and arithmetic skills share significant covariation both in grade 1 and in grade 7 in a sample of 1335 Finnish students (Korpipää et al., 2017). The correlation between reading comprehension and arithmetic performance is on average .55 and very consistent, as is shown by a meta-analysis conducted by Singer & Strassen on 68 individual study samples (Singer & Strasser, 2017). Moreover, Vukovic and Lesaux have found that language ability also predicts gains in

different mathematical fields, both for language majority and language minority speakers (Vukovic & Lesaux, 2013). It is in studies conducted on language minority students that the importance of good language comprehension not only becomes very clear, but also problematic. If the relationship between reading comprehension and success in mathematics is that consistent, then students that are less competent in the language of instruction and assessment of their school system will inevitably perform worse in mathematics than those who are competent users of the instruction language. Differences in language comprehension account for most of the performance differences between children with and without an immigration background (Kempert et al., 2016) in many areas including mathematics. In the same line of research, Paetsch and colleagues showed that performance differences in mathematics between children with a German background (language of instruction) and those with a foreign language background disappeared entirely after controlling for reading comprehension (Paetsch, Radmann, Felbrich, Lehmann, & Stanat, 2016). In the context of the present thesis, I conducted a similar study (study report 1) on the Luxembourgish school population and found the same pattern of results: Language minority pupils performed worse than their native peers both in measures of reading comprehension in the language of instruction and in mathematics and we showed that the differences in mathematics performance are largely or even entirely mediated by differences in reading comprehension.

The findings presented so far bring about a problematic situation when it comes to educational and psychometric assessment of mathematics in multilingual settings. If language competence is so closely related to mathematics performance, and if the tests evaluating mathematical skill draw heavily on verbal instructions and task content, then one can question how much of the performance in these tasks is effectively attributable to numerical reasoning and how much of the performance is due to varying levels of language competence. Beyond the empirical evidence

presented so far, the problem is very easy to illustrate. Imagine the following problem, written in Luxembourgish, a language that the international reader is very unlikely to understand: *Zwee Schwéngercher gin an de Stall. Et komme nach zwee Schwéngercher dobäi. Wéivill Schwéngercher sin elo am Stall?* This is a grade one level word problem. The underlying arithmetic problem is very simple: How much is two plus two. However, you could not solve it. It must mean that you are terrible at arithmetic, right? Of course, that is not the conclusion that you would draw as a reader. Sadly, it is the conclusion that many children with a foreign language background are faced with in multilingual settings in which their first language doesn't match the language of instruction and assessment. This realization isn't exactly new as a few authors have pointed to it before (e.g. Abedi, 2002; Abedi & Lord, 2001; Hickendorff, 2013). While the issue has been considered in educational settings, similar considerations arise when it comes to psychometric assessment and the diagnostics / screening of mathematical learning disabilities and dyscalculia.

# Chapter 3

## The (Mis)diagnosis of Dyscalculia

Dyscalculia is a learning disorder whose definitions are yet relatively unclear. The ICD-10 (World Health Organization, 1992) didn't use the word dyscalculia yet and spoke instead of a '*specific disorder of arithmetical skills*' that '*involves a specific impairment in arithmetical skills, which is not solely explicable on the basis of general mental retardation or of grossly inadequate schooling. The deficit concerns mastery of basic computational skills of addition, subtraction, multiplication, and division (rather than of the more abstract mathematical skills involved in algebra, trigonometry, geometry, or calculus.*' Two characteristics stand out in this definition. First, it's a specific disorder of only arithmetical skills. It must be independent from general mental retardation. Second, it's a disorder that affects only basic arithmetical skills and not the more complex mathematical concepts. This definition is relatively old and at the time of writing this text, but the 11<sup>th</sup> revision of the ICD is in the works. In an online preview version of the new classification (<https://icd.who.int/browse11/l-m/en>), the term dyscalculia still isn't used. Instead, it is now defined as a '*developmental learning disorder with impairment in mathematics*' and described as a disorder that '*is characterized by significant and persistent difficulties in learning academic skills related to mathematics or arithmetic, such as number sense, memorization of number facts, accurate calculation, fluent calculation, and accurate mathematical reasoning. The individual's performance in mathematics or arithmetic is markedly below what would be expected for chronological or developmental age and level of intellectual functioning and results in significant impairment in the individual's academic or occupational functioning. Developmental learning disorder with impairment in mathematics is not due to a disorder of intellectual*

*development, sensory impairment (vision or hearing), a neurological disorder, lack of availability of education, lack of proficiency in the language of academic instruction, or psychosocial adversity.* Several differences with the previous definition exist. First, the listing of affected competencies has been broadened and specified. Second, there is now a normative criterion concerning the relative performance of the affected person when compared to normally developing peers. Lastly and most importantly for the present thesis, the exclusion criteria have been widened and now include 'lack of proficiency in the language of instruction'. I will come back to this, but first I want to present the definitions of dyscalculia as offered by the two commonly used versions of the Diagnostic and Statistical Manual of Mental disorders, the de-facto classification system used in Psychology and Psychiatry around the world. In the fourth and still commonly used edition of the DSM (American Psychiatric Association, 1998), the term dyscalculia is also not used. Instead, it defines the *Mathematics disorder*, whose diagnostic criteria are established as follows:

- A. Mathematical ability, as measured by individually administered standardized tests, is substantially below that expected given the person's chronological age, measured intelligence, and age-appropriate education.
- B. The disturbance in Criterion A significantly interferes with academic achievement or activities of daily living that require mathematical ability
- C. If a sensory deficit is present, the difficulties in mathematical ability are in excess of those associated with it.

The two main characteristics of this definition are the normative criterion assessed by a standardized test on one hand and, indirectly, the exclusion criterion of general mental retardation. In that sense, the definition is similar to the one in ICD-10 while remaining rather imprecise when it comes to symptomatic manifestations of the disorder. Similarly to the ICD, the DSM has been

recently revised and a fifth edition was published in 2013. In this new version (American Psychiatric Association, 2013), a search for the term dyscalculia finally yields a result. The former *Mathematics disorder* is now classified as a *Specific learning disorder with impairment in mathematics*. Symptoms include problems with the person's number sense, memorization of arithmetic facts, accurate or fluent calculation and accurate mathematical reasoning. Dyscalculia is defined in a note stating that it is *an alternative term used to refer to a pattern of difficulties characterized by problems processing numerical information, learning arithmetic facts, and performing accurate or fluent calculations*. While the complete diagnostic criteria of specific learning disorders as defined in the DSM-V are too voluminous to be reported here, two criteria are specifically related to mathematics:

1. Difficulties mastering number sense, number facts, or calculation (e.g., has poor understanding of numbers, their magnitude, and relationships; counts on fingers to add single-digit numbers instead of recalling the math fact as peers do; gets lost in the midst of arithmetic computation and may switch procedures).
2. Difficulties with mathematical reasoning (e.g., has severe difficulty applying mathematical concepts, facts, or procedures to solve quantitative problems).

Another criterion of importance for all specific learning disorders including mathematics is that *'the affected academic skills are substantially and quantifiably below those expected for the individual's chronological age, and cause significant interference with academic or occupational performance, or with activities of daily living, as confirmed by individually administered standardized achievement measures and comprehensive clinical assessment.'* The most visible change from other definitions presented so far is that the exclusion of general mental retardation is not part of the diagnostic criteria in this newer definition.

Definitions of dyscalculia are thus manifold and comprise varying degrees of specificity. The common denominator, as pointed out by Gilmore and coauthors (Gilmore et al., 2018), is that a diagnosis of a mathematical disorder, or dyscalculia, is ‘*nearly always based on an individual’s performance on a standardized mathematical achievement test*’, while other explicatory factors for underachievement in mathematics need to be ruled out.

However, when looking at available psychometric test batteries for screening and diagnosing math learning difficulties, one can quickly see that all of them are based on verbal instructions and verbal task content to varying degrees. In Luxembourg, commonly used tests include but are not limited to the *Eggenberger Rechentest* (ERT) (Schaupp, Holzer, & Lenart, 2007), the *Osnabrücker Test zur Zahlbegriffsentwicklung* (OTZ) (van Luit, van de Rijt, & Hasemann, 2001), the *Neuropsychologische Testbatterie für Zahlenverarbeitung und Rechnen bei Kindern* (ZAREKI) (von Aster, Bzufka, & Horn, 2009), the *Test diagnostique des compétences de base en mathématiques* (TEDI-MATH) (Noël, Grégoire, & Nieuwenhoven, 2008) or the *Rechenfertigkeiten- und Zahlenverarbeitungs-Diagnostikum* (RZD) (Jacobs & Petermann, 2014). While these tools are of good quality when used appropriately, several issues arise when they are used in multilingual contexts. As should be clear after the literature presented so far, mathematical competence is far from independent from language proficiency, which is further underlined by the inclusion of ‘lack of proficiency in the language of instruction’ as an exclusion criteria in the definition provided by ICD-11. In other words, when the tested person doesn’t master the language of the test sufficiently, the test can be considered as neither objective, sensitive, reliable or generally valid in any form. For a more detailed critique of how this situation affects the classical quality criteria of a psychometric tests, I will refer the reader to the introduction of the last study report in this thesis. Above and beyond the language bias, a definition of a disorder that is vaguely

based on *some* standardized test battery is problematic. In practice, this leads to a situation in which there are as many different definitions of what does and what does not constitute dyscalculia as there are assessment batteries. While efforts have been made to propose a less dependent definition of math learning difficulties and dyscalculia (Geary, 1993, 2010), the problem persists and leads to large differences in the reports on the prevalence of dyscalculia, ranging from 1.3% to 10.3% (Devine, Soltész, Nobes, Goswami, & Szűcs, 2013) . This significantly hampers comparability between children, as a child could be diagnosed with dyscalculia using one battery while another conclusion could be derived using a different battery. The obvious ethical issues aside, this situation can have a dramatic impact on a child's access to intervention and support programs provided by any given educational system and might lead to circumstances in which a child might be assessed with different methods until *some* test provides the necessary criteria for inclusion. Finally, a vague definition of dyscalculia also leads to comparability issues when it comes to research on the normal and deviate development of mathematical abilities during childhood: Every study that targets a sample of dyscalculic children uses its own inclusion criteria, sometimes based on the standardized assessment of participants using a chosen test-battery during data collection, but most often based on reports of the diagnosis by participants which in turn are based on very different and incomparable test results. This severely limits the generalizability of many studies' results and constitutes a roadblock in the way of elucidating the factors underlying math learning difficulties.

Houston, we have a problem. And a solution! Faced with this situation, two possibilities for remediation come to mind. The first, theoretically ideal one, would be to have each test available in each and every language and to train highly polyglot practitioners. But one must quickly recognize that even only targeting this ideal would be hardly feasible both for practical and



financial reasons. The other solution is to remove language factors from standardized mathematics assessment altogether and thus, as the title of this thesis suggests, to remove language from the equation. This was the core idea that gave birth to the hereby presented research project and its resulting application. The objective of the project was twofold. First, by removing verbal components from the assessment process, the often adverse effects of language competence should diminish or vanish entirely, thus providing a less linguistically biased measure of basic number competence. Second, by developing an assessment method for basic number competence that is independent of the language skills of the assessed and the language context of its administration, I could provide a tool that can be used in many multilingual countries and provide a basis for interpersonal and international comparison of results. We choose to do this by developing NUMTEST, a web-application for tablet computers that replaces verbal task instructions and task content by video instructions and animated tasks. The basic principle underlying the method is that children get shown a video of a hand completing the task correctly, which is shown by a green happy smiley that appears at the end of each item. After watching the video, children then get to practice on their own on similar items before the actual assessment starts. For a more detailed description on how this works, I invite the reader to have a look at the methods sections of both the second and third study reports in this dissertation. In line with the main message of this thesis, I have also edited a video about the method and the project which can be watched under the following address (<https://www.youtube.com/watch?v=S1JhWr5DspE>).

Now that that it should be clear why I chose to develop an entirely nonverbal test battery for basic number competence, the remaining question is that of ‘How?’. What should be measured in this test? What are the most reliable predictors of math achievement at the beginning of the school

curriculum? Which of these can be assessed by using our novel methodology and the given technological framework? These questions will be addressed in the next chapter.

# Chapter 4

## Domain-specific predictors of mathematics achievement and NUMTEST task design

The tasks in the NUMTEST assessment were designed according to three guidelines. First, we wanted to include tasks for which there was empirical evidence that they would provide the best possible predictive power over later mathematics achievement. As stated before, mathematical knowledge is built hierarchically with its foundations in basic number skills learned during a child's preschool years. The currently most referenced theoretical model for children's development of numerical abilities is the four-step developmental model (von Aster & Shalev, 2007). The model describes the evolution of a child's numerical abilities, starting in its infancy before moving to preschool and into the first year of formal mathematics education. In this model, the development of numerical representations starts with genetically inherited core representations of magnitude and its related functions, namely subitizing and large quantity estimation. These functions are also often referred to as the number sense (Dehaene, 2011), a brain module designed to precisely account for quantities up to and including 4 (subitizing) and to estimate quantities above that threshold. The current state of empirical evidence suggests that this module is of ontogenetic origin and that humans share it with many other species ranging from closely related to humans like primates (Matsuzawa, 2009) to less related species like lions (Benson-Amram, Gilfillan, & McComb, 2018), fish (Agrillo, Dadda, Serena, & Bisazza, 2009), chicken (Rugani, Vallortigara, Priftis, & Regolin, 2015) and even insects like the honeybee (Howard, Avarguès-Weber, Garcia, Greentree, & Dyer, 2019). The model stipulates that, as working memory capacity evolves during early childhood, new mathematical knowledge is built upon these core

representations as they provide for the first basic meaning of number. In the next two steps, these core representations are then extended by two symbolic systems, one of verbal nature by providing number words for the core representations, another one in the form of the Arabic digits. In the final step, these symbolic representations provide the core for the development of what is called the mental number line *‘in which ordinality is represented as a second (acquired) core principal of number’* (von Aster & Shalev, 2007). According to this model, math learning difficulties and dyscalculia can result from deficient development of any of these four steps.

In the context of this theoretical framework, research in numerical cognition has tried and succeeded at identifying several of these basic numerical competencies that have high predictive power for later mathematics achievement. Indeed, there is a large body of evidence showing that the best predictor for mathematics achievement throughout the school curriculum and later life is performance on basic numerical tasks during or around the preschool years (Aunio & Niemivirta, 2010; Aunola, Leskinen, Lerkkanen, & Nurmi, 2004; Duncan et al., 2007; Hornung et al., 2014; Jordan, Glutting, & Ramineni, 2010; Krajewski & Schneider, 2009; Locuniak & Jordan, 2008; R. B. Martin, Cirino, Sharp, & Barnes, 2014; Nguyen et al., 2016; Watts et al., 2014). More specifically, these preschool competencies include counting ability (Aunola et al., 2004; Bartelet, Vaessen, Blomert, & Ansari, 2014; Hornung et al., 2014; R. B. Martin et al., 2014; Mou et al., 2018; Nguyen et al., 2016; Passolunghi, Lanfranchi, Altoè, & Sollazzo, 2015), seriation/ordering as measured for example by the number-line task (see (Schneider et al., 2018) for a meta-analysis of evidence) as well as symbol knowledge (Göbel, Watson, Lervåg, & Hulme, n.d.; Purpura, Baroody, & Lonigan, 2013) and symbolic magnitude comparison (Bartelet et al., 2014; De Smedt et al., 2009; Hawes, Nosworthy, Archibald, & Ansari, 2019; Lyons, Price, Vaessen, Blomert, & Ansari, 2014; Sasanguie, Göbel, Moll, Smets, & Reynvoet, 2013; Sasanguie, Van den Bussche, &

Reynvoet, 2012; Schneider et al., 2017; Vanbinst, Ansari, Ghesquière, & Smedt, 2016) . Additionally, there is a long-standing debate on the role of the approximate number system (ANS) (Feigenson, Dehaene, & Spelke, 2004) in the development of exact mathematical representations. While several studies found a statistically significant relationship between magnitude estimation abilities and later mathematics achievement (e.g. (Bartelet et al., 2014; Hornung et al., 2014; Mou et al., 2018)), recent meta-analyses of the available evidence (Chen & Li, 2014; Schneider et al., 2017) show that while the association exists, the correlations are much smaller than for symbolic magnitude comparison (Gilmore et al., 2018).

The second driving factor for NUMTEST's task design was that it should provide a measure of those skills that are defined in Luxembourg's governmental targets for the end of preschool (MENFP, 2011). The list overlaps with most of the basic number competencies listed so far such as counting, ordering and magnitude comparison, but also states that children should be able to perform basic non-symbolic arithmetic (addition & subtraction) based on images or tangible objects.

The last factor in designing tasks for our new approach to test instructions was the technological framework. NUMTEST was developed on the basis of OASYS, a proprietary large-scale assessment framework developed by the Luxembourg Centre for Educational Testing. OASYS is a web-based software framework that is used to build questions and provide a choice of answers for the standardized assessment of high-school students. As the project did not foresee a dedicated developer, I choose this option as it had been successfully used in similar research projects for developing educational and psychometric assessments. However, the framework also came with constraints. It originally provided tools only for building written questions accompanied by a multiple choice of answers but lacked any interactivity features for the user. The ability to

automatically display video stimuli instead of static questions, or the ability to drag & drop objects as a means of answering the questions, for example, were worked into the framework specifically for this project. Given these theoretical foundations and the technological constraints, several tasks were developed over three years of development, piloting and adaptations, resulting in what can be considered as the first (study report 2) and second (study report 3) version of the NUMTEST battery (See appendix for images of the different versions of the tasks.). All in all, I aimed for a battery that comprises a measure of magnitude comparison, magnitude ordering and simple non-symbolic and non-verbal addition and subtraction tasks while providing language-neutral instructions and tasks. As you will read in the second and third study report of this thesis, several other tasks were developed but are not part of the latest version of the battery.

### **Research Questions**

In summary, the hereby presented research project aims to answer three main questions, each addressed by a different research paper.

The first study (*The role of home language and comprehension of the instruction language in non-native speaker's relative underachievement in mathematics in a multilingual education system.*) in this thesis revolves around the role of language comprehension in third grade mathematics performance in Luxembourg. Using linear regression and mediation analyses on the Luxembourgish national school monitoring data, I explored the relationship between home language, competency in the language of instruction and mathematics achievement. I hypothesized that non-native speakers would perform below native speakers both in measures of language comprehension and mathematics (1), that language comprehension would be a significant and relevant predictor of mathematics performance (2) and that the performance differences in

mathematics between native and foreign speakers could be largely or entirely explained by differences in language comprehension (3).

After providing evidence for the impeding role of language comprehension in mathematics assessment in the multilingual context of Luxembourg, the second report (*Taking Language out of the Equation: The Assessment of Basic Math Competence Without Language*) revolved around the two pilot studies conducted at the beginning of the NUMTEST project. Can verbal instructions and task content be replaced by video instructions and animated stimuli? Do children in first grade reliably understand and solve tasks that use this new paradigm? And consequently, what are the strengths and limitations of the method? The second report tries to answer all the above questions and provides a proof of concept for the validity of the video instruction method.

Finally, based on the results of the two pilot studies, a second version of NUMTEST was designed and presented to a second sample of children at the beginning of first grade. While the pilot studies focussed on the methodological aspects of the video instruction, this study was designed to address the psychometric aspects of the tasks and explores the possibility of using NUMTEST as an early screening tool for basic math competence. Do NUMTEST's tasks provide valid, objective and reliable measures of basic numerical competence? How does the battery fare when compared to existing screening tools? What is the relationship between performance on NUMTEST and a standardized measure of symbolic arithmetic? Are the tasks of adequate difficulty or too easy or too difficult for screening at the lower end of the performance spectrum? These questions are addressed and answered in the final study report of this thesis, titled *Assessing basic math competence without language: First steps towards psychometric validation*.

**Study report 1: The role of home language and comprehension of the instruction language  
in non-native speaker's relative underachievement in mathematics in a multilingual  
education system.**

*(in preparation)*

Max Greisen, Caroline Hornung & Christine Schiltz

**Abstract**

The aim of the hereby presented study was to examine the role of a series of domain-general predictors in children's mathematics performance in third grade in Luxembourg, a highly multilingual country with a multilingual education system. As available research in other countries with comparable school demographics suggested that children's performance in early fundamental school mathematics was highly related to their verbal abilities in the school's language, we decided to investigate this relationship in data provided by Luxembourg's national school monitoring program. The results suggest not only that reading comprehension in the language of instruction is the strongest of all considered domain-general predictors for mathematics performance in third grade, but that reading comprehension largely or completely mediates performance differences in mathematics between native and non-native speakers. Implications for practice and policy making are discussed.



## **Introduction**

Mathematical reasoning is not a purely abstract computational skill. Instead, it is at least partially determined by verbal abilities. The idea that performance in mathematics depends on language competence has been supported by a vast body of correlational research (see (Singer & Strasser, 2017) for a meta-analysis). Language proficiency is of essence when learning and exchanging on mathematics in and beyond school. In practice, the content of the mathematical curriculum is most often transferred orally from the teacher to the pupils, while training and assessment of the curriculum are done through textbooks and written assessments. Whereas core representations and computational skills seem to be less affected by language (Gelman & Butterworth, 2005), in many circumstances the mathematical problem needs to be derived from oral or textual input. This is naturally the case for de facto word problems (Hickendorff, 2013; Peng & Lin, 2019; Vilenius-Tuohimaa, Aunola, & Nurmi, 2008) but extends to any situation in which a mathematical problem is presented verbally. Insufficient mastery of the language of presentation will thus inevitably lead to difficulties in solving the mathematical problem. While this is true for any school population, the relationship between linguistic and mathematical ability becomes a true cause for concern in multilingual settings.

Indeed, linguistic competence in the language of instruction and assessment of mathematics varies significantly between native and immigrant populations (see e.g.(Bos, Tarelli, Bremerich-Vos, & Schwippert, 2012) for data from Germany or (R. Martin, Ugen, & Fischbach, 2013) for data from Luxembourg). These differences are consistently reported from around the globe. In South Africa for example, it has been shown that pupil's proficiency in English, which was not their native language, was a strong predictor of their performance in mathematics (Howie, 2003). Similar effects have been reported in Spanish-English dual language learners in the United states as early

as preschool (Méndez, Hammer, Lopez, & Blair, 2019). A German study by Saalbach and colleagues found that proficiency in the instructional language was an important predictor of mathematics achievement especially for pupils of low socio-economic status, which was confounded with having an immigration background, (Saalbach, Gunzenhauser, Kempert, & Karbach, 2016). Another study in Germany has led to the conclusion that not only performance in mathematics, but also learning gains over two years (i.e. grade four to grade six) are predicted by pupil's reading comprehension (Paetsch, Radmann, Felbrich, Lehmann, & Stanat, 2016). Moreover, differences in mathematics achievement between native and non-native speakers disappeared once performance in reading comprehension was controlled for. The results of these studies suggest that non-native population's underachievement in mathematics is the result of insufficient mastery of the vehicle language rather than insufficient mathematical reasoning abilities.

The hereby presented study sought to further explore the relationship between reading comprehension and math performance by examining this link in the context of multilingual school system. To this aim we conducted the current study in Luxembourg. Luxembourg is a highly multilingual country, using three official languages (Luxembourgish, German & French) for press, administrative, judicial and everyday communication. Beyond the three official ones, many more languages are spoken in Luxembourg's population due to a high immigration rate (47.5% foreigners in 2019, (STATEC, 2019)). These demographics are equally reflected in Luxembourg's public-school system. Indeed, only 36% of Luxembourg's school population speaks Luxembourgish at home (Ministère de l'Education nationale et de la Formation professionnelle, n.d.), with the second most spoken language being Portuguese (28%). On the other hand, primary education is held predominantly in German and Luxembourgish, an originally German dialect that

has since drawn significant influence from French, leading to an increasingly difficult language situation for most of the school population.

Reports from the EPSTAN Program and the Ministry of Education have repeatedly shown that the immigrant population consistently underachieves not only in learning the languages taught in school, but also in every other subject when compared to the Luxembourgish reference population, which achieves the best educational outcomes across the board (R. Martin et al., 2013; Ministère de l'Éducation nationale et de la Formation professionnelle, n.d.). This leads to a situation where immigrant children are prone to grade-retention (Organisation for Economic Co-operation and Development, 2015; Tillman, Guo, & Harris, 2006) and are more often oriented towards vocational curricula (28,9% vs. 17,9% of Luxembourgers, (Lenz, 2015)), thus restricting their access to higher education and limiting their chances of emancipation from a low socio-economic status. While this pattern has so far been attributed to the traditionally consistent relationship between SES and school achievement and the fact that immigrated populations are mostly located in the lower parts of the socio-economic spectrum in Luxembourg (R. Martin et al., 2013), the research presented so far suggests that competency in the school language(s) might play a greater role in the achievement differences than the unfavourable socio-economic starting situation.

Following up on these findings, we found it interesting and necessary to explore the aforementioned relationships between home language, school language and mathematics achievement in the multilingual setting of Luxembourg, using data from Luxembourg's national school monitoring program (Épreuves standardisées, EPSTAN, (R. Martin et al., 2013)). More specifically, we examined the influence of children's language profile on their language and math performances in third grade. Concretely, we compared the performance of three language groups (French, Portuguese and South-Slavic) to the Luxembourgish reference population on measures of

listening comprehension (German), reading comprehension (German) and mathematics (taught in German). Based on previously reported studies conducted in predominantly monolingual school systems, we hypothesized that non-native speakers compared to native speakers would underperform in German listening and reading comprehension as well as mathematics, that reading and listening comprehension would be crucial non-specific predictors of math performance and that the performance gaps in mathematics between native and non-native speakers would be driven to a large extent by differences in listening and reading comprehension in the language of instruction (German).

## **Methods**

### **Measures**

All measures stem from the EPSTAN national school monitoring data. Every student in Luxembourg participates in this large-scale assessment in first, third, fifth and ninth grade. In this study, we are looking at the data from pupils that were in first grade in 2014 and in third grade in 2016. This dataset was limited to students that did not repeat a grade from first to third grade. The EPSTAN metrics use a common scale with an average of 500 points and a standard deviation of 100 points for assessing all competencies. Assessment batteries were constructed by a group of experts including psychometricians and teachers based on the governmental learning goals for each grade and competence level.

**HISEI.** Highest socio-economic index of both parents. An ISEI-88 (Ganzeboom, De Graaf, & Treiman, 1992) Questionnaire was administered to both parents during data collection. HISEI reflects the highest of both parent's scores.

**Reading comprehension (READ).** Scores on standardized EPSTAN reading comprehension assessment. The reading comprehension assessment was comprised of a series of text-based short stories followed by multiple choice questions.

**Listening comprehension (LIST).** Scores on standardized EPSTAN listening comprehension assessment. The listening comprehension assessment was comprised of a series of pre-recorded short stories followed by pre-recorded multiple-choice questions. Answers were given on a paper booklet that contained answer choices for each question.

**Mathematics (MATH G1).** Scores on standardized EPSTAN mathematics assessment, first grade.

**Mathematics (MATH G3).** Scores on standardized EPSTAN mathematics assessment, third grade.

**Language background.** This variable was assessed by a student-questionnaire and reflects the language spoken with the student's mother. The questionnaire is comprised of a single question (*Which language do you speak most with your mother?*) and offers multiple choices. Only the Luxembourgish (LUX), French (FRE), Portuguese (PORT) and South Slavic (SLA) language groups were retained for this study. Other language groups were not sufficiently represented in the sample for robust analysis.

### **Sample**

In the initial dataset we counted 3941 students. First, we removed all students that had missing values in the language background, which left us with 3506 students. We then removed all students belonging to language groups other than Luxembourgish, French, Portuguese and South Slavic. This left us with a sample of 3156 students. Finally, all cases with missing data on any measure we considered for analysis were deleted listwise, leaving 2649 cases for analysis.

### **Analyses**

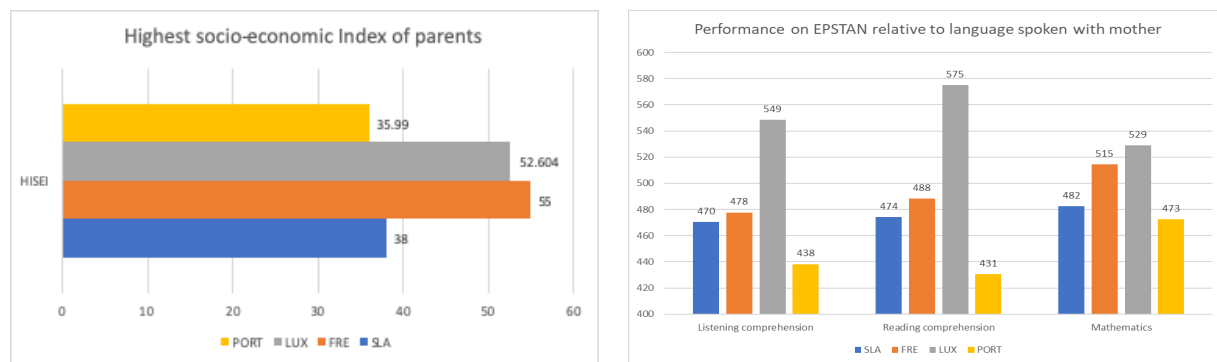
Regression modelling was done using JASP (JASP Team, 2018). Mediation analyses were completed using the PROCESS Macro for SPSS (Hayes, 2017).

## Results

TABLE 1. DESCRIPTIVES

|                  | HISEI (G3) |        |        |        | Listening comprehension (G3) |         |         |         | Reading comprehension (G3) |         |         |         | Mathematics (G3) |         |         |         |
|------------------|------------|--------|--------|--------|------------------------------|---------|---------|---------|----------------------------|---------|---------|---------|------------------|---------|---------|---------|
|                  | SLA        | FRE    | LUX    | PORT   | SLA                          | FRE     | LUX     | PORT    | SLA                        | FRE     | LUX     | PORT    | SLA              | FRE     | LUX     | PORT    |
| Mean             | 38.005     | 54.849 | 52.604 | 35.990 | 470.421                      | 477.705 | 548.543 | 438.257 | 474.265                    | 488.286 | 575.148 | 430.514 | 482.420          | 514.500 | 529.004 | 472.579 |
| Std. Deviation   | 12.273     | 14.010 | 14.608 | 13.702 | 64.667                       | 75.935  | 70.112  | 72.306  | 114.967                    | 127.929 | 129.444 | 108.297 | 95.905           | 100.864 | 104.830 | 94.280  |
| Skewness         | 0.894      | -0.663 | -0.399 | 0.760  | 0.215                        | 0.355   | 0.051   | 0.230   | 0.206                      | 0.163   | -0.139  | 0.477   | 0.311            | 0.275   | 0.152   | 0.213   |
| Kurtosis         | 0.729      | -0.783 | -1.105 | 0.168  | 0.669                        | 0.439   | 0.857   | -0.012  | 0.195                      | 0.356   | 0.043   | 0.501   | -0.193           | 0.089   | 0.685   | 0.384   |
| N (Valid)        | 126        | 433    | 1367   | 749    | 176                          | 499     | 1536    | 923     | 176                        | 500     | 1540    | 918     | 177              | 501     | 1544    | 925     |
| Mean age (years) | 8.43       | 8.33   | 8.35   | 8.39   | -                            | -       | -       | -       | -                          | -       | -       | -       | -                | -       | -       | -       |

FIGURE 1.1 & 1.2: HISEI, PERFORMANCE IN GERMAN LISTENING AND READING COMPREHENSION & MATHEMATICS BY LANGUAGE BACKGROUND.



A first look at the descriptive performance summaries revealed large differences in mean performance relative to language background for each assessed competence: The Luxembourgish native population outperformed the French language group, followed by the South-Slavic language group and finally the Portuguese language group. HISEI distribution followed the same pattern, except for the French language group which has a higher HISEI score than the Luxembourgish reference population. In a first step, using stepwise multiple linear regression, we examined the predictive role of German listening comprehension, reading comprehension, socio-economic status and language spoken with the mother on mathematics performance in third grade. Additionally, as the best predictor for future performance is previous performance, we included mathematics performance in grade one in order to provide the most complete regression model.

**TABLE 1.1: MODEL FIT MEASURES**

| MODEL | R    | R <sup>2</sup> | Adjusted R <sup>2</sup> | AIC      | BIC      | RMSE  | Overall Model Test |     |      |       |
|-------|------|----------------|-------------------------|----------|----------|-------|--------------------|-----|------|-------|
|       |      |                |                         |          |          |       | F                  | df1 | df2  | p     |
| 1     | 0.62 | 0.39           | 0.39                    | 30601.34 | 30618.98 | 79.66 | 1682.81            | 1   | 2637 | <.001 |
| 2     | 0.64 | 0.42           | 0.42                    | 30487.16 | 30510.68 | 77.93 | 938.26             | 2   | 2636 | <.001 |
| 3     | 0.72 | 0.52           | 0.52                    | 29993.61 | 30023.00 | 70.95 | 935.86             | 3   | 2635 | <.001 |
| 4     | 0.72 | 0.52           | 0.52                    | 29981.52 | 30016.79 | 70.76 | 708.91             | 4   | 2634 | <.001 |
| 5     | 0.72 | 0.52           | 0.52                    | 29959.09 | 30012.00 | 70.38 | 413.08             | 7   | 2631 | <.001 |

**TABLE 1.2: MODEL COMPARISONS**

| MODEL | Model | ΔR <sup>2</sup> | F      | df1 | df2  | p     |
|-------|-------|-----------------|--------|-----|------|-------|
| 1     | -     | 0.03            | 118.64 | 1   | 2636 | <.001 |
| 2     | -     | 0.10            | 544.31 | 1   | 2635 | <.001 |
| 3     | -     | 0.00            | 14.10  | 1   | 2634 | <.001 |
| 4     | -     | 0.01            | 9.50   | 3   | 2631 | <.001 |

**TABLE 1.3: MODEL COEFFICIENTS**

| MODEL | Predictor | Estimate | SE    | T     | p     | Stand. Estimate |
|-------|-----------|----------|-------|-------|-------|-----------------|
| 1     | Intercept | 151.81   | 8.96  | 16.94 | <.001 |                 |
|       | MATH G1   | 0.70     | 0.02  | 41.02 | <.001 | 0.62            |
| 2     | Intercept | 121.58   | 9.20  | 13.22 | <.001 |                 |
|       | MATH G1   | 0.66     | 0.02  | 38.87 | <.001 | 0.59            |
|       | HISEI     | 1.04     | 0.10  | 10.89 | <.001 | 0.17            |
| 3     | Intercept | 87.09    | 8.50  | 10.24 | <.001 |                 |
|       | MATH G1   | 0.52     | 0.02  | 31.22 | <.001 | 0.46            |
|       | HISEI     | 0.35     | 0.09  | 3.87  | <.001 | 0.06            |
|       | READ      | 0.27     | 0.01  | 23.33 | <.001 | 0.37            |
| 4     | Intercept | 66.03    | 10.17 | 6.49  | <.001 |                 |
|       | MATH G1   | 0.51     | 0.02  | 31.11 | <.001 | 0.46            |
|       | HISEI     | 0.30     | 0.09  | 3.26  | 0.001 | 0.05            |
|       | READ      | 0.24     | 0.01  | 16.14 | <.001 | 0.32            |
|       | LIST      | 0.08     | 0.02  | 3.75  | <.001 | 0.07            |
| 5     | Intercept | 27.99    | 12.61 | 2.22  | 0.027 |                 |
|       | MATH G1   | 0.51     | 0.02  | 30.89 | <.001 | 0.46            |
|       | HISEI     | 0.38     | 0.10  | 3.84  | <.001 | 0.06            |
|       | READ      | 0.24     | 0.01  | 16.62 | <.001 | 0.33            |
|       | LIST      | 0.13     | 0.02  | 5.47  | <.001 | 0.11            |
|       | FRE       | 16.68    | 4.20  | 3.97  | <.001 | 0.06            |
|       | PORT      | 18.96    | 4.11  | 4.61  | <.001 | 0.08            |
| SLA   | 17.91     | 6.87     | 2.61  | 0.009 | 0.04  |                 |

In the first model, we included only mathematics performance (grade 1) as predictor, resulting in 39% explained variance in third grade mathematics performance. In the second model, we added socio-economic status, leading to 42% explained variance. In the third model, we added reading comprehension to the prediction, leading to a 10% increase in explained variance. In the fourth and fifth model, we added listening comprehension and background language to the prediction, resulting in a statistically significant but only minimal increase in explained variance (<1%). The results of the final model (5) showed that after prior mathematics performance, reading comprehension was the strongest predictor of math performance in grade 3. Listening comprehension is the third largest predictor of math performance, but, due to its strong correlation



( $r=.7$ ,  $p<.001$ ) with reading comprehension, it did not improve the model by a relevant amount. All other predictors (background language, socio-economic status) were statistically significant, but their contribution to the prediction was only minimal once mathematics performance in first grade as well as reading and listening comprehension in third grade have been controlled for. A closer look at the average performance of each language group revealed that performance differences were much smaller in mathematics than they were in reading and listening comprehension. Considering that the mathematics tasks in third grade were presented with German written instructions, we tested if the differences in mathematics performance relative to language background would be mediated to some extent by performance in reading comprehension. Based on the final regression model, we then estimated a mediation model with background language as predictor, reading comprehension as mediator and mathematics performance (G3) as outcome variable.

TABLE 2: MEDIATION MODEL

| <i>Indirect effects</i>    | Effect        | Bootstrapped SE | Bootstrapped LLCI | Bootstrapped ULCI |
|----------------------------|---------------|-----------------|-------------------|-------------------|
| <i>PORT → READ → MATH*</i> | -66.09        | 2.84            | -71.84            | -60.67            |
| <i>FRE → READ → MATH*</i>  | -39.75        | 3.14            | -46.05            | -33.85            |
| <i>SLA → READ → MATH*</i>  | -46.16        | 4.45            | -55.11            | -37.58            |
| <i>Direct effects</i>      | <i>Effect</i> | <i>SE</i>       | <i>LLCI</i>       | <i>ULCI</i>       |
| <i>PORT → MATH*</i>        | 9.57          | 3.90            | 1.91              | 17.22             |
| <i>FRE → MATH*</i>         | 25.31         | 4.43            | 16.63             | 33.99             |
| <i>SLA → MATH</i>          | -1.24         | 6.76            | -14.48            | 12.01             |

Notes: \*=  $p<.05$ ; Unstandardized effects; 5000 Bootstrap samples for indirect effects.

The resulting model showed that there is was significant negative indirect effect of background language on mathematics performance, mediated by performance in reading comprehension. The indirect effect was larger than the direct effect for each language group. The indirect effect was strongest for the Portuguese language group, followed by the South-Slavic and finally the French language group. For the Portuguese and French language groups, there remained a significant

positive direct effect of language background while no significant direct effect remained for the South-Slavic group.

## **Discussion**

### **Summary of findings**

The aim of this study was to explore the predictive power of German reading and listening comprehension on mathematics performance in third grade in a sample of students that were in first grade in 2014 and third grade in 2016, i.e. who did not repeat a grade during their first two years of schooling. Concerning our first hypothesis, we found that the average performance of non-native speakers was below performance of their Luxembourgish peers in German listening and reading comprehension as well as mathematics. Performance differences were larger in the linguistic measures than they were in mathematics. Based on these observations and previous research (Méndez et al., 2019; Paetsch et al., 2016; Saalbach et al., 2016) we hypothesized that linguistic measures would be significant and relevant predictors of mathematics performance. Using stepwise multiple linear regression, we showed that after controlling for prior mathematics performance, reading comprehension was the sole statistically significant and practically relevant predictor of mathematics performance in third grade. While background language, listening comprehension and socio-economic status were statistically predictive, their contribution to the prediction of mathematics performance in third grade can be considered residual. Finally, we hypothesized that the differences between mathematics performance of non-native language speakers and Luxembourgers would be driven by performance differences on measures of reading and listening comprehension in the language of instruction and assessment. After the initial regression model, we thus estimated a mediation model with background language as predictor, reading comprehension as mediator and math performance in grade three as outcome variable. Due

to its strong correlation with reading comprehension and its marginal improvement of the regression model, listening comprehension was not included in the mediation model.

The results of the mediation showed that the relationship between language background and mathematics performance was indeed mediated by performance in reading comprehension across all non-native language groups. The mediation was complete for South-Slavic speakers as no significant direct effect of background language on grade three math performance remained. There remained a significant positive direct effect for French and Portuguese speakers after the mediation through reading comprehension. At first glance, this result was surprising as it suggests that after controlling for the indirect effects of reading comprehension, French and Portuguese speakers in this sample would outperform Luxembourgish native speakers in grade three mathematics. However, considering the fact that this sample only included those children that went from grade one to grade three without grade repetition and that being of foreign language background is a strong predictor for grade repetition (Lenz, 2015; Organisation for Economic Co-operation and Development, 2015), our results suggested that non-native speakers in our sample must have compensated their a-priori disadvantage by relying on other cognitive resources to succeed in school. In other words, the non-native speakers in our sample seemed to be inherently stronger in mathematics than their average Luxembourgish peer, an advantage that was in sum suppressed by the dominating negative effect of reading comprehension. This observation will be easy to verify once the complete dataset including grade-repeaters becomes available for analysis.

Our results thus corroborated previously presented studies which have shown that competency in the school's language(s) is a significant and strong predictor of mathematics performance and that this relationship is especially unfavourable for non-native speakers (Méndez et al., 2019; Paetsch et al., 2016; Saalbach et al., 2016) . Additionally and in line with (Paetsch et al., 2016), we have

shown that performance differences in mathematics were heavily reduced (Portuguese and French speakers) or disappeared entirely (South-Slavic speakers) when reading comprehension was controlled for. In other words, differences in grade three mathematics performance between non-native speakers and native Luxembourgers seem to be largely or entirely due to their underachievement in German reading comprehension.

### **Implications and future studies**

Finally, and strikingly, our results seem to suggest that only non-native speakers that show above average competency in mathematics succeed in Luxembourg's primary education without grade repetition, while still underperforming due to the suppressing effects of their reading comprehension in German. While our sample only included grade non-repeaters, we hypothesize that the damaging effects of instructional language (in)comprehension are likely to be stronger in grade-repeaters and that they contribute significantly to grade repetition itself. Indeed, many studies have shown that children with an immigration background and/or low socioeconomic status have less vocabulary knowledge in the language of instruction than their native peers (Biemiller & Slonim, 2001; Hart & Risley, 1995; Perfetti, McKeown, & Kucan, 2010; Rathvon, 2008). The same children are also largely overrepresented in grade retention, both in Luxembourg (Klapproth & Schaltz, 2015; Lenz, 2015; R. Martin et al., 2013) as well as internationally (Organisation for Economic Co-operation and Development, 2015; Tillman et al., 2006). In future studies, it would thus be crucial to identify the skills and strategies used by those that succeed despite their linguistic shackles in order to improve educational outcomes for all foreign speakers in multilingual educational settings.

More generally and beyond mathematics, it is easy to imagine that linguistic competence plays a similar role in many other school subjects across the elementary and secondary curriculum and

that insufficient mastery of the school's vehicle language can have long lasting and severe negative effects on motivation to learn and to participate. Indeed, a study conducted in a large sample (N=3261) of American high school students found that grade retention had negative effects on academic self-concept, self-esteem, homework completion and presence in school (A. J. Martin, 2011).

### **Conclusion**

Taken together, available research and our own results have shown that insufficient mastery of the language of instruction leads to academic underachievement both in measures of reading comprehension as well as mathematics. Immigrant populations are most affected and are thus overrepresented in grade retention. Sadly, grade retention leads to lower self-esteem and motivation to take part in the learning process. The resulting situation is a negative feedback loop that originates in insufficient language mastery and culminates in overall worse educational outcomes for those that speak a different language than their school. Considering that a democratic school system should aim to counteract disadvantageous starting conditions, significant investments in matching non-native speaker's performance in the language of instruction before and during primary education is thus likely to result in cumulative beneficial effects on non-native speaker's academic achievement and overall well-being and motivation during their school curriculum.

## References

- Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology, 93*(3), 498.
- Bos, W., Tarelli, I., Bremerich-Vos, A., & Schwippert, K. (2012). *IGLU 2011 Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Waxmann Verlag.
- Ganzeboom, H. B., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21*(1), 1–56.
- Gelman, R., & Butterworth, B. (2005). Number and language: How are they related? *Trends in Cognitive Sciences, 9*(1), 6–10. <https://doi.org/10.1016/j.tics.2004.11.004>
- Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.
- Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Publications.
- Hickendorff, M. (2013). The Language Factor in Elementary Mathematics Assessments: Computational Skills and Applied Problem Solving in a Multidimensional IRT Framework. *Applied Measurement in Education, 26*(4), 253–278. <https://doi.org/10.1080/08957347.2013.824451>
- Howie, S. J. (2003). Language and other background factors affecting secondary pupils' performance in Mathematics in South Africa. *African Journal of Research in Mathematics, Science and Technology Education, 7*(1), 1–20. <https://doi.org/10.1080/10288457.2003.10740545>
- JASP Team. (2018). *JASP (Version 0.9)[Computer software]*. Retrieved from <https://jasp-stats.org/>
- Klapproth, F., & Schaltz, P. (2015). Who is retained in school, and when? Survival analysis of predictors of grade retention in Luxembourgish secondary school. *European Journal of Psychology of Education, 30*(1), 119–136.
- Lenz, T. (2015). *Bildungsbericht Luxemburg*.
- Martin, A. J. (2011). Holding back and holding behind: Grade retention and students' non-academic and academic outcomes. *British Educational Research Journal, 37*(5), 739–763. <https://doi.org/10.1080/01411926.2010.490874>
- Martin, R., Ugen, S., & Fischbach, A. (2013). *Épreuves Standardisées—Bildungsmonitoring Luxemburg*.
- Méndez, L. I., Hammer, C. S., Lopez, L. M., & Blair, C. (2019). Examining language and early numeracy skills in young Latino dual language learners. *Early Childhood Research Quarterly, 46*, 252–261. <https://doi.org/10.1016/j.ecresq.2018.02.004>
- Ministère de l'Éducation nationale et de la Formation professionnelle. (n.d.). *The key figures of the national education* (No. 978-99959-1-129-4). Retrieved from <http://www.men.public.lu/catalogue-publications/themes-transversaux/statistiques-analyses/chiffres-cles/2016-2017/en.pdf>
- Organisation for Economic Co-operation and Development. (2015). *Immigrant students at school: Easing the journey towards integration*. OECD Publishing.
- Paetsch, J., Radmann, S., Felbrich, A., Lehmann, R., & Stanat, P. (2016). Sprachkompetenz als Prädiktor mathematischer Kompetenzentwicklung von Kindern deutscher und nicht-

- deutscher Familiensprache. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 48(1), 27–41. <https://doi.org/10.1026/0049-8637/a000142>
- Peng, P., & Lin, X. (2019). The relation between mathematics vocabulary and mathematics performance among fourth graders. *Learning and Individual Differences*, 69, 11–21. <https://doi.org/10.1016/j.lindif.2018.11.006>
- Perfetti, C., McKeown, M., & Kucan, L. (2010). Decoding, vocabulary, and comprehension. *Bringing Reading Research to Life*, 291–303.
- Rathvon, N. (2008). *Effective school interventions: Evidence-based strategies for improving student outcomes*. Guilford Press.
- Saalbach, H., Gunzenhauser, C., Kempert, S., & Karbach, J. (2016). Der Einfluss von Mehrsprachigkeit auf mathematische Fähigkeiten bei Grundschulkindern mit niedrigem sozioökonomischen Status. *Frühe Bildung*, 5(2), 73–81. <https://doi.org/10.1026/2191-9186/a000255>
- Singer, V., & Strasser, K. (2017). The association between arithmetic and reading performance in school: A meta-analytic study. *School Psychology Quarterly*, 32(4), 435–448. <https://doi.org/10.1037/spq0000197>
- STATEC. (2019). *Le Luxembourg en chiffres* (No. 1019–6471). Retrieved from <https://statistiques.public.lu/catalogue-publications/luxembourg-en-chiffres/2019/luxembourg-chiffres.pdf>
- Tillman, K. H., Guo, G., & Harris, K. M. (2006). Grade retention among immigrant children. *Social Science Research*, 35(1), 129–156. <https://doi.org/10.1016/j.ssresearch.2004.07.001>
- Vilenius-Tuohimaa, P. M., Aunola, K., & Nurmi, J.-E. (2008). The association between mathematical word problems and reading comprehension. *Educational Psychology*, 28(4), 409–426. <https://doi.org/10.1080/01443410701708228>

## **Study report 2: Taking language out of the equation: The assessment of basic math competence without language<sup>1</sup>**

*(published in *Frontiers in Psychology, Developmental Psychology*, 2018)*

Max Greisen, Caroline Hornung, Tanja Gabriele Baudson, Claire Muller, Romain Martin &  
Christine Schiltz

### **Abstract**

While numerical skills are fundamental in modern societies, some estimated 5–7% of children suffer from mathematical learning difficulties (MLD) that need to be assessed early to ensure successful remediation. Universally employable diagnostic tools are yet lacking, as current test batteries for basic mathematics assessment are based on verbal instructions. However, prior research has shown that performance in mathematics assessment is often dependent on the testee's proficiency in the language of instruction which might lead to unfair bias in test scores. Furthermore, language-dependent assessment tools produce results that are not easily comparable across countries. Here we present results of a study that aims to develop tasks allowing to test for basic math competence without relying on verbal instructions or task content. We implemented video and animation-based task instructions on touchscreen devices that require no verbal explanation. We administered these experimental tasks to two samples of children attending the first grade of primary school. One group completed the tasks with verbal instructions while another group received video instructions showing a person successfully completing the task. We assessed task comprehension and usability aspects both directly and indirectly. Our results suggest that the non-verbal instructions were generally well understood as the absence of explicit verbal instructions did not influence task performance. Thus we found that it is possible to assess basic math competence without verbal instructions. It also appeared that in some cases a single word in a verbal instruction can lead to the failure of a task that is successfully completed with non-verbal instruction. However, special care must be taken during task design because on rare occasions non-verbal video instructions fail to convey task instructions as clearly as spoken language and thus the latter do not provide a panacea to non-verbal assessment. Nevertheless, our findings provide an encouraging proof of concept for the further development of non-verbal assessment tools for basic math competence.

---

<sup>1</sup> The prettier, published print version of this report is available here:  
<https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01076/pdf>



## Introduction

Basic counting and arithmetic skills are necessary to manage many aspects of life. Although primary education focuses on these subjects, 5–7% of the general population suffer from mathematical learning difficulties (MLD) (Butterworth, Varma, & Laurillard, 2011), often leading to dependence on other people or technology.

Early diagnostic is key to remedying MLD (Gersten, Jordan, & Flojo, 2005). Basic mathematical skills, e.g., counting, quantity comparison, ordering, and simple arithmetic are the strongest domain-specific predictors for mathematical performance in later life (Desoete, Ceulemans, Roeyers, & Huylebroeck, 2009; Hornung, Schiltz, Brunner, & Martin, 2014; Jordan, Glutting, & Ramineni, 2010; LeFevre et al., 2010). Valid MLD assessments exist in various forms and for all ages (Aster, Bzufka, & Horn, 2009; Haffner, Baro, Parzer, & Resch, 2005; Noël, Grégoire, & Nieuwenhoven, 2008; Ricken, Fritz, & Balzer, 2011; Schaupp, Holzer, & Lenart, 2007; van Luit, van de Rijt, & Hasemann, 2001). However, all of them rely on verbal instructions and (in part) verbal tasks.

This is a problem. First, performance in mathematical tests is predicted by the pupils' proficiency in the instruction language (Abedi & Lord, 2001; Hickendorff, 2013; Paetsch, Radmann, Felbrich, Lehmann, & Stanat, 2016). Others have shown that the complexity of mathematical language content of items is predictive of performance (Haag, Heppt, Stanat, Kuhl, & Pant, 2013; Purpura & Reid, 2016). Diagnostic tools for MLD relying on language may therefore significantly bias performance in test-takers that are not proficient in the test language, leading to invalid results (see Ortiz & Dynda, 2005; Scarr-Salapatek, 1971 for similar considerations concerning intelligence testing). Furthermore, the match between math learners' language profiles and the linguistic

context in which mathematical learning takes place plays a critical role in the acquisition and use of basic number knowledge. Matching language contexts improve bilinguals' arithmetic performance in their second language (Van Rinsveld et al., 2016), and neural activation patterns of bilinguals solving additions differ depending on the language they used, suggesting different problem-solving processes (Van Rinsveld, Dricot, Guillaume, Rossion, & Schiltz, 2017).

In linguistically homogeneous societies, where the mother tongue of most primary school children matches the language of instruction and assessment tools, this is less of a problem. It is however critical in societies with high immigration and, therefore, linguistically diverse primary school populations. In Luxembourg, for instance, where the present project is located, currently 62% of the primary school students are not native Luxembourgish speakers (Ministère de l'éducation nationale de l'enfance et de la Jeunesse, 2015). Due to migration, multilingual classrooms are steadily becoming the rule rather than the exception (e.g. from 42% foreign speakers in 2004 to 62% in 2014) (Ministère de l'éducation nationale de l'enfance et de la Jeunesse, 2015), likely increasing the urgency of the problem in the future.

Even in traditionally multilingual contexts, diagnostic tools for the assessment of basic numerical abilities in early childhood are available in a few selected languages only, usually those that are best understood by most, yet not necessarily all students. As described above, this leads to invalid conclusions about non-native speakers' ability. In addition, comparisons between different tools and even different linguistic versions of the same tool are difficult because the norms they are based on are usually collected in linguistically homogenous populations and can thus not be extrapolated to populations with different linguistic profiles.

The present study originated in a project that aims to develop a test of basic numerical competencies which circumvents linguistic interference by relying on nonverbal instructions and task content. In the field of intelligence assessment, the acknowledgement of language interference has led to the development of numerous nonverbal test batteries (Cattell & Cattell, 1973; Feis, 2010; Lohman & Hagen, 2001; Naglieri, 2003). However, these tools tackle only the problem of verbal tasks, not of verbal instructions. The same is true for numeracy assessment. Although many test batteries (e.g. Tedi-MATH, Zareki-R, ERT0+, OTZ, Marko-D, to name a few) use nonverbal and non-symbolic tasks (e.g., arithmetic, counting, or logical operations on numbers), they still rely on verbal instructions, which may limit the testee's access to the content. Linguistic simplification of mathematics items can improve performance for language minority students (Haag, Heppt, Roppelt, & Stanat, 2014). However, we think that for many simple tasks, verbal content and instructions can be avoided altogether. These tasks that children of (above-) average ability usually solve easily are crucial to the diagnosis of MLD, as they allow for a differentiation of children's numerical abilities at the bottom end of the ability distribution. Hence, nonverbal assessment of basic mathematical skills may help identify children in need of intervention at an early age and independently of their linguistic abilities, thus reducing the bias that common assessments often suffer from. Comparable approaches have been taken in the field of intelligence testing for the hearing-impaired, in which pantomime instructions for the Wechsler performance scale have been explored (Braden & Hannah, 1998; Courtney, Hayes, Couch, & Frick, 1984).

With this goal in mind, using available test batteries and the official study plan (MENFP, 2011) as a reference for task content and design, we developed different task types for which a valid nonverbal computerized implementation was possible. Governmental learning goals for preschool mathematics include but are not limited to: Ability to represent numbers with concrete material,

ordering abilities (range 0-20), definition, resolution & interpretation of an arithmetical (addition/subtraction) problem based on images and mental addition/subtraction (range 0-20).

The tasks we developed encompass and measure all the above competencies: Quantity representation, ordering abilities as well as symbolic and non-symbolic arithmetic. We chose to add a quantity comparison task as it has been found to be one of the most consistent predictors of later math performance (e.g. Brankaer, Ghesquière, & De Smedt, 2017; De Smedt, Verschaffel, & Ghesquière, 2009; Nosworthy, Bugden, Archibald, Evans, & Ansari, 2013; Sasanguie, Van Den Bussche, & Reynvoet, 2012; see Schneider et al., 2017 for a meta-analysis). Instead of using verbal instructions, we convey task requirements with the use of videos that show successful task completion and interactions with the tasks from a first-person point of view. Prior research has shown improved performance in a computerized number-line estimation task for participants who viewed videos of a model participant's eye gaze or mouse movements, compared to control conditions both with and without anchor points (Gallagher-Mitchell, Simms, & Litchfield, 2017).

The aims of the present study were to evaluate whether basic math competence can be assessed on a tablet PC without language instructions and whether the mode of instruction affects performance. To this end, we designed a set of computerized tasks based on validated assessments measuring basic non-symbolic and symbolic mathematical abilities, which were administered either nonverbally (using computer-based demonstrations; experimental condition) or traditionally (using verbal instructions; control condition). Because young school children's attention span is limited (Pellegrini & Bohn, 2005), some of the tasks were administered to one sample (Sample 1) in a first study and the remainder to another sample (Sample 2) in a second study 5 months later. First, considering that the nonverbal mode of instruction was new, we examined possible difficulties both directly (understanding of feedback and navigation) and indirectly (repeated

practice sessions). Second, though tasks were derived from field-tested assessments, performance on the new tasks was correlated with performance on two standardized and one self-developed measure in order to ensure task validity. Third, we examined students' performance compared by condition and overall. Considering the novelty of the nonverbal task administration, we did not specify directed hypotheses but examined this question exploratively.

## Methods

### Participants

Table 1: Participant demographics, language background and SES.

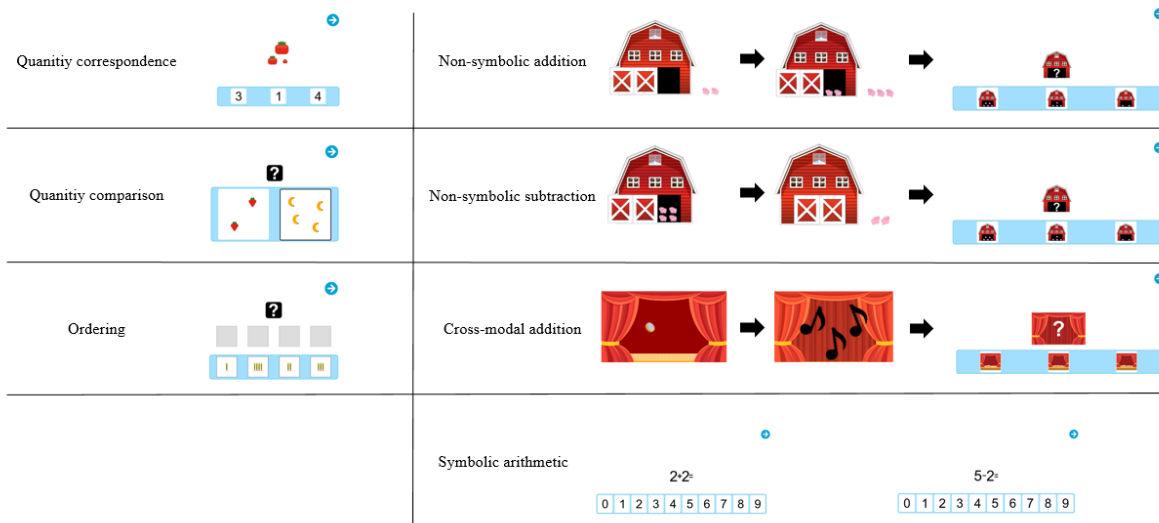
| Sample   | N   | % girls | Age          | Schooling          | Language |      |              | ISEI       |
|----------|-----|---------|--------------|--------------------|----------|------|--------------|------------|
|          |     |         |              |                    | % RO     | % LG | % OT         |            |
|          |     |         | <i>M(SD)</i> |                    |          |      | <i>M(SD)</i> |            |
| Sample 1 | 96  | 53.1    | 6y;7m (4m)   | Grade 1 (5 weeks)  | 30.2     | 55.2 | 14.6         | 50 (6.3)   |
| Sample 2 | 141 | 48.2    | 7y;2m (4m)   | Grade 1 (28 weeks) | 55.3     | 34.8 | 9.9          | 47.9 (7.2) |

*Notes.* % RO = percentage of Romance language speaking children (French, Portuguese, Italian, Spanish). % LG = percentage of children speaking Luxembourgish or German. %OT = percentage of children with other language backgrounds (Slavic, English). ISEI = International Socio-Economic Index of Occupational Status.

Table 1 shows participant demographics, language background and socio-economic status. The ISEI is the International Socio-Economic Index of Occupational Status, used in large scale assessments. It ranges from 16 (e.g. agricultural worker) to 90 (e.g. judge). An average ISEI of 50 will thus indicate above average socio-economic status. As we could not directly assess socio-economic status in our studies, ISEI was estimated based on the communes in which the studies took place. This data is publicly available and in Luxembourg the communes average ISEI ranges from 35 to 65. All participants were recruited from first grade in Luxembourg's primary schools with the authorization of the Ministry of Education and the directors of the participating school sectors. Participants from the first sample were tested after 5 weeks of schooling while participants from the second sample were tested after 28 weeks of schooling. Teachers interested to participate in the study with their classes received information and consent letters for the pupil's legal representatives. Only pupils whose parents consent was obtained participated in this study. All children in Luxembourg spend two obligatory years in preschool and about a third of them participate in an optional third year of preschool prior to the two mandatory years (Lenz, 2015).

## Materials

Figure 1. Example Images of the experimental tasks



**Experimental Tasks.** As mentioned, the two samples received different types of tasks. In the following, all task types will be described in order of their administration. The number in parentheses after each task name indicates the sample it was administered to. Example images for each task are presented in figure 1.

**Quantity correspondence (S1).** The first task required determination of the exact quantity of the target display and choosing the response display with the corresponding quantity (both ranging from 1 to 9). Each item consisted of a target quantity displayed at the centre of the screen (stimulus). The nature of the quantity was varied and was either non-symbolic (based on real objects [fruit], abstract [dot collections]) or symbolic (Arab numerals). In the lower part of the screen, three different quantities were displayed to the participant from which he/she was to choose the one corresponding to the stimulus (multiple-choice images). The item pool consisted of five subgroups of items containing four items each:

1. Non-symbolic, *identical* objects for stimulus and multiple-choice images
2. Non-symbolic, *different* objects for stimulus and multiple-choice images
3. Non-symbolic, collections of black *dots* of variable sizes and configurations
4. Symbolic, *Arab numerals* in both stimulus and multiple-choice images
5. Mixed (combinations of the preceding characteristics)

Image characteristics (object area, total occupied area, etc.) were manually randomized but not systematically controlled for.

**Quantity comparison (S1).** The second task required determining and choosing the larger of two quantities (range: 1–9) displayed at the centre of the screen. The nature of the quantities was varied similarly to the first task:

1. Non-symbolic, each quantity being composed of *different objects* (4 items)
2. Non-symbolic, each quantity being composed of collections of *black dots* of variable sizes and configurations (4 items)
3. Symbolic, at least one of the two displays showing an *Arabic numeral* (4 items)

**Ordering (S1).** The third task required reordering 4 images by increasing quantities (range 1–9). The characteristics were divided into 2 subgroups, represented by 4 items each:

1. Ordering based on non-symbolic quantity
2. Ordering based on numerical symbols (Arabic digits)

**Non-symbolic addition (S2).** The first task required to solve a non-symbolic addition problem. Participants saw an animation of 1–5 pigs entering a barn. The barn door closed. Then, the door opened again, and 1–5 more pigs entered the barn. The door closed again. The result range included the numbers from 3 to 8 only. In the **non-symbolic answer** version of this task (3 items),



participants were then presented with three images containing an open barn with pigs inside. Their task was to choose the image showing the total number of pigs left in the barn. In the **symbolic answer** version of the task (3 items), participants selected the correct number of pigs from an array of numerals from 1 to 9 in ascending order to choose from.

**Non-symbolic subtraction (S2).** The second task required solving a non-symbolic subtraction problem using the same pigs-and-barn setting described above. Participants were shown an animation of an open barn containing some pigs, after which some pigs left and the barn door closed. The minimum number of pigs displayed in a group was 2, the maximum was 9. The result range was from 1 to 6. Symbolic and non-symbolic answer versions (3 items each) were the same as above.

**Crossmodal addition (S2).** The third task for Sample 2 required solving a crossmodal addition problem using visual and auditory stimuli. Participants saw an animation of coins dropping on the floor, each one making a distinctive sound. A curtain was then closed in front of the coins. More coins dropped, but the curtain remained closed. Participants could only hear but not see the second set of coins falling. Their task was to choose the total amount of coins on the floor, both the ones they saw and heard and the ones they only heard but did not see falling. The minimum number of coins displayed / heard was 1, the maximum was 5. The result range was from 3 to 7. In the **non-symbolic answer** version of this task (3 items), participants were presented with three images showing coins on the floor with an open curtain. Their task was to choose the image showing the total number of coins that are now on the floor. In the **symbolic answer** version of the task (3 items), participants were presented with an array of numerals from 1 to 9 in ascending order to choose from.

This task aimed to assess numerical processing at a crossmodal level, requiring a higher level of abstraction than unimodal tasks like the non-symbolic addition and subtraction tasks where only visual information is processed before answering the question. The addition of discrete sounds as stimuli adds a layer of abstraction that is not present in the other addition tasks (symbolic or non-symbolic) and ensures that responses must be based on a *truly abstract number sense, capable of representing any set of discrete elements* (Barth, Kanwisher, & Spelke, 2003), independently from its physical nature and prior cultural learning of number symbols.

**Symbolic arithmetic: addition & subtraction (S2).** In this task, participants had to solve traditional symbolic arithmetic problems in the range of 0 to 9, both addition (6 items) and subtractions (6 items), shown at the centre of the screen. The answer format in this task was symbolic only, i.e., participants were presented with an array of numerals from 1 to 9 in ascending order below the problem to choose their answer from.

### **Observation and Interview Sheets**

To examine the usability of instructions and task presentation, test administrators collected information about participants' behaviour during testing through semi-structured observation and interview sheets. Of special interest were the observations about the general use of the tablet and the tool's navigational features as well as participants' understanding of both video and verbal instructions and feedback elements in both groups.

The following questions (yes-no format) were answered for each participant and task: (1) Did the participant understand the purpose of the smiley? (2) Did the participant understand the use of the blue arrow as a navigational tool? To this aim, the test administrators asked the participants to

describe the task, the role of the smiley, and the role of the arrow and evaluated that answer as a 'Yes' or a 'No'. These questions were followed by empty space for comments.

### **Demographics and Criterion Validation Tasks**

After completion of the digitally administered tasks, all children received a paper notebook containing a demographic questionnaire as well as some control tasks. The questionnaire collected basic demographic data (age, gender, language spoken with mother). Control tasks were included to examine the criterion validity of the experimental tasks and were administered to both samples. The paper pencil control tasks were:

- ***TTR (Tempo Test Rekenen)*** (De Vos, 1992): a classical standardized measure of speeded arithmetic performance. Participants had 60 seconds for each subtest. Arithmetic difficulty increased systematically within each subtest list, with operands and results in the range of 1–100. As multiplication and division were not part of the participant's curriculum at that age, we used the addition and subtraction subtests only.
- ***"How many animals?"(Counting and transcoding)***: Since all of our experimental task assume basic counting skills, we included this self-developed counting task, in which ten paper sheets displaying a randomly arranged variable number of animals (range: 3–19) were presented successively to the participants, who reported how many animals they saw. Their oral answer was noted on a coding sheet by the test administrators. Furthermore, participants wrote down their answer on a separate coding sheet included in the participant notebook. This resulted in two separate measures: one for counting (oral) and one for transcoding ability (written).

- ***SYMP (Symbolic magnitude processing test)*** (Brankaer et al., 2017): a standardized measure of symbolic number comparison performance (1- and 2-digit, ranging from 1–10 and from 12–99, respectively). It includes a motor speed control task requiring participants to cross out the black shape in pairs of black/white shapes. Participants had 30 seconds for each subtest. Although number comparison abilities assessed by the SYMP test do not strictly constitute a measure of curricular learning goals, we choose to include it due to its well-recognized power to predict later differences in standardized mathematical tests and distinguish children with MLD from typically developing peers (see Schneider et al., 2017 for a meta-analysis). In contrast to the TTR scales and the counting task, correlation with the SYMP does not inform on the ability of our tasks to predict children’s achievement on higher level learning goals but allows to compare performance in our tasks to another low-level predictor of later math competence.

## **Design and Procedure**

**Experimental Design.** To evaluate comprehensibility and effectiveness of the video instructions in comparison to classical verbal instructions, we implemented a between-group design in the two samples. All children solved the tasks on tablet computers, but under two different conditions. In the experimental condition (“nonverbal condition”), instructions were conveyed through a video of a person performing specific basic mathematical tasks, followed by a green smiley indicating successful solution of the task. Importantly, children did not receive any verbal instructions in the experimental condition. In the control condition (“verbal condition”), children received verbal instructions in German, the official instruction language for Mathematics in elementary schools in Luxembourg. Analogous to usual classroom conditions, test administrators read the instructions aloud to the children. In both conditions, tasks were presented visually on tablet computers, either

through static images or animated “short stories”. In both samples, one group was allocated to the experimental “nonverbal” condition without language instructions and the other group was assigned to the “verbal” condition, respectively.

**Task Presentation.** The three main tasks for Sample 1 were presented on iPads using a borderless browser window. Two children were tested simultaneously. They were connected to a local server through a secured wireless network set up by the research team at each school to store and retrieve data. The tasks were implemented using proprietary web-based assessment-building software under development by the Luxembourg Centre for Educational Testing. Sample 2 worked on Chromebooks instead of iPads. The advantage of Chromebooks is that they are relatively inexpensive, are optimized for web applications, and provide both touchscreen interactivity and a physical keyboard when necessary. Four children were tested simultaneously to speed up data collection.

After the initial setup of the hardware (server, wireless connection), participants were called into the test room in groups of two (Sample 1) or four (Sample 2) and seated individually on opposite sides of the room, allowing to run multiple test sessions simultaneously. Participants were randomly assigned to one of two groups. A trained test administrator supervised each participant during the test session. Since the tasks for Sample 2 used audio material, participants were provided with headphones, which they wore during the video instructions and the tasks.

Both samples were presented with either nonverbal or verbal instructions. In the nonverbal condition (experimental group), each participant was shown three items, with the exception of the comparison task, where ten instruction items were given to account for the less salient nature of the implicit “Where is more?” instruction. The video also clarified how to proceed to the next item

by the person touching a blue arrow pointing rightwards on the top right corner of the screen, after which a new item was loaded. In the verbal condition (control group), the test administrator read the standardized oral instructions to the participant in German, thus mimicking traditional teaching and test situations. The instruction was repeated by the test administrator while the first practice item was displayed to facilitate the hands-on understanding of the task. After the instruction, participants were given three practice items with the same smiley-type feedback they had just witnessed (a happy green face for correct answers, an unhappy red face for wrong answers). After successful completion of the three practice items, the application moved on to the test items. If one or more answers were wrong, all three practice items were repeated once, including those that had been solved correctly in the first trial. At the end of this second run, the application moved on to the test items, even if one or more practice items had still been answered incorrectly. After each practice session, an animation showing a traffic light switching from red to green was displayed to notify children that the test was about to start.

At the end of the three tasks, a smiley face was displayed thanking the participants for their efforts. At the end of the individual testing sessions, all participants were regrouped in their classroom to complete the pen-and-paper measures instructed orally by the test administrators.

**Scoring.** Scores from symbolic and non-symbolic subgroups of items in most *experimental tasks* were averaged and operationalized as POMP (percentage of maximum performance) scores (Cohen, Cohen, Aiken, & West, 1999), giving rise to two scores in each task. The exception was the symbolic arithmetic task in Sample 2, which by its nature included only symbolic answer formats, but offered both addition and subtraction items, producing one score for each operation type. All scores from the *criterion validation tasks* are expressed as POMP scores.

## Results

In line with our research questions outlined in the introduction, we will first report findings on participants' difficulties by experimental condition, as usability represents an important prerequisite. Results on the directly assessed difficulties will focus on understanding of feedback and navigation, whereas indirectly assessed difficulties comprise findings on repeated practice. This is followed by descriptive analyses including scale quality, tests of normality, and scale intercorrelations. As we also examined the convergent validity of our tasks (another prerequisite), which were based on existing measures, we subsequently report findings on the correlations with the external measures, i.e., the paper pencil tests (see Materials section). Finally, we will compare performance by experimental condition.

### Observation data

Table 2: Directly assessed difficulties by experimental condition

| Sample | Task type                | Condition | Smiley | $\chi^2$ | df | p     | Navigation | $\chi^2$ | df | p     |
|--------|--------------------------|-----------|--------|----------|----|-------|------------|----------|----|-------|
| 1      | Quantity correspondence  | Verbal    | 46/46  | 1.01     | 1  | 0.315 | 45/46      | 6.03     | 1  | 0.014 |
|        |                          | Nonverbal | 45/46  |          |    |       | 38/46      |          |    |       |
|        | Quantity comparison      | Verbal    | 46/46  | 1.01     | 1  | 0.315 | 46/46      | 1.01     | 1  | 0.315 |
|        |                          | Nonverbal | 45/46  |          |    |       | 45/46      |          |    |       |
|        | Ordering                 | Verbal    | 46/46  | 1.03     | 1  | 0.309 | 46/46      | 1.06     | 1  | 0.304 |
|        |                          | Nonverbal | 44/45  |          |    |       | 43/44      |          |    |       |
| 2      | Non-symbolic addition    | Verbal    | 70/70  | 3.02     | 1  | 0.082 | 69/70      | 5.82     | 1  | 0.016 |
|        |                          | Nonverbal | 68/71  |          |    |       | 62/70      |          |    |       |
|        | Non-symbolic subtraction | Verbal    | 70/70  | n.a.     |    |       | 68/69      | 1.02     | 1  | 0.312 |
|        |                          | Nonverbal | 71/71  |          |    |       | 70/70      |          |    |       |
|        | Cross-modal addition     | Verbal    | 70/70  | 1.02     | 1  | 0.312 | 68/69      | 1.04     | 1  | 0.309 |
|        |                          | Nonverbal | 71/71  |          |    |       | 70/70      |          |    |       |
|        | Symbolic arithmetic      | Verbal    | 69/69  | n.a.     |    |       | 68/68      | n.a.     |    |       |
|        |                          | Nonverbal | 71/71  |          |    |       | 71/71      |          |    |       |

Note: n.a. = not applicable due to 1-level factor.

**Directly assessed difficulties: understanding of feedback and navigation.** The following results are based on the observation sheets for each task. Table 2 shows the number of participants that understood the smiley as a feedback symbol and the number of participants that understood the

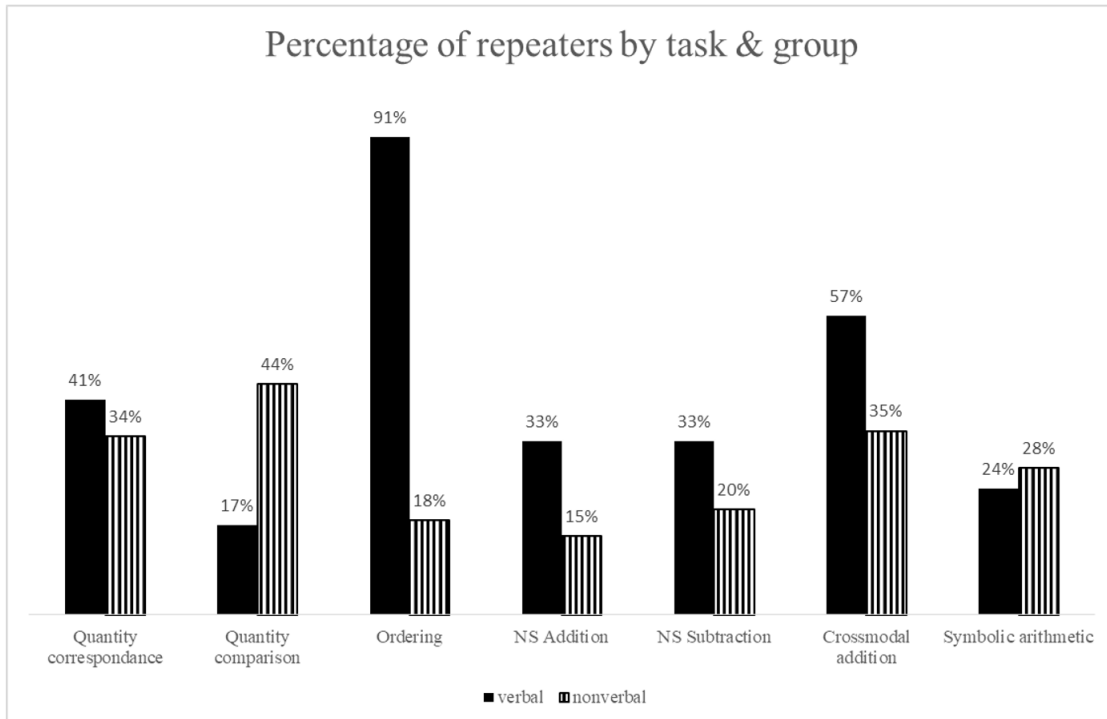
arrow as a navigational interface element. Discrepancies in the total number of participants are due to missing data points for some participants.

Summarily, we observed that all but a few participants had correctly understood the feedback symbols and the navigation arrow from the start.

Table 3: Indirectly assessed difficulties (practice repetition) by experimental condition

| Sample | Task type                | Condition | Repeater |     | $\chi^2$ | df | p     |
|--------|--------------------------|-----------|----------|-----|----------|----|-------|
|        |                          |           | no       | yes |          |    |       |
| 1      | Quantity correspondence  | Verbal    | 27       | 19  | 0.55     | 1  | .46   |
|        |                          | Nonverbal | 33       | 17  |          |    |       |
|        | Quantity comparison      | Verbal    | 38       | 8   | 7.90     | 1  | .005  |
|        |                          | Nonverbal | 28       | 22  |          |    |       |
|        | Ordering                 | Verbal    | 4        | 42  | 51.70    | 1  | <.001 |
|        |                          | Nonverbal | 41       | 9   |          |    |       |
| 2      | Non-symbolic addition    | Verbal    | 47       | 23  | 5.81     | 1  | .016  |
|        |                          | Nonverbal | 60       | 11  |          |    |       |
|        | Non-symbolic subtraction | Verbal    | 47       | 23  | 3.14     | 1  | .076  |
|        |                          | Nonverbal | 57       | 14  |          |    |       |
|        | Cross-modal addition     | Verbal    | 30       | 40  | 6.82     | 1  | .009  |
|        |                          | Nonverbal | 46       | 25  |          |    |       |
|        | Symbolic arithmetic      | Verbal    | 53       | 17  | 0.27     | 1  | .6    |
|        |                          | Nonverbal | 51       | 20  |          |    |       |

Figure 2. Percentage of repeaters by task and experimental group





**Indirectly assessed difficulties: practice repetition.** As an indirect measure of usability, we examined whether the number of participants that repeated the practice session of each task differed by experimental condition. Table 3 presents contingency tables and  $\chi^2$  tests of association. Figure 2 presents percentage of repeaters per condition and task.

The number of participants that repeated the practice session did not vary significantly between conditions in the *Quantity correspondence task*, the *Non-symbolic subtraction task* and the *Symbolic arithmetic task*. Fewer participants repeated the practice session in the nonverbal condition of the *Ordering*, *Non-symbolic addition* and *Cross-modal addition* tasks. Inversely, more participants repeated the practice session in the nonverbal condition of the quantity comparison task.

## Task Descriptives

Table 4. Task performance, descriptives and nonverbal vs. verbal comparison

| Task type                    | Cond.     | N  | M(POMP) (SD) | Range  |        | Internal consistency |          | Skewness | S-W  |       | ANOVA (K-W) |      |
|------------------------------|-----------|----|--------------|--------|--------|----------------------|----------|----------|------|-------|-------------|------|
|                              |           |    |              | Theor. | Emp.   | $\alpha$             | $\omega$ |          | W    | p     | $\chi^2$    | p    |
| <i>Sample 1</i>              |           |    |              |        |        |                      |          |          |      |       |             |      |
| Quantity correspondence (NS) | verbal    | 46 | 0.86 (0.19)  | 0-1    | 0.08-1 | 0.813                | 0.852    | -1.77    | 0.69 | <.001 | 1.47        | >.05 |
|                              | nonverbal | 50 | 0.91 (0.16)  | 0-1    | 0.17-1 |                      |          | -2.54    |      |       |             |      |
| Quantity correspondence (S)  | verbal    | 46 | 0.96 (0.09)  | 0-1    | 0.63-1 | 0.746                | 0.819    | -2.39    | 0.48 | <.001 | 1.63        | >.05 |
|                              | nonverbal | 50 | 0.92 (0.17)  | 0-1    | 0.25-1 |                      |          | -2.63    |      |       |             |      |
| Quantity comparison (NS)     | verbal    | 46 | 0.95 (0.13)  | 0-1    | 0.50-1 | 0.892                | 0.899    | -2.94    | 0.48 | <.001 | 1.62        | >.05 |
|                              | nonverbal | 50 | 0.87 (0.27)  | 0-1    | 0-1    |                      |          | -1.94    |      |       |             |      |
| Quantity comparison (S)      | verbal    | 46 | 0.96 (0.14)  | 0-1    | 0.25-1 | 0.697                | 0.737    | -3.54    | 0.48 | <.001 | 3.70        | >.05 |
|                              | nonverbal | 50 | 0.88 (0.24)  | 0-1    | 0-1    |                      |          | -2.00    |      |       |             |      |
| Ordering (NS)                | verbal    | 46 | 0.78 (0.26)  | 0-1    | 0.25-1 | 0.462                | 0.521    | -0.68    | 0.83 | <.001 | 0.60        | >.05 |
|                              | nonverbal | 50 | 0.74 (0.25)  | 0-1    | 0-1    |                      |          | -0.80    |      |       |             |      |
| Ordering (S)                 | verbal    | 46 | 0.92 (0.22)  | 0-1    | 0-1    | 0.735                | 0.763    | -2.78    | 0.53 | <.001 | 3.02        | >.05 |
|                              | nonverbal | 50 | 0.88 (0.23)  | 0-1    | 0-1    |                      |          | -2.08    |      |       |             |      |
| <i>Sample 2</i>              |           |    |              |        |        |                      |          |          |      |       |             |      |
| NS Add. (NS)                 | verbal    | 46 | 0.78 (0.24)  | 0-1    | 0-1    | 0.495                | 0.553    | -0.82    | 0.75 | <.001 | 1.94        | >.05 |
|                              | nonverbal | 50 | 0.84 (0.19)  | 0-1    | 0.33-1 |                      |          | -0.69    |      |       |             |      |
| NS Add. (S)                  | verbal    | 46 | 0.66 (0.31)  | 0-1    | 0-1    |                      |          | -0.40    | 0.84 | <.001 | 0.15        | >.05 |
|                              | nonverbal | 50 | 0.67 (0.32)  | 0-1    | 0-1    |                      |          | -0.63    |      |       |             |      |
| NS Sub. (NS)                 | verbal    | 46 | 0.88 (0.21)  | 0-1    | 0-1    | 0.593                | 0.618    | -1.88    | 0.61 | <.001 | 0.00        | >.05 |
|                              | nonverbal | 50 | 0.89 (0.19)  | 0-1    | 0.33-1 |                      |          | -1.51    |      |       |             |      |
| NS Sub (S)                   | verbal    | 46 | 0.62 (0.35)  | 0-1    | 0-1    |                      |          | -0.50    | 0.83 | <.001 | 1.27        | >.05 |
|                              | nonverbal | 50 | 0.69 (0.33)  | 0-1    | 0-1    |                      |          | -0.78    |      |       |             |      |
| Cross-modal Add. (NS)        | verbal    | 46 | 0.79 (0.26)  | 0-1    | 0-1    | 0.439                | 0.480    | -0.90    | 0.75 | <.001 | 0.03        | >.05 |
|                              | nonverbal | 50 | 0.79 (0.27)  | 0-1    | 0-1    |                      |          | -1.07    |      |       |             |      |
| Cross-modal Add. (S)         | verbal    | 46 | 0.62 (0.33)  | 0-1    | 0-1    |                      |          | -0.49    | 0.86 | <.001 | 1.37        | >.05 |
|                              | nonverbal | 50 | 0.57 (0.31)  | 0-1    | 0-1    |                      |          | -0.28    |      |       |             |      |
| Symbolic arithmetic (Add.)   | verbal    | 46 | 0.95 (0.14)  | 0-1    | 0-1    | 0.880                | 0.888    | -4.76    | 0.32 | <.001 | 0.07        | >.05 |
|                              | nonverbal | 50 | 0.94 (0.21)  | 0-1    | 0-1    |                      |          | -4.05    |      |       |             |      |
| Symbolic Arithmetic (Sub.)   | verbal    | 46 | 0.85 (0.23)  | 0-1    | 0-1    | 0.787                | 0.803    | -2.04    | 0.66 | <.001 | 0.01        | >.05 |
|                              | nonverbal | 50 | 0.83 (0.27)  | 0-1    | 0-1    |                      |          | -1.84    |      |       |             |      |

Note: POMP = Percentage of maximum performance; S-W = Shapiro-Wilk test of normality; K-W = Kruskal-Wallis ANOVA on ranks

**Internal consistency.** Internal consistency of the experimental tasks in the first sample ranged from good to questionable (see Table 4). Only the *Ordering* task with non-symbolic answers showed unacceptable internal consistency. Due to the low number of items in each task, we estimated internal consistency without differentiation as to answer format in the second sample. While the *Symbolic arithmetic* task provided acceptable (Subtraction) to good (Addition) internal consistency, the three other tasks only reached poor to questionable consistency.

**Tests for normality.** All task scores showed ceiling effects (somewhat less pronounced in Sample 2), independently from experimental group or the symbolic nature of the task, thus deviating significantly from the normal distribution (statistical tests for all subtests are reported in Table 4). Skewed distributions were expected considering the test was designed to differentiate at the bottom

end of the ability distribution. Consequently, the Shapiro-Wilks tests showed substantial non-normality. Therefore, we conducted non-parametric analysis of variance to examine possible group differences in task performance.

Table 5: Scale intercorrelations: Sample 1

| Scale intercorrelations<br>Sample 1 |          | Quantity correspondence<br>(S) | Quantity<br>comparison (NS) | Quantity<br>comparison (S) | Ordering<br>(NS) | Ordering<br>(S) |
|-------------------------------------|----------|--------------------------------|-----------------------------|----------------------------|------------------|-----------------|
| Quantity correspondence (NS)        | Rho      | 0.516                          | 0.424                       | 0.189                      | 0.436            | 0.296           |
|                                     | <i>p</i> | <.001                          | <.001                       | 0.065                      | <.001            | 0.003           |
| Quantity correspondence (S)         | Rho      |                                | 0.374                       | 0.212                      | 0.294            | 0.215           |
|                                     | <i>p</i> |                                | <.001                       | 0.038                      | 0.004            | 0.036           |
| Quantity comparison (NS)            | Rho      |                                |                             | 0.612                      | 0.443            | 0.381           |
|                                     | <i>p</i> |                                |                             | <.001                      | <.001            | <.001           |
| Quantity comparison (S)             | Rho      |                                |                             |                            | 0.125            | 0.147           |
|                                     | <i>p</i> |                                |                             |                            | 0.225            | 0.153           |
| Ordering (NS)                       | Rho      |                                |                             |                            |                  | 0.519           |
|                                     | <i>p</i> |                                |                             |                            |                  | <.001           |

Note: (S)=Symbolic answer format; (NS)=Non-symbolic answer format; Rho = Spearman's rho

**Scale intercorrelations.** In Sample 1 performances on almost all experimental tasks correlated significantly among each other (see Table 5). The exception was the *Quantity comparison* task (symbolic format), which did not correlate significantly with the *Quantity correspondence* task (non-symbolic format) and with the *Ordering* task (both formats).

Table 6: Scale intercorrelations: Sample 2

| Scale intercorrelations<br>Sample 2 |          | NS Add.<br>(S) | NS Sub.<br>(NS) | NS Sub.<br>(S) | Cross. Add.<br>(NS) | Cross. Add.<br>(S) | Sym. Arith.<br>(Add) | Sym. Arith.<br>(Sub) |
|-------------------------------------|----------|----------------|-----------------|----------------|---------------------|--------------------|----------------------|----------------------|
| NS Add. (NS)                        | Rho      | 0.257          | 0.157           | 0.317          | 0.306               | 0.162              | 0.138                | 0.046                |
|                                     | <i>p</i> | 0.002          | 0.063           | <.001          | <.001               | 0.056              | 0.102                | 0.590                |
| NS Add. (S)                         | Rho      |                | 0.335           | 0.372          | 0.244               | 0.260              | 0.236                | 0.417                |
|                                     | <i>p</i> |                | <.001           | <.001          | 0.003               | 0.002              | 0.005                | <.001                |
| NS Sub. (NS)                        | Rho      |                |                 | 0.342          | 0.197               | 0.042              | 0.241                | 0.231                |
|                                     | <i>p</i> |                |                 | <.001          | 0.019               | 0.619              | 0.004                | 0.006                |
| NS Sub. (S)                         | Rho      |                |                 |                | 0.249               | 0.290              | 0.193                | 0.352                |
|                                     | <i>p</i> |                |                 |                | 0.003               | <.001              | 0.022                | <.001                |
| Cross. Add. (NS)                    | Rho      |                |                 |                |                     | 0.301              | 0.091                | 0.165                |
|                                     | <i>p</i> |                |                 |                |                     | <.001              | 0.282                | 0.051                |
| Cross. Add. (S)                     | Rho      |                |                 |                |                     |                    | 0.207                | 0.211                |
|                                     | <i>p</i> |                |                 |                |                     |                    | 0.014                | 0.012                |
| Sym. Arith. (Add& Sub) (S)          | Rho      |                |                 |                |                     |                    |                      | 0.262                |
|                                     | <i>p</i> |                |                 |                |                     |                    |                      | 0.002                |

Note: (S)=Symbolic answer format; (NS)=Non-symbolic answer format; Rho = Spearman's rho

The reported correlations in the following paragraph are all significant (see Table 6). Letters in parentheses indicate the answer format (NS=non-symbolic; (S)=symbolic). In Sample 2,

performances in *Symbolic arithmetic* (addition and subtraction) correlated with each other and with performance in all other tasks having a symbolic response format (i.e. *Non-symbolic addition*, *Non-symbolic subtraction* and *Cross-modal addition*). Performance in *Symbolic arithmetic* did not correlate with performance in tasks requiring non-symbolic output, except for the *Non-symbolic subtraction* task. Performances in *Non-symbolic addition* and *subtraction (S)* correlated with performance on all other tasks. Performances in the two *Non-symbolic arithmetic (NS)* did not correlate with each other. Performance in *Cross-modal addition (S)* correlated with performance in all other tasks, except *Non-symbolic arithmetic* (i.e. *Non-symbolic addition* and *Non-symbolic subtraction*) with non-symbolic response formats. Performance in *Cross-modal addition (NS)* correlated with performance in all other tasks, except *Symbolic arithmetic*.

### Criterion validity

Table 7: Criterion validity

| Criterion validity             |          | TTR+  | TTR-  | Counting (oral) | Counting (written) | SYMP (one digit) | SYMP (two digit) |
|--------------------------------|----------|-------|-------|-----------------|--------------------|------------------|------------------|
| <i>Sample 1</i>                |          |       |       |                 |                    |                  |                  |
| Average test score (all tasks) | Rho      | .453  | .349  | .279            | .475               | .308             | .111             |
|                                | <i>p</i> | <.001 | <.001 | .006            | <.001              | .002             | .28              |
| <i>Sample 2</i>                |          |       |       |                 |                    |                  |                  |
| Average test score (all tasks) | Rho      | .431  | .355  | .441            | .499               | .409             | .26              |
|                                | <i>p</i> | <.001 | <.001 | <.001           | <.001              | <.001            | .002             |

Note: Rho = Spearman's rho

In Sample 1, average performance (all experimental tasks combined) correlated significantly with all criterion validity tasks (see Table 7) except with the two-digit SYMP test.

In Sample 2, average performance (all experimental tasks combined) correlated significantly with all criterion validity tasks.

### **Comparison of task performance: Verbal versus nonverbal instructions**

Analyses of variance (Kruskal-Wallis) on task scores with experimental group (verbal versus nonverbal) as between-subjects factor revealed no significant differences in any of the tasks, neither in Sample 1 nor in Sample 2 (see Table 4). Overall performances were very high in the nonverbal and in the verbal condition (ranging between 57% and 96%), indicating that children succeeded comparably well in both conditions.

## Discussion

The purpose of the present study was to explore the possibility of measuring basic math competence in young children without using verbal instructions. To this aim we developed a series of computerized tasks presented on tablet-computers either verbally, using traditional language instructions or nonverbally, using video instructions repeatedly showing successful task completion and assessed whether the instruction type influenced task performance.

### Usability aspects

To check whether this new mode of instruction was effective, we assessed the comprehensibility of the tasks both directly and indirectly. Regarding the prior, the feedback symbols (the green happy and the red sad smiley faces during the instruction and practice phase) were easily understood by most if not all participants. The same is true for the navigation symbol (the arrow to both save the answer and switch to the next item).

As an indirect assessment of task comprehension, we examined differences in the number of participants that repeated the practice session of each task. Given the low difficulty level of the tasks presented during instruction and practice, we assumed that children who did not get the practice items right in their first attempt had not understood the purpose of the task at first and therefore needed a second run. In three tasks (*Quantity correspondence (S1)*, *Non-symbolic subtraction (S2)* and *Symbolic arithmetic (S2)*), the number of repeaters did not vary significantly, suggesting that nonverbal instructions can be understood as well as verbal ones. On the other hand, we observed significantly less repeaters in three other tasks (*Non-symbolic addition (S2)*, *Ordering (S1)* and *Cross-modal addition (S2)*) when children were instructed nonverbally, implying that nonverbal instructions can be more effective than verbal ones in these situations. This tendency

was especially pronounced in the *Ordering* task. Finally, we found an inverse difference in repeaters in the *Quantity comparison* task. Significantly more participants repeated the practice session of the *Quantity comparison* task when they received nonverbal instructions. Conveying “choose the side that has more” through a video showing successful task completion repeatedly seems to have worked less well than simply giving the participants an explicit verbal instruction to do so, even though we displayed more repetitions in this task than in the other tasks. This shows that not every task instruction can be easily replaced by nonverbal videos without adding unnecessary complexity. This result stands in stark contrast with our observations concerning the *Ordering* task, which was understood much better following non-verbal instructions. Because the verbal instruction requested to order items from left to right, the extreme difference in repeaters (91% vs. 18%) could possibly be attributed to the fact that reliable left /right distinction has not been achieved by children of this age. Notwithstanding, this observation illustrates well that a single word in the instruction can lead to a complete failure to understand the task at hand and that this can be easily avoided by using nonverbal video instructions. Taken together, our results based on the repetition of practice items suggest that nonverbal instructions are an efficient alternative to the classically used verbal instructions and might in some cases even be more direct and effective. However, they do not provide a universally applicable solution, because on rare occasions they fail to convey task instructions as clearly and unequivocally as spoken language.

Anecdotally, it appeared that children were generally highly motivated to complete our tasks and many asked if they could do them again. This might be due to the video-game-like appearance of the assessment tool, which differs considerably from the paper-and-pencil material that they encounter in everyday math classes, which probably helped to promote task compliance and motivation (Lumsden, Edwards, Lawrence, Coyle, & Munafò, 2016).

## Validity aspects

Scale intercorrelations indicate that performance in the three tasks assessed in Sample 1 (i.e. *Quantity correspondence*, *Quantity comparison*, *Ordering*) largely correlated, which may reflect the fact that they rely, at least in part, on the same basic numerical competences. While performance on the non-symbolic version of the *Quantity comparison* task did correlate with performance on most other experimental tasks, performance on the symbolic version of the *Quantity comparison task* shows less consistent correlations with performance on other tasks. Most strikingly, the latter does not correlate significantly with performance on the *Ordering* task, both symbolic and non-symbolic versions. This stands in contrast with most findings in recent literature that report strong correlation between performance on tasks measuring cardinality (*Quantity comparison task*) and ordinality (*Ordering task*) (e.g. Lyons, Price, Vaessen, Blomert, & Ansari, 2014; Sasanguie, Lyons, De Smedt, & Reynvoet, 2017; Sasanguie & Vos, 2018). This might be due to reporting correlations for the whole sample without distinguishing instruction type: a large proportion of participants in the video condition of the task did not seem to correctly understand its purpose, which could explain the absence of correlation between its performance and any other task. Accordingly, the *Quantity comparison* task will need to be adapted in future studies. Sample 2 consisted of calculation tasks that were either presented in classical symbolic or more unusual non-symbolic and/or cross-modal format (i.e. *Symbolic addition and subtraction*, *Non-symbolic addition and subtraction*, *Cross-modal addition*). In this sample, performance in symbolic arithmetic correlated with performance in those tasks having a symbolic response format, but not those requiring non-symbolic answers. This points towards a special role of number symbol processing, in line with the importance of this ability for mathematics (e.g. Bugden & Ansari, 2011; Bugden, Price, McLean, & Ansari, 2012). Interestingly, and in line with the importance of



number symbols, performance in non-symbolic arithmetic tasks with symbolic output formats also correlated with all calculation tasks of Sample 2. While validating the main expectations concerning our task and their properties, conclusions concerning scale intercorrelations remain provisional at this stage, since all tasks could not be correlated with each other in the present design due to two different participant samples.

Considering the overall medium reliability of our experimental tasks, special care should be taken to include more items assessing performance in the different tasks in further developments of this project.

Finally, we observed that average performance of all experimental tasks combined correlated significantly with performance in most (Sample 1) to all (Sample 2) control tasks. The control tasks were chosen to cover the most established measures of basic math competences in young children, known to predict later differences in standardized mathematical tests and distinguish children with mathematical learning difficulties from typically developing peers. We therefore included tasks assessing children's abilities to count (Geary, Hoard, & Hamson, 1999; Goldman, Pellegrino, & Mertz, 1988; Hornung et al., 2014; Passolunghi, M & Siegel, L, 2004; Willburger, Fussenegger, Moll, Wood, & Landerl, 2008), to compare symbolic magnitudes (Brankaer et al., 2017; De Smedt et al., 2013, 2009) and to calculate (De Vos, 1992; Geary, 1993; Geary, 2010; Klein & Bisanz, 2000; Locuniak & Jordan, 2008). The non-significant correlation between performance of the tasks in the first sample with performance in the two-digit symbolic number comparison task can be attributed to participant's lack of knowledge on two-digit numbers at the time of data collection (approx. 5 weeks of schooling) (Martin, Ugen, & Fischbach, 2013; MENFP, 2011).

### **Task performance compared by experimental group**

Type of instruction prior to the test did not affect participants' performance in any of the experimental tasks. We observed high average performance in both samples and similar performances in both experimental conditions. This leads us to conclude that instruction type does not seem have an observable effect on future task performance. In other words, explicit verbal instructions can be replaced by videos showing successful task completion for children to understand the functioning and purpose of the numerical and mathematical tasks. This is an important result when put in the context of multilingual settings in particular, where the language of instruction can have considerable negative effects on task performance. Indeed, video instructions seem to work as well as traditional verbal instructions while taking language out of the equation.

At this point, we want to stress that we do not claim that mathematics and language can be assessed independently (Dowker & Nuerk, 2016). Indeed, prior research has shown that while the logic and procedures of counting are stored independently from language, the learning of even small number words relies on linguistic skills (Wagner, Kimura, Cheung, & Barner, 2015). Also, languages inverting the order of units and tens in number words negatively affect the learning of number concepts and arithmetic (Gobel et al., 2014; Imbo, Vanden Bulcke, De Brauwer, & Fias, 2014; Zuber, Pixner, Moeller, & Nuerk, 2009). Other studies have highlighted that proficiency in the language of instruction (Abedi & Lord, 2001; Hickendorff, 2013; Paetsch et al., 2016; Saalbach, Gunzenhauser, Kempert, & Karbach, 2016) and, more specifically, the mastery of mathematical language are essential predictors of mathematics performance (Purpura & Reid, 2016). It also becomes increasingly clear that test language modulates the neuronal substrate of mathematical cognition (Salillas, Barraza, & Carreiras, 2015; Salillas & Carreiras, 2014; Van Rinsveld et al.,

2017). On the other hand, we do claim that a testee's access to the assessment tools should not be limited by proficiency in a certain language. Although most existing tasks already use images to minimize linguistic load, they still rely on some form of verbal instruction or vocabulary that needs to be fully understood to solve the task correctly. We thus think that it is not sufficient to minimize language load in mathematics items, but that it would be preferential to remove linguistic demands altogether. Our results show that this can be achieved by using implicit video instructions that rely on participant's nonverbal cognitive skills.

### **Limitations and future studies**

A first limitation for the interpretation of our results are the medium internal consistency scores of many of our tasks. We aimed to explore as many tasks as possible using nonverbal instructions, while keeping total test time under 40 minutes due to children's limited attention span (Manly et al., 2001). This led to some psychometric compromises by offering only a few items per task and subscale (i.e. symbolic and non-symbolic answer format), especially for the tasks in the second sample. In the future, we will select the tasks with the highest potential of differentiating in the lower spectrum of ability and supplement them with more items.

To further differentiate experimental conditions, it would have been possible to present only word problems and exclude all animations in the verbal instruction group whenever possible. For example, instead of showing pigs moving into a barn, the animation could be replaced with a written/spoken story on pigs going into a barn before offering three possible answers. We expect that such a contrasted design would lead to more significant differences in task comprehension and would be particularly interesting to investigate differences in item functioning in relationship to the participant's language background. In order to provide a robust proof of concept for the valid

use of video instructions we decided here to adapt a more conservative approach with minimal differences between the video and verbal conditions. However, it would be interesting to use also more contrasted conditions in future studies.

Additionally, we anecdotally observed that touchscreen responsiveness seemed to be an issue with more impulsive participants. Indeed, when the touchscreen did not react to a first touch by showing a bold border around the selected image, these participants switched to another answer. We speculate that they interpreted the non-response of the tool as a wrong answer on their part and choose to try another one. This is an unfortunate but important technical limitation that will be addressed in future versions of the application, as impulsivity and attention issues are strongly correlated with mathematical abilities, especially in the target population for this test (LeFevre et al., 2013). Finally, we want to stress the difference in participant's age between the two sets of tasks presented here. In future developments of this project, homogenous groups of children from the first half of the first grade should be targeted.

## **Conclusion**

Taken together, these preliminary results show that explicit verbal instructions do not seem to be required for assessing basic math competencies when replaced by instructional videos. While variations depending on the task and the quality of experimental instructions are present, video instructions seem to constitute a valid alternative to traditional verbal instructions. In addition, the video-game-like aspect of the present assessment tool was well received, contributing positively to children's task compliance and motivation. All in all, the results of this study provide an important and encouraging proof of concept for further developments of language neutral and fair tests without verbal instructions.

## References

- Abedi, J., & Lord, C. (2001). The Language Factor in Mathematics Education. *Applied Measurement in Education, 14*(3), 219–234. [https://doi.org/10.1207/s15324818ame1403\\_2](https://doi.org/10.1207/s15324818ame1403_2)
- Aster, M. G. von, Bzufka, M. W., & Horn, R. R. (2009). *ZAREKI-K. Neuropsychologische Testbatterie für Zahlenverarbeitung und Rechnen bei Kindern: Kindergartenversion: Manual* (Harcourt).
- Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition, 86*(3), 201–221. [https://doi.org/10.1016/S0010-0277\(02\)00178-6](https://doi.org/10.1016/S0010-0277(02)00178-6)
- Braden, J. P., & Hannah, J. M. (1998). 9 - Assessment of Hearing-Impaired and Deaf Children with the WISC-III A2 - Prifitera, Aurelio. In D. H. B. T.-W.-I. C. U. and I. Saklofske (Ed.), *Practical Resources for the Mental Health Professional* (pp. 175–201). San Diego: Academic Press. <https://doi.org/https://doi.org/10.1016/B978-012564930-8/50010-3>
- Brankaer, C., Ghesquière, P., & De Smedt, B. (2017). Symbolic magnitude processing in elementary school children: A group administered paper-and-pencil measure (SYMP Test). *Behavior Research Methods, 49*(4), 1361–1373. <https://doi.org/10.3758/s13428-016-0792-3>
- Bugden, S., & Ansari, D. (2011). Individual differences in children's mathematical competence are related to the intentional but not automatic processing of Arabic numerals. *Cognition, 118*(1), 35–47. <https://doi.org/10.1016/j.cognition.2010.09.005>
- Bugden, S., Price, G. R., Mclean, D. A., & Ansari, D. (2012). Developmental Cognitive Neuroscience The role of the left intraparietal sulcus in the relationship between symbolic number processing and children's arithmetic competence. *Accident Analysis and Prevention, 2*(4), 448–457. <https://doi.org/10.1016/j.dcn.2012.04.001>
- Butterworth, B., Varma, S., & Laurillard, D. (2011). Dyscalculia: From brain to education. *Science, 332*(6033), 1049–1053. <https://doi.org/10.1126/science.1201536>
- Cattell, R. B., & Cattell, A. K. S. (1973). *Culture Fair Intelligence Tests: CFIT*. Institute for Personality & Ability Testing.
- Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research, 34*(3), 315–346. [https://doi.org/10.1207/S15327906MBR3403\\_2](https://doi.org/10.1207/S15327906MBR3403_2)
- Courtney, A. S., Hayes, F. B., Couch, K. W., & Frick, M. (1984). Administration of the WISC-R Performance Scale to Hearing-Impaired Children Using Pantomimed Instructions. *Journal of Psychoeducational Assessment, 2*(1), 1–7. <https://doi.org/10.1177/073428298400200101>
- De Smedt, B., Noël, M. P., Gilmore, C., Ansari, D., Noël, M.-P., Gilmore, C., & Ansari, D. (2013). How do symbolic and non-symbolic numerical magnitude processing skills relate to individual differences in children's mathematical skills? A review of evidence from brain and behavior. *Trends in Neuroscience and Education, 2*(2), 48–55.

<https://doi.org/10.1016/j.tine.2013.06.001>

- De Smedt, B., Verschaffel, L., & Ghesquière, P. (2009). The predictive value of numerical magnitude comparison for individual differences in mathematics achievement. *Journal of Experimental Child Psychology*, 103(4), 469–479.  
<https://doi.org/10.1016/j.jecp.2009.01.010>
- De Vos, T. (1992). *Tempo-Test-Rekenen. Handleiding.[Tempo Test Arithmetic. Manual]*. Nijmegen: Berkhout.
- Desoete, A., Ceulemans, A., Roeyers, H., & Huylebroeck, A. (2009). Subitizing or counting as possible screening variables for learning disabilities in mathematics education or learning? *Educational Research Review*, 4(1), 55–66.
- Dowker, A., & Nuerk, H.-C. (2016). Linguistic influences on mathematics. *Frontiers in Psychology*, 7, 1035.
- Feis, Y. F. (2010). *Raven's Progressive Matrices. Encyclopedia of Cross-Cultural School Psychology*. Western Psychological Services. [https://doi.org/10.1007/978-0-387-71799-9\\_344](https://doi.org/10.1007/978-0-387-71799-9_344)
- Gallagher-Mitchell, T., Simms, V., & Litchfield, D. (2017). Learning from where 'eye' remotely look or point: impact on number line estimation error in adults. *The Quarterly Journal of Experimental Psychology*, 0218(May), 1–30.  
<https://doi.org/10.1080/17470218.2017.1335335>
- geary, lunn. (1993). SPEED-OF-PROCESSING ACROSS MONOLINGUAL, WEAK BILINGUAL, AND STRONG BILINGUAL ADULTS - geary, lunn, 1993.pdf.
- Geary, D. C. (2010). Mathematical disabilities: Reflections on cognitive, neuropsychological, and genetic components. *Learning and Individual Differences*, 20(2), 130–133.  
<https://doi.org/10.1016/j.lindif.2009.10.008>
- Geary, D. C., Hoard, M. K., & Hamson, C. O. (1999). Numerical and Arithmetical Cognition: Patterns of Functions and Deficits in Children at Risk for a Mathematical Disability. *Journal of Experimental Child Psychology*, 74(3), 213–239.  
<https://doi.org/10.1006/jecp.1999.2515>
- Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities*, 38(4), 293–304.
- Gobel, S. M., Moeller, K., Pixner, S., Kaufmann, L., Nuerk, H.-C. C., Göbel, S. M., ... Nuerk, H.-C. C. (2014). Language affects symbolic arithmetic in children: the case of number word inversion. *Journal of Experimental Child Psychology*, 119(1), 17–25.  
<https://doi.org/10.1016/j.jecp.2013.10.001>
- Goldman, S. R., Pellegrino, J. W., & Mertz, D. L. (1988). Extended Practice of Basic Addition Facts: Strategy Changes in Learning-Disabled Students. *Cognition and Instruction*, 5(3), 223–265. [https://doi.org/10.1207/s1532690xci0503\\_2](https://doi.org/10.1207/s1532690xci0503_2)
- Haag, N., Heppt, B., Roppelt, A., & Stanat, P. (2014). Linguistic simplification of mathematics

- items: effects for language minority students in Germany. *European Journal of Psychology of Education*, 30(2), 145–167. <https://doi.org/10.1007/s10212-014-0233-6>
- Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction*, 28, 24–34. <https://doi.org/10.1016/j.learninstruc.2013.04.001>
- Haffner, J., Baro, K., Parzer, P., & Resch, F. (2005). Heidelberger Rechentest (HRT 1-4) [Heidelberg Calculation Test]. *Göttingen: Hogrefe*.
- Hickendorff, M. (2013). The Language Factor in Elementary Mathematics Assessments: Computational Skills and Applied Problem Solving in a Multidimensional IRT Framework. *Applied Measurement in Education*, 26(4), 253–278. <https://doi.org/10.1080/08957347.2013.824451>
- Hornung, C., Schiltz, C., Brunner, M., & Martin, R. (2014). Predicting first-grade mathematics achievement: The contributions of domain-general cognitive abilities, nonverbal number sense, and early number competence. *Frontiers in Psychology*, 5(APR), 272. <https://doi.org/10.3389/fpsyg.2014.00272>
- Imbo, I., Vanden Bulcke, C., De Brauwer, J., & Fias, W. (2014). Sixty-four or four-and-sixty? The influence of language and working memory on children's number transcoding. *Frontiers in Psychology*, 5, 313.
- Jordan, N. C., Glutting, J., & Ramineni, C. (2010). The importance of number sense to mathematics achievement in first and third grades. *Learning and Individual Differences*, 20(2), 82–88. <https://doi.org/10.1016/j.lindif.2009.07.004>
- Klein, J. S., & Bisanz, J. (2000). Preschoolers doing arithmetic: The concepts are willing but the working memory is weak. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 54(2), 105.
- LeFevre, J.-A., Berrigan, L., Vendetti, C., Kamawar, D., Bisanz, J., Skwarchuk, S.-L., & Smith-Chant, B. L. (2013). The role of executive attention in the acquisition of mathematical skills for children in Grades 2 through 4. *Journal of Experimental Child Psychology*, 114(2), 243–261.
- LeFevre, J., Fast, L., Skwarchuk, S., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. *Child Development*, 81(6), 1753–1767.
- Lenz, T. (2015). *Bildungsbericht Luxemburg*.
- Locuniak, M. N., & Jordan, N. C. (2008). Using kindergarten number sense to predict calculation fluency in second grade. *Journal of Learning Disabilities*, 41(5), 451–459.
- Lohman, D. F., & Hagen, E. P. (2001). Cognitive abilities test (Form 6). *Rolling Meadows, IL: Riverside*.
- Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafò, M. R. (2016). Gamification of cognitive assessment and cognitive training: a systematic review of applications and

- efficacy. *JMIR Serious Games*, 4(2).
- Lyons, I. M., Price, G. R., Vaessen, A., Blomert, L., & Ansari, D. (2014). Numerical predictors of arithmetic success in grades 1-6. *Developmental Science*, 17(5), 714–726. <https://doi.org/10.1111/desc.12152>
- Manly, T., Anderson, V., Nimmo-Smith, I., Turner, A., Watson, P., & Robertson, I. H. (2001). The Differential Assessment of Children's Attention: The Test of Everyday Attention for Children (TEA-Ch), Normative Sample and ADHD Performance. *Journal of Child Psychology and Psychiatry*, 42(8), 1065–1081. <https://doi.org/10.1111/1469-7610.00806>
- Martin, R., Ugen, S., & Fischbach, A. (2013). *Épreuves Standardisées - Bildungsmonitoring Luxemburg*.
- MENFP. (2011). Plan d'études: Ecole fondamentale.
- Ministère de l'éducation nationale de l'enfance et de la Jeunesse. (2015). *Luxembourgish Education System in Key Figures School year 2014/2015*. Luxembourg.
- Naglieri, J. A. (2003). Naglieri Nonverbal Ability Tests. In *Handbook of nonverbal assessment* (p. 175). Springer.
- Noël, M.-P., Grégoire, J., & Nieuwenhoven, V. (2008). *Test diagnostique des compétences de base en mathématiques* (Editions d).
- Nosworthy, N., Bugden, S., Archibald, L., Evans, B., & Ansari, D. (2013). A Two-Minute Paper-and-Pencil Test of Symbolic and Nonsymbolic Numerical Magnitude Processing Explains Variability in Primary School Children's Arithmetic Competence. *PLoS ONE*, 8(7), e67918. <https://doi.org/10.1371/journal.pone.0067918>
- Ortiz, S. O., & Dynda, A. M. (2005). Use of Intelligence Tests with Culturally and Linguistically Diverse Populations. In *Contemporary Intellectual Assessment: Theories, Tests, and Issues*. (pp. 545–556). New York, NY, US: Guilford Press.
- Paetsch, J., Radmann, S., Felbrich, A., Lehmann, R., & Stanat, P. (2016). Sprachkompetenz als Prädiktor mathematischer Kompetenzentwicklung von Kindern deutscher und nicht-deutscher Familiensprache. *Zeitschrift Fur Entwicklungspsychologie Und Padagogische Psychologie*, 48(1), 27–41. <https://doi.org/10.1026/0049-8637/a000142>
- Passolunghi, M. C., & Siegel, L. S. (2004). Working Memory and access to numerical information in children with disability in mathematics. *Journal of Experimental Child Psychology*, 88(4), 348–367.
- Pellegrini, A. D., & Bohn, C. M. (2005). The Role of Recess in Children's Cognitive Performance and School Adjustment. *Educational Researcher*, 34(1), 13–19. <https://doi.org/10.3102/0013189X034001013>
- Purpura, D. J., & Reid, E. E. (2016). Mathematics and language: Individual and group differences in mathematical language skills in young children. *Early Childhood Research Quarterly*, 36, 259–268. <https://doi.org/10.1016/j.ecresq.2015.12.020>



- Ricken, G., Fritz, A., & Balzer, L. (2011). Mathematik und Rechnen–Test zur Erfassung von Konzepten im Vorschulalter (MARKO-D)–ein Beispiel für einen niveaurorientierten Ansatz. *Empirische Sonderpädagogik*, 3(3), 256–271.
- Saalbach, H., Gunzenhauser, C., Kempert, S., & Karbach, J. (2016). Der Einfluss von Mehrsprachigkeit auf mathematische Fähigkeiten bei Grundschulkindern mit niedrigem sozioökonomischen Status. *Frühe Bildung*, 5(2), 73–81. <https://doi.org/10.1026/2191-9186/a000255>
- Salillas, E., Barraza, P., & Carreiras, M. (2015). Oscillatory brain activity reveals linguistic prints in the quantity code. *PloS One*, 10(4), e0121434.
- Salillas, E., & Carreiras, M. (2014). Core number representations are shaped by language. *Cortex*, 52, 1–11.
- Sasanguie, D., Lyons, I. M., De Smedt, B., & Reynvoet, B. (2017). Unpacking symbolic number comparison and its relation with arithmetic in adults. *Cognition*, 165, 26–38. <https://doi.org/10.1016/j.cognition.2017.04.007>
- Sasanguie, D., Van Den Bussche, E., & Reynvoet, B. (2012). Predictors for Mathematics Achievement? Evidence From a Longitudinal Study. *Minds, Brain, and Education*, 6(3), 119–128. <https://doi.org/10.1111/j.1751-228X.2012.01147.x>
- Sasanguie, D., & Vos, H. (2018). About why there is a shift from cardinal to ordinal processing in the association with arithmetic between first and second grade. *Developmental Science*, 1–48. <https://doi.org/10.1111/desc.12653>
- Scarr-Salapatek, S. (1971). Race, social class, and IQ. *Science*, Vol. 174(4016), 1285–1295. <https://doi.org/10.1126/science.174.4016.1285>
- Schaupp, H., Holzer, N., & Lenart, F. (2007). ERT 1+. Eggenberger Rechentest 1+. *Diagnostikum Für Dyskalkulie Für Das Ende Der, 1*.
- Schneider, M., Beeres, K., Coban, L., Merz, S., Susan Schmidt, S., Stricker, J., & De Smedt, B. (2017). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: a meta-analysis. *Developmental Science*, 20(3), 1–16. <https://doi.org/10.1111/desc.12372>
- van Luit, J. E. H., van de Rijt, B. A. M., & Hasemann, K. (2001). *Osnabrücker Test zur Zahlbegriffsentwicklung: OTZ*. Hogrefe, Verlag für Psychologie.
- Van Rinsveld, A., Dricot, L., Guillaume, M., Rossion, B., & Schiltz, C. (2017). Mental arithmetic in the bilingual brain: Language matters. *Neuropsychologia*, 101(May), 17–29. <https://doi.org/10.1016/j.neuropsychologia.2017.05.009>
- Van Rinsveld, A., Schiltz, C., Brunner, M., Landerl, K., Ugen, S., Rinsveld, A. Van, ... Ugen, S. (2016). Solving arithmetic problems in first and second language: Does the language context matter? *Learning and Instruction*, 42, 72–82. <https://doi.org/10.1016/j.learninstruc.2016.01.003>
- Wagner, K., Kimura, K., Cheung, P., & Barner, D. (2015). Why is number word learning hard?

Evidence from bilingual learners. *Cognitive Psychology*, 83, 1–21.  
<https://doi.org/10.1016/j.cogpsych.2015.08.006>

Willburger, E., Fussenegger, B., Moll, K., Wood, G., & Landerl, K. (2008). Naming speed in dyslexia and dyscalculia. *Learning and Individual Differences*, 18(2), 224–236.  
<https://doi.org/10.1016/j.lindif.2008.01.003>

Zuber, J., Pixner, S., Moeller, K., & Nuerk, H. C. (2009). On the language specificity of basic number processing: Transcoding in a language with inversion and its relation to working memory capacity. *Journal of Experimental Child Psychology*, 102(1), 60–77.  
<https://doi.org/10.1016/j.jecp.2008.04.003>

**Study report 3: Assessing basic math competence without language: First steps towards psychometric validation.**

*(Submitted to Psychological and Educational Measurement)*

Max Greisen, Caroline Hornung, Tanja Gabriele Baudson, Claire Muller & Christine Schiltz

**Abstract**

The present study's aim was to assess the psychometric characteristics of NUMTEST, a language-neutral assessment battery for basic math competence that replaces verbal instructions and task content with instructional videos and animations. In this study, we investigated task and item performance, scale and test reliability as well as concurrent validity. Additionally, we provide an overview of NUMTEST's classification performance when compared to a validated screener for basic math competence. Additionally, we investigate if the claim of language-neutrality can be upheld on the basis of the data we collected in this sample. In summary, the results show that the battery is of adequate difficulty, that it provides good to excellent reliability and that it is highly and significantly correlated to a standardized measure of arithmetic. Detection performance of the battery is comparable but not identical to that of existing screeners. Bayesian performance comparison between language groups provides evidence to the claim that performance on NUMTEST is independent from the home language of the test taker. The results of this study constitute a first essential and encouraging step in the validation of the NUMTEST battery, while future studies will need to assess missing aspects such as predictive validity.

## Introduction

Basic arithmetic skills are fundamental to independent adult living in most societies. Yet an estimated 5–10 % of people (see (Devine, Soltész, Nobes, Goswami, & Szűcs, 2013), for a review of prevalence studies) suffer from numerical learning difficulties. This often leads to dependency on other people and technology in many walks of life, ranging from grocery shopping to personal finance, time management, or even voting (see (Geary, 2011) for a review). Early diagnostic and intervention are therefore essential to counteract these learning difficulties and their cumulative effects.

While diverse diagnostic tools exist, all of them share the verbal nature of their instructions and task content. However, mathematical and linguistic abilities are not independent from each other, and their complex relationship has been thoroughly studied on different levels. Whereas core representations of number seem to be independent from linguistic influence (see (Gelman & Butterworth, 2005) for a review; but see Salillas and Carreiras, 2015; Salillas et al., 2015) language plays a significant role in accessing and processing numerical representations (Brybaert, Fias, & Noël, 1998; Macizo, Herrera, Román, & Martín, 2010). For instance, intransparent number-word structure has detrimental effects on math performance (Göbel, Moeller, Pixner, Kaufmann, & Nuerk, 2014; Imbo, Vanden Bulcke, De Brauwer, & Fias, 2014; Lonnemann & Yan, 2015; Pixner, Moeller, Hermanova, Nuerk, & Kaufmann, 2011; Pixner, Zuber, et al., 2011; Van Rinsveld & Schiltz, 2016; Zuber, Pixner, Moeller, & Nuerk, 2009). In a similar vein, research on bilinguals has shown that arithmetic performance is closely related to the language in which it is performed (Van Rinsveld, Brunner, Landerl, Schiltz, & Ugen, 2015; Van Rinsveld, Schiltz, Landerl, Brunner, & Ugen, 2016). These findings have been corroborated by neuro-imagery studies (Lin, Imada, & Kuhl, 2012; Van Rinsveld, Dricot, Guillaume, Rossion, & Schiltz, 2017; Venkatraman, Siong,

Chee, & Ansari, 2006), thereby underlining the notion that exact processing of numerical objects taps into language processing networks.

The interdependence of math and language has further been corroborated by correlational studies on the link between the performance during mathematics and language assessment in general (Chow & Ekholm, 2019; Chow & Jacobs, 2016; Korpipää et al., 2017; Singer & Strasser, 2017; Vukovic & Lesaux, 2013b; Zhang, 2016) and particularly in multilingual settings (Howie, 2003; Kempert, Saalbach, & Hardy, 2008; Méndez, Hammer, Lopez, & Blair, 2019; Paetsch, Felbrich, & Stanat, 2015; Paetsch, Radmann, Felbrich, Lehmann, & Stanat, 2016; Vukovic & Lesaux, 2013a).

Other studies highlight the importance of good knowledge of the language of mathematics beyond general language skills (Purpura, Napoli, & King, 2019; Purpura & Reid, 2016) and, more specifically, mathematical vocabulary (Riccomini, Smith, Hughes, & Fries, 2015). Especially performance in word problems is strongly affected by testees' mastery of mathematical vocabulary (Hickendorff, 2013; Hornburg, Schmitt, & Purpura, 2018; Peng & Lin, 2019). It is closely related to reading comprehension (Vilenius-Tuohimaa, Aunola, & Nurmi, 2008) and vocabulary knowledge (Sepeng & Madzorera, 2014), especially in multilingual settings (Kempert, Saalbach, & Hardy, 2011)

Taken together, these discoveries have significant implications for the assessment of mathematical abilities in multilingual contexts because they indicate a potential negative bias towards populations that do not master the language in which mathematics are instructed and measured (Abedi, 2002; Abedi & Lord, 2001). For example, in Luxembourg, where primary school mathematics are taught and assessed in German, results of the national school monitoring program

(Martin, Ugen, & Fischbach, 2013) consistently show that foreign language speakers perform significantly lower in mathematics than their native-language peers.

Traditionally, a standardized psychometric test's quality relies on its objectivity, reliability, validity, and sensitivity (Eid & Schmidt, 2014). Considering the research presented so far, existing batteries that are based on verbal instructions in a language not mastered by the testees fail to meet several of these criteria. First and foremost, their validity comes into question when testees fail to understand the instructions to a task. Imagine being presented with a mathematical problem in a foreign language which you do not master: Would you conclude that your failure to resolve the task is due to your lack of mathematical abilities? Of course, you would not; yet in practice, this conclusion is common in multilingual societies like Luxembourg. Second, objectivity is affected by the fact that the testing situation is not the same for all testees. When neither the test nor its administrator can convey task instructions in a language that the tested person understands, performance depends on the (variable) compatibility between the testee's and the test's language. Additionally, test administrators faced with this situation might consciously or subconsciously be inclined to adapt the instructions to each testee, which further calls the tool's objectivity into question. Consequently, the sensitivity criterion is not met either: A good test should reliably classify problematic (subnormal achievement) and unproblematic performance (i.e., in the normal range). However, if performance on verbal assessment tools can be attributed (at least partially) to failure of understanding the task at hand, gatekeeping occurs: The tested person is prevented from accessing the very content that is supposed to measure his or her ability, resulting in failure of the task for reasons unrelated to the primary aim of the assessment. This leads to false positives in the process of screening for sub-normal performance, ultimately producing invalid assessments and unhelpful intervention strategies that fail to address the real issues. Finally, as the different quality

criteria depend on each other, a test that fails to be objective can hardly be considered for reliable, sensitive, or even valid ability assessment.

Faced with this situation, two possible solutions suggest themselves. The first one would be linguistic adaptation. This would, first, require the translation (and back-translation) of existing batteries into many different languages, second, proof that the different versions are both equivalent and statistically invariant, and third, highly polyglot test administrators (see, e.g., Van de Vijver & Hambleton, 1996;). While this solution would be ideal, it has considerable limitations from an economical and practical point of view.

The second solution, which we chose to explore, is that of removing linguistic barriers to assessment altogether. Building on the adage that a picture is worth a thousand words, we developed an alternative method for assessing basic mathematical abilities based on video instructions and animated task content. Our previous study (Greisen et al., 2018) showed that the method represents a valid alternative to traditional paper-and-pencil assessment using verbal instruction, even when testing first-grade children without or with little formal tuition on arithmetic. As the focus of the previous report lay on exploring the methodology, the results showed relatively poor overall reliability of the measures, calling for systematic selection and composition of the tasks regarding design and contents. This was the reason for the present follow-up study in which we streamlined task design, aiming to pilot a first complete version of NUMTEST (Name of the project and working name of the test battery).

The selection of task design and content was driven by two considerations. First, we wanted to include common task types which research has shown to be highly predictable of future mathematics performance in children and which thus represent ideal candidate items for screening purposes. One of these measures is symbolic and, to a lesser extent, non-symbolic quantity

comparison (see (Schneider et al., 2017) for a meta-analysis), which has been used, e.g., in the Numeracy Screener (Nosworthy, Bugden, Archibald, Evans, & Ansari, 2013), a test which was also included in the present study for validation purposes. Other studies (e.g., (Lyons & Beilock, 2011; Lyons, Price, Vaessen, Blomert, & Ansari, 2014; Reynvoet & Sasanguie, 2016; Stock, Desoete, & Roeyers, 2009)) have shown that performance on numerical ordering tasks (a generalization and extension of the comparison tasks) is also predictive of mathematics performance in later grades, which led us to include an ordering task in this battery, too. However, to provide a more differentiated view on children's basic numerical competence, we wanted to go beyond the content of classical screening measures. One common shortcoming of screeners like the Numeracy Screener is that they are limited to one single sub-competence, which, while statistically predictive, has limited its usefulness for practitioners aiming to identify a child's strengths and weaknesses. In order to broaden the set of tasks, and taking into consideration Luxembourg's governmental learning goals for preschool, which include the ability to solve image-based arithmetic problems in the range of 0 to 10, we therefore also included nonverbal addition and subtraction tasks, resulting in two sets of tasks in NUMTEST. A first set (comparison and ordering) is comprised of precursor abilities, while a second set (addition and subtraction) covers applied abilities. As this test's purpose is to screen at the lowest ability range, we limited the content of each task to the range of 0 to 10.

According to the most influential model of the development of children's numerical cognition (von Aster & Shalev, 2007), children in the target population (i.e., at the beginning of first grade) should be able to represent numbers non-symbolically using concrete quantities. They should also be situated in the beginning stages of matching these non-symbolical representations to the Arabic digits that they will predominantly encounter during their primary schooling. Therefore, each task



included both non-symbolic and symbolic answering options, allowing children to answer the task's demands at the best of their ability, regardless of their level of knowledge of Arabic digits. On the following pages, we will first describe the subtasks that were used and the modifications that we implemented after two previous exploratory studies (Greisen et al., 2018). We will then present the test's quality criteria, an exploratory factor structure as well as its classification performance compared to the Numeracy Screener in order to provide evidence of NUMTEST's psychometric validity. Finally, we will examine our core claim that NUMTEST's language-independent assessment provides for greater fairness.

## Methods

### Participants

Table 1: Participant demographics.

| <i>N</i> | % girls | Age ( <i>SD</i> ) | Duration of Schooling  | Language |       | ISEI          |
|----------|---------|-------------------|------------------------|----------|-------|---------------|
|          |         |                   |                        | % NL     | % NNL | <i>M (SD)</i> |
| 158      | 51.9    | 7y3m (6m)         | 4–7 weeks into Grade 1 | 50.6     | 49.4  | 46.3 (6.41)   |

*Notes.* % NL = percentage of native language speaking children (Luxembourgish and German). % NNL = percentage of children speaking non-native languages (Portuguese, French, Italian, Spanish, Slavic languages, English and others). ISEI = International Socio-Economic Index of Occupational Status.

Table 1 shows participant gender, age, duration of schooling, first language, and socio-economic status. Participants were recruited on a voluntary basis in participating primary schools under the authorization of Luxembourg’s Ministry of Education. Teachers were contacted directly for participation by information letters. Upon acceptance of participation, they were sent further information and consent declarations to be signed by the participants’ legal guardians. Only children whose parents had provided consent were tested and included in the present sample. Participants’ age could not be inquired directly due to strict personal data privacy regulations. We instead reported their age in this study based on the average age of the population during the national school monitoring assessment, which happened to take place within two weeks of the present study.

### Materials

**Experimental Tasks.** Tasks will be described in order of administration. Each task existed in a non-symbolic or a symbolic answer version. Both were based on identical quantities and operations.

**Non-symbolic addition (NSADD).** The first task required the children to solve a non-symbolic addition problem. Participants were shown an animation of 1–5 pigs entering a stable. The stable’s door was closed and then reopened for 1–5 additional pigs to move into the stable. Finally, the door closed again. In the **non-symbolic answer** version of this task (5 items), participants were subsequently shown three pictures of an open stable with a certain amount of pigs in it. The task was to select the picture showing the correct sum of all pigs in the stable. In the **symbolic answer** version of the task (5 items), the pictures with pigs in the stable were replaced by an array of numerals ranging from 1 to 9 from which participants chose their answer. The result range included the numbers from 3 to 8 only. Going up to 10 was not possible because cramming more than 8 pigs into the small pictures made them difficult to read.

**Non-symbolic subtraction (NSSUB).** The second task was similar to the previous one, with the difference that the animation started with an open stable showing pigs, some of which left the barn. The door was closed before the participants were asked to say how many remained. The range of pigs leaving the barn was 2-3. The result range (remaining pigs) was from 1 to 6. The non-symbolic answer version (5 items) again used pictures of pigs in a stable. The symbolic answer version of the task again used an array displaying numerals from 1 to 9. (5 items)

**Cross-modal addition (CMADD).** The third task required solving a non-symbolic addition problem using visual *and* auditory stimuli. Participants watched coins dropping on a scene floor, each one accompanied by a distinctive sound. After a curtain closed, hiding the coins already dropped, more coins fell down behind the closed curtains such that participants could hear, but not see the second series of coins hitting the floor. They were then asked to select the corresponding sum of coins on the floor while considering both the coins they had seen and heard and those they had heard only. The number of coins at each step ranged from 1 to 5, with totals in the range of 3

to 7. Non-symbolic and symbolic answer versions of this task were designed identically to those of the previous tasks.

**Ordering (ORD).** The fourth task required reordering 4 pictures by ascending numerosity. Pictures showed numerosities non-symbolically (dot arrays, 5 items) or symbolically (Arab digits, 5 items) in a range from 1 to 9.

**Quantity comparison (COMP).** This task required children to choose the larger of two numerosities, ranging from 1 to 9 and displayed in the middle of the screen. The type of numerosity (non-symbolic vs. symbolic) was varied in the same way as in the other tasks but represented by 6 items for each type to symmetrically counterbalance correct answers in the left /right side of the screen.

The tasks were similar to those developed for the first “proof of concept” study of the NUMTEST (Greisen et al., 2018) but methodologically improved for the present version. First and foremost, stimulus design was streamlined and reduced in variability across all tasks. For instance, while the first version of the tasks often used different depictions of objects (fruit, people, school accessories, etc.) as non-symbolic stimuli, we limited them to depictions of arrays of more abstract black dots in this revised version to minimize task-irrelevant distraction. Moreover, we increased the number of items per task to five to increase each scale’s reliability, except for the comparison task, which required 6 items for left-right counterbalancing of correct answer positions, as mentioned above. The animations used in the first three tasks were also slowed down significantly to allow participants more time for evaluation, thereby reducing the task’s demands on rapid visual processing. Finally, the comparison task was entirely redesigned to facilitate the understanding of task instructions. In doing so, we limited the stimuli to depictions of dot arrays and Arabic digits

and used only the easiest ratios between the left and right stimulus during the instruction and practice phase

### **Criterion Validation Tasks**

After completing each of the aforementioned tasks, participants received a paper notebook containing questions on demographics (gender, language background) as well as tasks for the criterion (concurrent) validation of the NUMTEST. The following pen & paper tasks were administered in the following order:

- ***TTR (Tempo Test Rekenen)*** (De Vos, 1992): The TTR assesses mental arithmetic performance under time constraints (60 seconds) through 8 subsets of symbolic arithmetic problems with increasing difficulty for each subsequent set and operands in the range of 1 to 100. We only used the addition scale as symbolic subtraction, multiplication and division have not yet been taught at the beginning of grade 1. The tool's reliability has been established in several studies (Desoete, 2008; Ghesquière & Ruijsenaars, 1994).
- ***Quantity comparison:*** The Numeracy Screener (Nosworthy et al., 2013) is a validated and reliable (Hawes, Nosworthy, Archibald, & Ansari, 2019) screener for future mathematics performance. It requires participants to cross out the larger of two numerosities in a total of 2 x 56 numerosity pairs. The test comprises two parts, Part 1 using non-symbolic dot arrays and Part 2 using symbolic digits for rapid numerosity comparison. Participants were given 1 minute to correctly solve as many items as possible. We used the results to compare performance classification of the Numeracy Screener with that of our experimental test battery.

## **Procedure**

Participants were seated in front of touchscreen computers and given headphones to wear. Instructions were administered in the same way for each task. First, a video was displayed showing the same computer that the participants were using with an item on display. Participants then saw a hand pushing on the correct answer in the following screen, followed by a green smiley face. This was shown three times altogether with different items. The instruction video also showed that pushing a blue arrow in the corner of the screen confirms the given answer and proceeds to the next item. Participants then moved on to the first of three practice items that were similar, but not identical to the ones they had seen in the video. In this phase, participants received the same smiley-face feedback as seen in the instruction video if their answer was correct and a red frowning face if the answer was wrong. Since the instruction video showed correct answers only in order not to confuse participants, the red frowning face could be discovered upon error during the practice phase only. If the participants had solved three practice items correctly, the application proceeded to the actual testing part. In case of one or more incorrect answers, a second run including all practice items was offered to the participants. After this, the application started the test phase, regardless of participant's performance in the second practice run. To make the transition clearer, a traffic light switching from red to green was shown to inform participants that the test session was about to start. During the test phase, participants received no feedback on the correctness of their answers.

After completion of each task, an image was displayed thanking the children for their participation. Finally, the children completed the notebooks comprising the demographics questions and the remaining tasks (TTR, Numeracy Screener) in a group setting with a test administrator.

## **Results**

Results will be presented according to the criteria of classical test theory: objectivity, reliability, and validity. Furthermore, fairness was assessed to examine the impact of children's first language.

### **Objectivity**

NUMTEST's objectivity is ascertained by its automated and standardized computerized administration without any verbal instructions. Except for occasional encouragement, the test administrators did not interact with participants during task completion.

### **Reliability**

Table 2 shows the average scores on the experimental tasks, expressed as POMP scores (percentage of maximum performance, i.e. the number of correct answers divided by the total number of answers)(Cohen, Cohen, Aiken, & West, 1999) and Cronbach's alphas for the subscales and the full scale. As they were constrained in time, only total performance was considered for validation tasks and is reported as raw value sums. Detailed item statistics are given in Appendix. All item and scale statistics were calculated using Jamovi (Love, Dropmann, & Selker, 2019).

**Table 2. Scale reliability and performance**

|   | $\alpha$ | Average performance<br>( <i>SD</i> ) | Empirical range |      | Theoretical range | Skew  | Kurtosis |
|---|----------|--------------------------------------|-----------------|------|-------------------|-------|----------|
| <b>Experimental tasks: Non-symbolic answers</b> |          |                                      |                 |      |                   |       |          |
| NSADD   | .69      | 66% (32%)                            | 0-100%          |      | 0–100%            | -.60  | -.78     |
| NSSUB   | .68      | 67% (31%)                            | 0-100%          |      | 0–100%            | -.60  | -.75     |
| CMADD   | .69      | 70% (30%)                            | 0-100%          |      | 0–100%            | -.77  | -.49     |
| ORD   | .81      | 77% (30%)                            | 0-100%          |      | 0–100%            | -1.40 | .96      |
| COMP  | .84      | 94% (18%)                            | 0-100%          |      | 0–100%            | -3.23 | 10.27    |
| NUMTEST (NS)                                    | .86      | 75% (19%)                            | 16-100%         |      | 0–100%            | -.68  | -.21     |
| <b>Experimental tasks: Symbolic answers</b>     |          |                                      |                 |      |                   |       |          |
| NSADD   | .82      | 50% (38%)                            | 0%              | 100% | 0–100%            | -.05  | -1.54    |
| NSSUB   | .79      | 52% (37%)                            | 0%              | 100% | 0–100%            | .02   | -1.44    |
| CMADD   | .71      | 48% (34%)                            | 0%              | 100% | 0–100%            | .07   | -1.19    |
| ORD   | .81      | 87% (25%)                            | 0%              | 100% | 0–100%            | -2.22 | 4.35     |
| COMP  | .82      | 93% (19%)                            | 0%              | 100% | 0–100%            | -2.95 | 8.48     |
| NUMTEST (S)                                     | .88      | 66% (22%)                            | 17%             | 100% | 0–100%            | -.34  | -.88     |
| <b>Validation tasks</b>                         |          |                                      |                 |      |                   |       |          |
| TTR Addition                                    |          | 5.06 (2.83)                          | 0               | 12   | 0–40              | 0.21  | -0.48    |
| Numeracy Screener                               |          | 29.43 (7.3)                          | 5.5             | 44   | 0–56              | -0.86 | 0.58     |

**Note.** NS = Non-symbolic answer; S= Symbolic answer

## Validity

### Face validity

Face validity was established by expert discussion between the authors and colleagues working on the standardized mathematics assessment for Luxembourg’s school monitoring program (Martin et al., 2013) and confirmed through feedback from participating teachers.

### Convergent validity

Convergent validity was shown by significant correlations between the NUMTEST total score and two validated measures of basic number competence (TTR (Additions):  $r=.56, p<.05$ , Numeracy Screener:  $r=.46, p<.05$ ). Table 3 shows the correlations between NUMTEST subtask and average



scores (for the non-symbolic and the symbolic version, respectively) and the TTR as well as with the Numeracy Screener.

**Table 3. Sub-scale correlations with validation tasks**

| <b>Non-symbolic answers</b> | <b>NSADD</b> | <b>NSSUB</b> | <b>CMADD</b> | <b>ORD</b> | <b>COMP</b> | <b>NUMTEST (NS)</b> |
|-----------------------------|--------------|--------------|--------------|------------|-------------|---------------------|
| TTR Addition                | .33*         | .37*         | .33*         | .42*       | .34*        | .53*                |
| Numeracy Screener           | .20*         | .35*         | .23*         | .33*       | .31*        | .41*                |
| <b>Symbolic answers</b>     | <b>NSADD</b> | <b>NSSUB</b> | <b>CMADD</b> | <b>ORD</b> | <b>COMP</b> | <b>NUMTEST (S)</b>  |
| TTR Addition                | .41*         | .41*         | .41*         | .35*       | .33*        | .55*                |
| Numeracy Screener           | .35*         | .31*         | .28*         | .40*       | .34*        | .46*                |

**Note.** \*  $p < .05$ ; NS = Non-symbolic answer; S= Symbolic answer

### **Factorial Validity**

In order to explore NUMTEST’s factor structure we performed exploratory factor analysis. Two tasks (ordering and comparison) showed strong ceiling effects (see table 2), performance on the three other tasks approximated normal distribution. In order to avoid the resulting bias in exploratory factor analysis, we subjected the performance data of each task to a  $\log_{10}$  transformation. Because some participants scored 0 on some tasks and  $\log_{10}(0)$  is not defined, the following formula was used:  $\text{score\_log} = \log_{10}(1 + \text{brute\_score})$ . These transformed scores were used for exploratory factor analysis. The resulting model is significantly better than the null model ( $\chi^2 = 20.84$ ,  $df = 11$ ,  $p < .05$ ) and proposes four factors that line up perfectly with our tasks. Factor 1 (arithmetic: addition) underlies both non-symbolic addition tasks (NSADD and CMADD), factor 2 (quantity comparison) underlies the comparison task (COMP), factor 3 (quantity ordering) underlies the ordering task (ORD) and finally factor 4 (arithmetic: subtraction) underlies the non-symbolic subtraction task (NSSUB). Table 4.1 shows factor loadings while table 4.2 shows factor correlations. Model goodness-of-fit tends towards good (RMSEA = .079, TLI = .935) but closely

misses the criteria. Although the sample was technically too small for a robust factor analysis, we nevertheless decided to include these results at this stage of test development to obtain an indicator of what to expect in the next test development phases.

**Table 4.1: Factor Loadings**

|            | <b>Factor 1</b> | <b>Factor 2</b> | <b>Factor 3</b> | <b>Factor 4</b> | <b>Uniqueness</b> |
|------------|-----------------|-----------------|-----------------|-----------------|-------------------|
| NSADD (NS) | 0.876           | .               | .               | .               | 0.386             |
| NSADD (S)  | 0.891           | .               | .               | .               | 0.160             |
| CMADD (NS) | 0.540           | .               | .               | .               | 0.613             |
| CMADD (S)  | 0.583           | .               | .               | .               | 0.534             |
| COMP (NS)  | .               | 0.787           | .               | .               | 0.304             |
| COMP (S)   | .               | 0.973           | .               | .               | 0.142             |
| ORD (NS)   | .               | .               | .               | 0.674           | 0.438             |
| ORD (S)    | .               | .               | .               | 0.876           | 0.319             |
| NSSUB (NS) | .               | .               | 0.898           | .               | 0.276             |
| NSSUB (S)  | .               | .               | 0.656           | .               | 0.412             |

*Note.* Applied rotation method is promax. The number of factors was determined through parallel analysis. NS= non-symbolic answers; S= symbolic answers.

**Table 4.2: Factor Correlations**

|          | Factor 1 | Factor 2 | Factor 3 | Factor 4 |
|----------|----------|----------|----------|----------|
| Factor 1 | 1.000    |          |          |          |
| Factor 2 | 0.388    | 1.000    |          |          |
| Factor 3 | 0.715    | 0.320    | 1.000    |          |
| Factor 4 | 0.554    | 0.509    | 0.588    | 1.000    |

### **Comparison of performance classification**

We compared NUMTEST’s performance classification with the performance rankings of the Numeracy Screener to shed light on its screening capabilities at this stage of development. We choose to look at the 7% and 25% lowest performers according to the performance classification proposed by David Geary (Geary, 1993, 2010)

13 participants were identified by the Numeracy Screener performance as the 7% lowest performers. As this score is averaged based on both non-symbolic and symbolic numerical processing, we used the NUMTEST average performance with no distinction between the non-symbolic and symbolic answer performance. From these 13 participants, 4 ranked within the 7<sup>th</sup> percentile on NUMTEST scores. 6 more ranked within in the 25<sup>th</sup> percentile. The 3 remaining ones ranked over the 25<sup>th</sup> percentile on NUMTEST scores and even had good to excellent performance (68, 94 and 96% correct answers respectively).

### Fairness: Language-independence of NUMTEST Performance

The major objective of NUMTEST is to provide language-independent results, due to the language-free video instruction. In order to examine this claim, we compared two groups: first, those speaking a native language at home (Luxembourgish and/or German, native group;  $n = 79$ ), which is also the language of instruction of mathematics of our sample, and second, those that speak a foreign language (all others, non-native group;  $n = 77$ ). As frequentist inference cannot conclude on the absence of a difference, we performed a Bayesian independent samples  $t$ -test using JASP (JASP Team, 2018) on total NUMTEST performance to compare the aforementioned groups. Tables 5.1 and 5.2 show the results.

**Table 5.1: Descriptives**

|                           | Group      | $n$ | $M$   | $SD$  | $SE$  | 95% Confidence Interval |       |
|---------------------------|------------|-----|-------|-------|-------|-------------------------|-------|
|                           |            |     |       |       |       | Lower                   | Upper |
| NUMTEST total performance | Native     | 79  | 0.700 | 0.188 | 0.021 | 0.658                   | 0.742 |
|                           | Non-native | 77  | 0.714 | 0.202 | 0.023 | 0.668                   | 0.760 |

**Table 5.2: Bayesian Independent Samples  $t$ -Test**

|                           | $BF_{01}$ | error %   |
|---------------------------|-----------|-----------|
| NUMTEST total performance | 5.316     | 7.478e -6 |

A Bayes factor (BF01) of 5.32 indicates that it is more than five times more likely that our data supports the null hypotheses (no difference) than it is to support the alternative hypothesis (performance difference between language groups). We thus have moderate evidence that total performance on NUMTEST is independent from the language of the assessed.

## **Discussion**

The aim of the present study was to establish the first psychometric validation step for the tasks and the total scale of the computer based NUMTEST screening of early mathematical abilities and to provide evidence for its fairness for young students from different linguistic backgrounds.

### **Summary of the findings**

#### **Performance and reliability**

Overall performance on the NUMTEST subscales was high (Non-symbolic addition, Crossmodal addition, Non-symbolic subtraction) to very high (Ordering and Comparison), which is a positive sign for a screening at the lower end of the ability spectrum. For screening purposes, some of the more difficult items could even be removed in future versions of the tasks. However, care must be taken to not diminish the test's reliability in the process, which was overall excellent for the full scale and ranged from good to excellent when looking at the subtask-based scales. On one hand, performance on the non-symbolic versions was generally higher than on the symbolic versions, as is to be expected considering the developmental stage of our test population. On the other hand, the non-symbolic tasks seem to be of lesser (though comparable) reliability. We included non-symbolic versions of each task in the battery because the children in our sample were in an intermediate stage of their numerical development (von Aster & Shalev, 2007). The higher performance on non-symbolic versions suggests that some children were not yet competent in using Arabic digits to convey their representation of number and thus performed better when

offered concrete, task-dependent answer possibilities. This observation was corroborated by the high standard deviation in TTR performance, which is a symbolic arithmetic task: While some children already have adequate knowledge on number symbols at the beginning of the first grade, some are still acquiring it. According to the literature, performance on symbolic tasks of basic math competence is more predictive for future performance on mathematics (Schneider et al., 2017; Schwenk et al., 2017) . Nevertheless, in the context of early screening for math learning difficulties, we think that it is necessary to allow participants to express their response at the best of their abilities. Additionally, a better performance on the non-symbolic than the symbolic answer version of the same scale could be interpreted in terms of a child being in a certain developmental stage. However, low performance on both versions of the same scale could be interpreted as indicative of developmental delay. In conclusion, NUMTEST will comprise both non-symbolic and symbolic answer formats in the future.

Even though performance on the three arithmetic tasks (NSADD, CMADD and NSSUB) was average to good, performance on the ordering and comparison (ORD and COMP) task was showing ceiling effects. The design of both tasks was constrained by the limited range of numbers that could be expected to be known by participants of that age. One way of increase the variability in performance on these tasks would be to add a time constraint, similarly to the approach taken in the Numeracy Screener. So far, it was not possible to use time-constraints or measure reaction times in the current implementation of the tasks due to technical limitations of the framework that was used but will be implemented in future versions. This will considerably improve the predictive validity of the battery. Research has indeed shown that not only performance, but also solving speed is a robust predictor of later performance in mathematics (Schwenk et al., 2017).

The crossmodal addition task (CMADD) was included as an experimental alternative to the other non-symbolic addition task (NSADD). The objective was to design a task that took advantage of the tablet-computer used to present the task with both visual and auditory stimuli as operands of an addition in order to evoke a truly crossmodal and consequently abstract addition computation process. However, performance on both addition tasks was very similar. Additionally, item design for the CMADD seems heavily constrained by the limits of children's phonological working memory span. Indeed, average performance on the item  $3 + 4$  was the lowest of all items in both versions of the task, suggesting that presenting 4 consecutive sounds was already too demanding for many children of that age. Moreover, there was a 22% performance difference when comparing non-symbolic and symbolic answer versions of the CMADD task, the largest of all observed differences, suggesting that although item content was identical, the two versions of the task could be measuring different things. Finally, the CMADD tasks presents a much lower factor loading for the addition factor than the NSADD tasks, again suggesting that performance on this task is determined by more than arithmetic competencies, at least when compared to the NSADD tasks. In conclusion, we consider this task, while interesting in nature, to be ultimately redundant and too complicated for screening at the lower range of the performance spectrum. It will thus be discarded in future developments of the NUMTEST battery.

### **Validity**

We examined different facets of scale validity in this study and the results were overall very satisfying. Convergent validity with the TTR and the Numeracy Screener was very good as both versions of the scale correlate highly and significantly. Predictive validity and test sensitivity could not be measured in the current study but are essential for complete psychometric validation of

NUMTEST as a screening tool. This will be tackled in a future and final validation study using a larger sample.

### **Factor structure**

According to the results of the exploratory factor analysis, performance on NUMTESTs total scale is based on four underlying factors, which we labelled according to the tasks that presented high factor loadings. As was noted before, existing screeners are limited to certain sub competencies that, while statistically predictive of future math performance, have limited usefulness to practitioners aiming to assess a child's strengths and weaknesses. While these results are limited by the relatively small sample size for factor-analysis, they provide preliminary evidence that NUMTEST reliably measures four different subskills of basic numerical competence.

### **Performance classification**

We compared performance classification on NUMTEST with rankings on the Numeracy Screener and found that NUMTEST identified less participants as performing poorly than the Numeracy Screener. Surprisingly, three participants identified by the Numeracy Screener as performing very poorly had good to excellent performance on NUMTEST. We attribute these differences to the fact that the Numeracy Screener was completed at the very end of the testing session, which lasted around 1 hour. This, in combination with the less attractive paper and pencil format used might have contributed to underestimating the participants' real performance, which could explain why some participants that were identified as ranking in the 7<sup>th</sup> percentile on the Numeracy Screener performed significantly better in our experimental tasks. In the next study, order of administration will be counterbalanced in order to address this issue and we are confident that performance classifications will then be more aligned.

## **Language independence**

Finally, the major objective of this project was to avoid performance differences due to incompatibilities between the language of test instruction and the language of the testee by removing verbal instructions and task content. According to our results, this claim can be safely upheld at this point of NUMTEST's development, since no performance differences between native and non-native language groups could be observed using Bayesian analyses. Literature indicates that language-based performance differences are mostly seen in word-problem type tasks. By presenting problems without words and including visual information only, the tasks we designed indeed thus allowed to measure basic math competence without language interference. This finding has considerable implications, as it lays the foundation for universally valid assessment methods of which performance is largely independent of its linguistic context. It could also contribute to the establishment of a standardized criterion for math learning disability, which at this point is dependent on the country and assessment tool that is used (see (Devine et al., 2013) for a discussion of the issue), hampering comparability between experimental studies and performance on existing assessment tools.

## **Conclusion**

NUMTEST's development was driven by two main goals. First, we wanted to provide a measuring tool for basic numerical skills at the lower end of the performance spectrum that provides reliable results independently from the first language of the assessed. Although there is room for improvement, our findings suggest that this target has been met: NUMTEST is reliable and its performance doesn't seem to depend on the first language of the test-taker. Moreover, the automated computer-based assessment allows for a highly standardized test-situation in which the language spoken by the test administrator is equally irrelevant. Next, we wanted to design a



screening tool that includes measures beyond the ones used by available screeners to provide a more diversified image of the assessed competencies for practitioners. Our findings concerning NUMTEST's factor structure suggest that this target has been met as well.

Considering the importance of early identification of math problems and the problem that linguistic abilities affect the validity of available test's results, the evidence for NUMTEST's overall quality in general and its language neutrality are promising for future refinement of the test and its application in practice.

## Appendix

### Item statistics

| <b>Non-symbolic Addition (non-symbolic answers)</b> |                      |   |  |                                  |
|---|----------------------|---|--|----------------------------------|
| <b>Item</b>   | <b><i>M (SD)</i></b> | <b>Corrected item-total correlation</b> | <b><math>\alpha</math> if item dropped</b> | <b>Scale <math>\alpha</math></b> |
| 2+3   | .67 (.47)            | .53                                     | .6   | .69                              |
| 4+4   | .68 (.47)            | .50                                     | .62  | .69                              |
| 3+3   | .72 (.45)            | .46                                     | .64  | .69                              |
| 4+2   | .58 (.5)             | .36                                     | .68  | .69                              |
| 1+4   | .65 (.48)            | .39                                     | .66  | .69                              |
| <b>Non-symbolic Addition (symbolic answers)</b>     |                      |   |  |                                  |
| <b>Item</b>   | <b><i>M (SD)</i></b> | <b>Corrected item-total correlation</b> | <b><math>\alpha</math> if item dropped</b> | <b>Scale <math>\alpha</math></b> |
| 2+3   | .5 (.5)              | .64                                     | .78  | .82                              |
| 4+4   | .44 (.5)             | .60                                     | .79  | .82                              |
| 3+3   | .56 (.5)             | .62                                     | .79  | .82                              |
| 4+2   | .46 (.5)             | .55                                     | .8   | .82                              |
| 1+4   | .54 (.5)             | .66                                     | .77  | .82                              |

| <b>Non-symbolic Subtraction (non-symbolic answers)</b> |                      |   |  |                                  |
|--|----------------------|---|--|----------------------------------|
| <b>Item</b>  | <b><i>M (SD)</i></b> | <b>Corrected item-total correlation</b> | <b><math>\alpha</math> if item dropped</b> | <b>Scale <math>\alpha</math></b> |
| 4-2  | .81 (.39)            | .38                                     | .66  | .68                              |
| 5-2  | .65 (.48)            | .47                                     | .62  | .68                              |
| 3-1  | .71 (.46)            | .45                                     | .63  | .68                              |
| 4-3  | .68 (.47)            | .47                                     | .62  | .68                              |
| 5-3  | .5 (.5)              | .43                                     | .64  | .68                              |
| <b>Non-symbolic Subtraction (symbolic answers)</b>     |                      |   |  |                                  |
| <b>Item</b>  | <b><i>M (SD)</i></b> | <b>Corrected item-total correlation</b> | <b><math>\alpha</math> if item dropped</b> | <b>Scale <math>\alpha</math></b> |
| 4-2  | .64 (.48)            | .50                                     | .77  | .79                              |
| 5-2  | .5 (.5)              | .68                                     | .71  | .79                              |
| 3-1  | .54 (.5)             | .50                                     | .77  | .79                              |
| 4-3  | .46 (.5)             | .54                                     | .76  | .79                              |
| 5-3  | .45 (.5)             | .63                                     | .73  | .79                              |

| <b>Crossmodal addition (non-symbolic answers)</b> |                      |   |  |                                  |
|---|----------------------|---|--|----------------------------------|
| <b>Item</b>                                       | <b><i>M (SD)</i></b> | <b>Corrected item-total correlation</b> | <b><math>\alpha</math> if item dropped</b> | <b>Scale <math>\alpha</math></b> |
| 3+2   | .7 (.46)             | .31                                     | .69  | .69                              |
| 2+3   | .7 (.46)             | .53                                     | .59  | .69                              |
| 3+3   | .67 (.47)            | .48                                     | .62  | .69                              |
| 4+3   | .82 (.38)            | .50                                     | .62  | .69                              |
| 3+4   | .63 (.48)            | .41                                     | .65  | .69                              |
| <b>Crossmodal addition (symbolic answers)</b>     |                      |   |  |                                  |
| <b>Item</b>                                       | <b><i>M (SD)</i></b> | <b>Corrected item-total correlation</b> | <b><math>\alpha</math> if item dropped</b> | <b>Scale <math>\alpha</math></b> |
| 3+2   | .53 (.5)             | .48                                     | .65  | .71                              |
| 2+3   | .48 (.5)             | .48                                     | .65  | .71                              |
| 3+3   | .54 (.5)             | .50                                     | .64  | .71                              |
| 4+3   | .47 (.5)             | .50                                     | .64  | .71                              |
| 3+4   | .37 (.48)            | .35                                     | .7   | .71                              |

| <b>Ordering (non-symbolic answers)</b> |                      |   |  |                                  |
|--|----------------------|---|--|----------------------------------|
| <b>Item</b>                            | <b><i>M (SD)</i></b> | <b>Corrected item-total correlation</b> | <b><math>\alpha</math> if item dropped</b> | <b>Scale <math>\alpha</math></b> |
| 3 2 5 4                                | .83 (.38)            | .57                                     | .76  | .80                              |
| 6 4 3 5                                | .85 (.35)            | .67                                     | .73  | .80                              |
| 5 6 4 7                                | .80 (.4)             | .58                                     | .76  | .80                              |
| 5 7 6 8                                | .79 (.4)             | .69                                     | .72  | .80                              |
| 6 9 8 7                                | .59 (.49)            | .43                                     | .82  | .80                              |
| <b>Ordering (symbolic answers)</b>     |                      |   |  |                                  |
| <b>Item</b>                            | <b><i>M (SD)</i></b> | <b>Corrected item-total correlation</b> | <b><math>\alpha</math> if item dropped</b> | <b>Scale <math>\alpha</math></b> |
| 3 2 5 4                                | .82 (.39)            | .34                                     | .84  | .80                              |
| 6 4 3 5                                | .88 (.35)            | .65                                     | .74  | .80                              |
| 5 6 4 7                                | .86 (.35)            | .60                                     | .75  | .80                              |
| 5 7 6 8                                | .89 (.31)            | .67                                     | .73  | .80                              |
| 6 9 8 7                                | .89 (.31)            | .70                                     | .72  | .80                              |

Note. Item denominations refer to the starting configuration.

| <b>Comparison (non-symbolic answers)</b> |                      |   |  |                                  |
|--|----------------------|---|--|----------------------------------|
| <b>Item</b>                              | <b><i>M (SD)</i></b> | <b>Corrected item-total correlation</b> | <b><math>\alpha</math> if item dropped</b> | <b>Scale <math>\alpha</math></b> |
| 4 3                                      | .95 (.22)            | .51                                     | .83  | .84                              |
| 6 5                                      | .95 (.22)            | .55                                     | .82  | .84                              |
| 4 6                                      | .93 (.27)            | .72                                     | .79  | .84                              |
| 3 5                                      | .93 (.26)            | .49                                     | .84  | .84                              |
| 7 2                                      | .93 (.26)            | .75                                     | .78  | .84                              |
| 3 8                                      | .96 (.21)            | .67                                     | .80  | .84                              |
| <b>Comparison (symbolic answers)</b>     |                      |   |  |                                  |
| <b>Item</b>                              | <b><i>M (SD)</i></b> | <b>Corrected item-total correlation</b> | <b><math>\alpha</math> if item dropped</b> | <b>Scale <math>\alpha</math></b> |
| 4 3                                      | .95 (.22)            | .49                                     | .80  | .82                              |
| 6 5                                      | .92 (.28)            | .39                                     | .83  | .82                              |
| 4 6                                      | .93 (.26)            | .52                                     | .80  | .82                              |
| 3 5                                      | .91 (.29)            | .67                                     | .77  | .82                              |
| 7 2                                      | .94 (.24)            | .68                                     | .77  | .82                              |
| 3 8                                      | .93 (.26)            | .77                                     | .75  | .82                              |

*Note.* Item denominations refer to displayed number / quantity pairs.

## References

- Abedi, J. (2002). Standardized Achievement Tests and English Language Learners: Psychometrics Issues. *Educational Assessment*, 8(3), 231–257. [https://doi.org/10.1207/S15326977EA0803\\_02](https://doi.org/10.1207/S15326977EA0803_02)
- Abedi, J., & Lord, C. (2001). The Language Factor in Mathematics Tests. *Applied Measurement in Education*, 14(3), 219–234. [https://doi.org/10.1207/S15324818AME1403\\_2](https://doi.org/10.1207/S15324818AME1403_2)
- Brysbaert, M., Fias, W., & Noël, M.-P. (1998). The Whorfian hypothesis and numerical cognition: Is 'twenty-four' processed in the same way as 'four-and-twenty'? *Cognition*, 66(1), 51–77. [https://doi.org/10.1016/S0010-0277\(98\)00006-7](https://doi.org/10.1016/S0010-0277(98)00006-7)
- Chow, J. C., & Ekholm, E. (2019). Language domains differentially predict mathematics performance in young children. *Early Childhood Research Quarterly*, 46, 179–186. <https://doi.org/10.1016/j.ecresq.2018.02.011>
- Chow, J. C., & Jacobs, M. (2016). The role of language in fraction performance: A synthesis of literature. *Learning and Individual Differences*, 47, 252–257. <https://doi.org/10.1016/j.lindif.2015.12.017>
- Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The Problem of Units and the Circumstance for POMP. *Multivariate Behavioral Research*, 34(3), 315–346. [https://doi.org/10.1207/S15327906MBR3403\\_2](https://doi.org/10.1207/S15327906MBR3403_2)
- De Vos, T. (1992). Tempo-Test-Rekenen. *Handleiding.[Tempo Test Arithmetic. Manual]. Nijmegen: Berkhout.*
- Desoete, A. (2008). Multi-method assessment of metacognitive skills in elementary school children: How you test is what you get. *Metacognition and Learning*, 3(3), 189. <https://doi.org/10.1007/s11409-008-9026-0>
- Devine, A., Soltész, F., Nobes, A., Goswami, U., & Szűcs, D. (2013). Gender differences in developmental dyscalculia depend on diagnostic criteria. *Learning and Instruction*, 27, 31–39. <https://doi.org/10.1016/j.learninstruc.2013.02.004>
- Eid, M., & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Hogrefe Verlag.
- Geary, D. C. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin*, 114(2), 345–362. <https://doi.org/10.1037/0033-2909.114.2.345>
- Geary, D. C. (2010). Mathematical disabilities: Reflections on cognitive, neuropsychological, and genetic components. *Learning and Individual Differences*, 20(2), 130–133. <https://doi.org/10.1016/j.lindif.2009.10.008>
- Geary, D. C. (2011). Consequences, Characteristics, and Causes of Mathematical Learning Disabilities and Persistent Low Achievement in Mathematics: *Journal of Developmental & Behavioral Pediatrics*, 32(3), 250–263. <https://doi.org/10.1097/DBP.0b013e318209edef>
- Gelman, R., & Butterworth, B. (2005). Number and language: How are they related? *Trends in Cognitive Sciences*, 9(1), 6–10. <https://doi.org/10.1016/j.tics.2004.11.004>
- Ghesquière, P., & Ruijsenaars, A. (1994). Vlaamse normen voor studietoetsen Rekenen en technisch lezen lager onderwijs [Flemish norms for school tests on mathematics and technical reading in elementary school]. *Leuven, Belgium: KUL-CSBO.*
- Göbel, S. M., Moeller, K., Pixner, S., Kaufmann, L., & Nuerk, H.-C. (2014). Language affects symbolic arithmetic in children: The case of number word inversion. *Journal of Experimental Child Psychology*, 119, 17–25. <https://doi.org/10.1016/j.jecp.2013.10.001>

- Greisen, M., Hornung, C., Baudson, T. G., Muller, C., Martin, R., & Schiltz, C. (2018). Taking Language out of the Equation: The Assessment of Basic Math Competence Without Language. *Frontiers in Psychology, 9*, 1076. <https://doi.org/10.3389/fpsyg.2018.01076>
- Hawes, Z., Nosworthy, N., Archibald, L., & Ansari, D. (2019). Kindergarten children's symbolic number comparison skills relates to 1st grade mathematics achievement: Evidence from a two-minute paper-and-pencil test. *Learning and Instruction, 59*, 21–33. <https://doi.org/10.1016/j.learninstruc.2018.09.004>
- Hickendorff, M. (2013). The Language Factor in Elementary Mathematics Assessments: Computational Skills and Applied Problem Solving in a Multidimensional IRT Framework. *Applied Measurement in Education, 26*(4), 253–278. <https://doi.org/10.1080/08957347.2013.824451>
- Hornburg, C. B., Schmitt, S. A., & Purpura, D. J. (2018). Relations between preschoolers' mathematical language understanding and specific numeracy skills. *Journal of Experimental Child Psychology, 176*, 84–100. <https://doi.org/10.1016/j.jecp.2018.07.005>
- Howie, S. J. (2003). Language and other background factors affecting secondary pupils' performance in Mathematics in South Africa. *African Journal of Research in Mathematics, Science and Technology Education, 7*(1), 1–20. <https://doi.org/10.1080/10288457.2003.10740545>
- Imbo, I., Vanden Bulcke, C., De Brauwer, J., & Fias, W. (2014). Sixty-four or four-and-sixty? The influence of language and working memory on children's number transcoding. *Frontiers in Psychology, 5*, 313.
- JASP Team. (2018). *JASP (Version 0.9)[Computer software]*. Retrieved from <https://jasp-stats.org/>
- Kempert, S., Saalbach, H., & Hardy, I. (2008). Der Zusammenhang zwischen mathematischer Kompetenz und Zweisprachigkeit bei türkischdeutschen Grundschulkindern. In J. Ramseger & M. Wagener (Eds.), *Chancenungleichheit in der Grundschule: Ursachen und Wege aus der Krise* (pp. 219–222). [https://doi.org/10.1007/978-3-531-91108-3\\_41](https://doi.org/10.1007/978-3-531-91108-3_41)
- Kempert, S., Saalbach, H., & Hardy, I. (2011). Cognitive benefits and costs of bilingualism in elementary school students: The case of mathematical word problems. *Journal of Educational Psychology, 103*(3), 547–561. <https://doi.org/10.1037/a0023619>
- Korpiää, H., Koponen, T., Aro, M., Tolvanen, A., Aunola, K., Poikkeus, A.-M., ... Nurmi, J.-E. (2017). Covariation between reading and arithmetic skills from Grade 1 to Grade 7. *Contemporary Educational Psychology, 51*, 131–140. <https://doi.org/10.1016/j.cedpsych.2017.06.005>
- Lin, J.-F. L., Imada, T., & Kuhl, P. K. (2012). Mental Addition in Bilinguals: An fMRI Study of Task-Related and Performance-Related Activation. *Cerebral Cortex, 22*(8), 1851–1861. <https://doi.org/10.1093/cercor/bhr263>
- Lonnemann, J., & Yan, S. (2015). Does number word inversion affect arithmetic processes in adults? *Trends in Neuroscience and Education, 4*(1–2), 1–5. <https://doi.org/10.1016/j.tine.2015.01.002>
- Love, J., Dropmann, D., & Selker, R. (2019). *The jamovi project (Version 1.0)*. Retrieved from <https://www.jamovi.org>
- Lyons, I. M., & Beilock, S. L. (2011). Numerical ordering ability mediates the relation between number-sense and arithmetic competence. *Cognition, 121*(2), 256–261. <https://doi.org/10.1016/j.cognition.2011.07.009>

- Lyons, I. M., Price, G. R., Vaessen, A., Blomert, L., & Ansari, D. (2014). Numerical predictors of arithmetic success in grades 1–6. *Developmental Science*, *17*(5), 714–726. <https://doi.org/10.1111/desc.12152>
- Macizo, P., Herrera, A., Román, P., & Martín, M. C. (2010). Second language acquisition influences the processing of number words. *Procedia - Social and Behavioral Sciences*, *9*, 1128–1134. <https://doi.org/10.1016/j.sbspro.2010.12.295>
- Martin, R., Ugen, S., & Fischbach, A. (2013). *Épreuves Standardisées—Bildungsmonitoring Luxemburg*.
- Méndez, L. I., Hammer, C. S., Lopez, L. M., & Blair, C. (2019). Examining language and early numeracy skills in young Latino dual language learners. *Early Childhood Research Quarterly*, *46*, 252–261. <https://doi.org/10.1016/j.ecresq.2018.02.004>
- Nosworthy, N., Bugden, S., Archibald, L., Evans, B., & Ansari, D. (2013). A Two-Minute Paper-and-Pencil Test of Symbolic and Nonsymbolic Numerical Magnitude Processing Explains Variability in Primary School Children’s Arithmetic Competence. *PLoS ONE*, *8*(7), e67918. <https://doi.org/10.1371/journal.pone.0067918>
- Paetsch, J., Felbrich, A., & Stanat, P. (2015). Der Zusammenhang von sprachlichen und mathematischen Kompetenzen bei Kindern mit Deutsch als Zweitsprache. *Zeitschrift Für Pädagogische Psychologie*, *29*(1), 19–29. <https://doi.org/10.1024/1010-0652/a000142>
- Paetsch, J., Radmann, S., Felbrich, A., Lehmann, R., & Stanat, P. (2016). Sprachkompetenz als Prädiktor mathematischer Kompetenzentwicklung von Kindern deutscher und nicht-deutscher Familiensprache. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *48*(1), 27–41. <https://doi.org/10.1026/0049-8637/a000142>
- Peng, P., & Lin, X. (2019). The relation between mathematics vocabulary and mathematics performance among fourth graders. *Learning and Individual Differences*, *69*, 11–21. <https://doi.org/10.1016/j.lindif.2018.11.006>
- Pixner, S., Moeller, K., Hermanova, V., Nuerk, H.-C., & Kaufmann, L. (2011). Whorf reloaded: Language effects on nonverbal number processing in first grade—A trilingual study. *Journal of Experimental Child Psychology*, *108*(2), 371–382. <https://doi.org/10.1016/j.jecp.2010.09.002>
- Pixner, S., Zuber, J., Heřmanová, V., Kaufmann, L., Nuerk, H.-C., & Moeller, K. (2011). One language, two number-word systems and many problems: Numerical cognition in the Czech language. *Research in Developmental Disabilities*, *32*(6), 2683–2689. <https://doi.org/10.1016/j.ridd.2011.06.004>
- Purpura, D. J., Napoli, A. R., & King, Y. (2019). Chapter 7—Development of Mathematical Language in Preschool and Its Role in Learning Numeracy Skills. In D. C. Geary, D. B. Berch, & K. Mann Koepke (Eds.), *Cognitive Foundations for Improving Mathematical Learning* (pp. 175–193). <https://doi.org/10.1016/B978-0-12-815952-1.00007-4>
- Purpura, D. J., & Reid, E. E. (2016). Mathematics and language: Individual and group differences in mathematical language skills in young children. *Early Childhood Research Quarterly*, *36*, 259–268. <https://doi.org/10.1016/j.ecresq.2015.12.020>
- Reynvoet, B., & Sasanguie, D. (2016). The Symbol Grounding Problem Revisited: A Thorough Evaluation of the ANS Mapping Account and the Proposal of an Alternative Account Based on Symbol–Symbol Associations. *Frontiers in Psychology*, *7*. <https://doi.org/10.3389/fpsyg.2016.01581>
- Riccomini, P. J., Smith, G. W., Hughes, E. M., & Fries, K. M. (2015). The Language of Mathematics: The Importance of Teaching and Learning Mathematical Vocabulary.

- Reading & Writing Quarterly*, 31(3), 235–252.  
<https://doi.org/10.1080/10573569.2015.1030995>
- Schneider, M., Beeres, K., Coban, L., Merz, S., Schmidt, S. S., Stricker, J., & Smedt, B. D. (2017). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: A meta-analysis. *Developmental Science*, 20(3), e12372.  
<https://doi.org/10.1111/desc.12372>
- Schwenk, C., Sasanguie, D., Kuhn, J.-T., Kempe, S., Doebler, P., & Holling, H. (2017). (Non-)symbolic magnitude processing in children with mathematical difficulties: A meta-analysis. *Research in Developmental Disabilities*, 64, 152–167.  
<https://doi.org/10.1016/j.ridd.2017.03.003>
- Sepeng, P., & Madzorera, A. (2014). Sources of Difficulty in Comprehending and Solving Mathematical Word Problems. *International Journal of Educational Sciences*, 6(2), 217–225. <https://doi.org/10.1080/09751122.2014.11890134>
- Singer, V., & Strasser, K. (2017). The association between arithmetic and reading performance in school: A meta-analytic study. *School Psychology Quarterly*, 32(4), 435–448.  
<https://doi.org/10.1037/spq0000197>
- Stock, P., Desoete, A., & Roeyers, H. (2009). Predicting Arithmetic Abilities: The Role of Preparatory Arithmetic Markers and Intelligence. *Journal of Psychoeducational Assessment*, 27(3), 237–251. <https://doi.org/10.1177/0734282908330587>
- Van Rinsveld, A., Brunner, M., Landerl, K., Schiltz, C., & Ugen, S. (2015). The relation between language and arithmetic in bilinguals: Insights from different stages of language acquisition. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00265>
- Van Rinsveld, A., Dricot, L., Guillaume, M., Rossion, B., & Schiltz, C. (2017). Mental arithmetic in the bilingual brain: Language matters. *Neuropsychologia*, 101, 17–29.  
<https://doi.org/10.1016/j.neuropsychologia.2017.05.009>
- Van Rinsveld, A., & Schiltz, C. (2016). Sixty-twelve = Seventy-two? A cross-linguistic comparison of children's number transcoding. *British Journal of Developmental Psychology*, 34(3), 461–468. <https://doi.org/10.1111/bjdp.12151>
- Van Rinsveld, A., Schiltz, C., Landerl, K., Brunner, M., & Ugen, S. (2016). Speaking two languages with different number naming systems: What implications for magnitude judgments in bilinguals at different stages of language acquisition? *Cognitive Processing*, 17(3), 225–241. <https://doi.org/10.1007/s10339-016-0762-9>
- Venkatraman, V., Siong, S. C., Chee, M. W. L., & Ansari, D. (2006). Effect of Language Switching on Arithmetic: A Bilingual fMRI Study. *Journal of Cognitive Neuroscience*, 18(1), 64–74. <https://doi.org/10.1162/089892906775250030>
- Vilenius-Tuohimaa, P. M., Aunola, K., & Nurmi, J.-E. (2008). The association between mathematical word problems and reading comprehension. *Educational Psychology*, 28(4), 409–426. <https://doi.org/10.1080/01443410701708228>
- von Aster, M. G., & Shalev, R. S. (2007). Number development and developmental dyscalculia. *Developmental Medicine & Child Neurology*, 49(11), 868–873.  
<https://doi.org/10.1111/j.1469-8749.2007.00868.x>
- Vukovic, R. K., & Lesaux, N. K. (2013a). The language of mathematics: Investigating the ways language counts for children's mathematical development. *Journal of Experimental Child Psychology*, 115(2), 227–244. <https://doi.org/10.1016/j.jecp.2013.02.002>



- Vukovic, R. K., & Lesaux, N. K. (2013b). The relationship between linguistic skills and arithmetic knowledge. *Learning and Individual Differences, 23*, 87–91.  
<https://doi.org/10.1016/j.lindif.2012.10.007>
- Zhang, X. (2016). Linking language, visual-spatial, and executive function skills to number competence in very young Chinese children. *Early Childhood Research Quarterly, 36*, 178–189. <https://doi.org/10.1016/j.ecresq.2015.12.010>
- Zuber, J., Pixner, S., Moeller, K., & Nuerk, H.-C. (2009). On the language specificity of basic number processing: Transcoding in a language with inversion and its relation to working memory capacity. *Journal of Experimental Child Psychology, 102*(1), 60–77.  
<https://doi.org/10.1016/j.jecp.2008.04.003>

# Discussion

The aim of the hereby presented research project was twofold. On one hand, we wanted to explore a new language-neutral methodology for assessing basic number skills in children. If proven successful, the other aim of the project was to evaluate if this new methodology could be used to reliably assess basic number competence in children entering the formal school curriculum. Current models of normal and problematic development of numerical abilities in children (e.g. von Aster & Shalev, 2007) suggest that the disorder is of developmental nature and takes root in the primary numerical and logical representations acquired during the preschool years. Indeed, the findings presented in the introduction show that performance in preschool number competence is the best and most reliable predictor of later achievement in fundamental school mathematics. It is thus important to detect developmental shortcomings as early as possible in order to remediate efficiently. In practice, this detection is accomplished by using standardized test batteries for number competence. A multitude of different tools exist and are used in practice, however, they come with caveats in view of their verbal load. Research on the interaction between linguistic and numerical abilities suggests that language competence influences performance in mathematics assessment in various ways (Dowker & Nuerk, 2016). These influences become problematic when the language competence of the assessed person is not sufficient for understanding the assessment's instruction or task content. Moreover, the negative influence of insufficient linguistic competence becomes dramatic in highly multilingual school contexts.

## **The relationship between home language, school language and mathematics achievement in a multilingual school setting**

To demonstrate how far reaching this influence can be, I started off the project by analyzing available large-scale data from the school monitoring program of a highly multilingual country, Luxembourg. Luxembourg has three official languages: German, French and Luxembourgish, a language rooted in a German regional dialect. The language of teaching and assessment in Luxembourg from grade 1 to 6 is German. Nevertheless, during preschool, Luxembourgish is the language of instruction for classroom communication and teaching. German is introduced during first grade. This is not problematic for Luxembourg's native speakers. The Luxembourgish language being very close to German, a quasi-automatic transfer from Luxembourgish to German can be assumed (although this assumption lacks any empirical evidence as of now). However, only 36% of Luxembourg's school population in 2018 speaks either German or Luxembourgish at home. This leads to an increasingly challenging learning environment for the majority of pupils in which they do not only have to learn a new language (Luxembourgish) in preschool, but also need to acquire knowledge in this new language at the same time. This challenge becomes more difficult as they enter the first grade of formal schooling and are required to learn a similar, yet different language and yet again need to acquire new knowledge through this new language while not benefitting from the same assumed language transfer that natives do.

The consequences of this challenging learning environment become clear when looking at performance metrics of the language populations. The data presented in the first study report showed that non-native speakers not only underperform in German reading comprehension, but also in mathematics. The results indicated that the achievement differences in mathematics between native and non-native speakers are largely or entirely mediated by their lacking reading

comprehension in the language of *instruction* and the language of *the instructions*. They also suggested that only non-native speakers with above average skill in numerical reasoning can progress without grade repetition in a multilingual school curriculum. In other words, if the children in the presented sample had similar levels of reading comprehension than the Luxembourgish natives, they would tend to outperform them in mathematics. These findings have considerable implications for teaching policies. If the biggest barrier to non-native speaker's success in mathematics is their lacking language competence, then policies should address this by first making sure that foreign speaker's competence in the school language is sufficient *before* integrating them into the regular curriculum with their native peers. While such policies might lead to longer curricula for foreign speakers than for their native peers, it should however enable them to fully participate in all school subjects during their career and to build the foundation for later educational success and quality of life. However, research findings are seldom easy and fast to implement in practice and policy making. In the meantime, I explored a complementary angle of attack in the context of the hereby presented thesis. Instead of trying to better match the language of the testee with the language of the assessment, I designed an assessment method that works without verbal content and whose outcomes should be less or not biased by linguistic factors. The first two studies on this new assessment method were presented in this thesis, the first one focusing more on the methodology itself, while the second one aimed to gather data on the psychometric properties of the tasks, its items and the battery as a whole.

### **NUMTEST and the video-instruction: a possible alternative?**

#### **Video Instructions**

NUMTEST's instructions work by showing, rather than explaining, how a task is performed correctly. In the second study report, I evaluated the efficacy of this method from different angles.

First, I compared performance between participants using this novel method and participants that used a version of the experimental tasks that did not include video instructions. Instead, I formulated a standard instruction in German, the school language, which was read to the participants by test administrators. I compared performance on the tasks in both groups with the hypothesis that if the video instruction failed to convey task instructions effectively, then one should expect a significantly lower average performance in the video group. I did not find a significant difference between performance in both groups, providing primary support for the idea that the video instruction seems to work as well as the verbal instructions. In hindsight, I would have approached this differently. Between the two experimental groups, only the nature of the instruction was controlled, not the task content per se. For some tasks however, like the nonverbal addition and subtraction tasks or the crossmodal addition task, I could have gone further by replacing the animated word problem by a verbal one. Indeed, in the case of word problems, the task content can be viewed as an integral part of the instruction and replacing all visual content by verbal content would have provided an interesting opportunity for observing potential differences between the groups where none were observed in this study. Indeed, I collected information on participant's language background and, while not reported in the published article, I found no significant performance differences between native speakers and non-native speakers in any of the tasks, neither in the verbal group nor in the video group. By translating all visual content into words and considering the findings from the first study report, I would have expected non-native speaker's performance to be lower in the verbal group than in the video group, further corroborating the problematic influence of language skills on mathematics performance.

Nevertheless, we used an additional measure for evaluating the efficacy of the video instruction, namely the repetition of the practice session. After the video instruction, participants could

complete three practice items with feedback allowing them to test if they understood the task as intended. This practice session was repeated once in its entirety when the participant gave a wrong answer on at least one of the three practice items. I thus categorized participants into groups of repeaters and non-repeaters with the idea that a participant who did not repeat the practice session and solved all items immediately after the video instruction must have understood the tasks correctly. The only other explanation would be that the participant would have had to choose the correct answer randomly three times in a row, which is rather unlikely. I then looked at the percentage of participants that repeated the practice session for each task and in each group (video or verbal) in order to test for differences. Summarily, in most tasks, a higher percentage of participants repeated the practice session when they were offered a verbal instruction than in the video instruction group. In some cases, there was no significant difference and, concerning the quantity comparison task, the pattern was reversed. This led to the redesign of the comparison task which will be discussed later. The case of the ordering task is also interesting to point out. Indeed, the difference between the percentage of repeaters in the video and verbal group for this task was much larger than for all other tasks. While this is the result of a suboptimal design choice concerning the verbal instructions (see study report 2 for the explanation), it provided for an excellent example as to how a single word in the instruction of a task can have overwhelming effects on the way it is interpreted and realized. The fact that most of the time, viewing a video instruction instead of hearing a verbal instruction led to less people repeating the practice session is another strong indicator that the video instruction works at least as well as an explicit verbal instruction. Taken together, these two findings provided for a good proof of concept for the video-instruction method. Video instructions seemed to be equivalent or better than verbal instruction,

both in terms of subsequent performance on the tasks and in terms of rapidity of comprehension, all while providing a way of assessing participants without language interference.

### **Practice phase**

As stated before, after the children viewed the video instruction for each task, they were presented with three items to try themselves. These items were very similar in difficulty to those seen during the instruction phase. After a correct answer, they were presented a green smiling face, whereas an incorrect answer is followed by the display of a red, unhappy smile. We expected that children at the beginning of first grade would intuitively understand the meaning of the smile faces as they are commonly used both in classrooms and software designed for children to signify *good* and *bad*. Our expectations were confirmed during the pilot studies, in which I asked the children to explain the meaning of the feedback symbols to me. Almost all children could convincingly explain to me or the other test administrators what they meant. The same is true for the little blue arrow on the top right of the screen, which the instruction showed as a mean to confirm your answer and to move on to the next item in the task. Again, there was hardly any child that could not explain the function of the arrow to me. However, I observed that a few more children did not understand the role of the arrow in the first task in the verbal group. This was probably due to the fact that the video instruction showed not only *what* the task was about, but also *how* to use the application in general as it showed a hand confirming the answer and moving to the next item by touching the blue arrow. The part on how to use the application was not present in the verbal instruction I gave to the children, which explains why some of them could not say what the arrow was about after the first task. Still, most children could intuitively explain the function of the arrow even without seeing it in action.

Generally speaking, the practice phase thus worked as intended. It provides an opportunity for the participants to experiment with their understanding of the instruction before moving to the part where their competence is effectively assessed. Nevertheless, the fact that all items were repeated in case of a single mistake has led to confusion in some children. Why would they have to repeat an item although there was a green smile? From observation I know that some children changed their correct answer to an incorrect one, likely because the application asked them to reply again on the same item. This was a design constraint of the OASYS framework as it does not provide for real test-branching capabilities. The software is designed to present the same test to all participants in a linear manner. The programmer of OASYS managed to implement a work-around for checking intermediate answers and adapting the next steps of the test as a function of the answers given by the participant. We used this method to provide for very basic branching: Either the participant provided only correct answers, and the application would move on to the assessment items, or, the participant replied incorrectly to at least one of the presented items, leading to the repetition of the entire practice block. This solution is far from ideal and should be improved in a future version of the NUMTEST battery. The way I imagine the instruction / practice interplay ideally is as follows. Participants are shown the video instruction, but at the rate of one item at a time. After each video-item, they would see a practice item to try themselves. If they reply correctly, a second practice item would appear, just to confirm that their first answer was not due to chance. If answered correctly, the application would then move on to the assessment part. In case of a wrong answer on the first practice item, the application would then show a second video instruction, again followed by another practice item. This would go on for as long as the participant is unable to solve two practice items correctly in a row. Compared to the paradigm used in the current version of NUMTEST, this would allow for a more dynamic and customized approach to



the instruction and practice phase of the test. Participants that require more trials before grasping the task at hand could have as many as they need, while children that are faster to understand the task would not be presented with unnecessary repetitions of items that they solved correctly before and would be able to move on to the assessment part faster. One could also imagine a way for children to replay an instruction video on their own. Indeed, in the current version, the three instruction videos are being played consequentially without possible stops between items. This provides for strict standardization of the instruction and practice phases but in practice we observed that children are not always ready to focus when we expect them to focus, and therefore some children miss the first instruction item, turn around to ask what to do, leading to them missing the next item and finally not grasping the task altogether. While this was a rare occurrence during the three studies using NUMTEST, it could be easily avoided by allowing children more control over the timing of the instructions. All in all, the instruction and practice phase worked as planned, but there is much space for improving the efficiency of the procedure in the future.

### **Changes to tasks and the battery over the course of development**

Over the course of NUMTEST's development, two different versions of the battery existed and, in this part, I want to present the rationale behind the modifications to both the tasks and the composition of the battery. The first version was designed with two aims in mind. The first aim was to offer not only different tasks but also many variations of the same task to check which ones would be understood best and how much abstraction I could expect from children of that age. Many different stimuli were used in all tasks, ranging from depictions of different fruits to depictions of people, black dots, tools or shapes. The primary idea behind this colorful presentation was to provide for interesting and thus engaging content for the children. In some tasks however, varying the form of the stimulus was also integral to the construct I aimed to measure: Numeracy

is an inherent property of any collection of objects and as such, the nature of the objects does not matter. However, with children this young, I quickly noticed that they would regularly lose focus on the numeracy aspect of the presented tasks and preferred to tell me which fruit they liked best.

Two tasks are worth mentioning specifically in this context. First, I want to discuss the quantity correspondence task. As the reader will have noticed in the second study report, there were five different variations of the same task. While the purpose of the task was always to match a centrally presented quantity with one of three quantities presented in the answer section, the nature of the stimuli varied so heavily that it was difficult for many children to grasp the essence of the task on the basis of three video demonstrations as these could only depict three of the five possible scenarios. Moreover, the task was in fact not always the same. For example, while using identical objects in the center and in the answers, the task measures counting ability and the principle of one-to-one correspondence. Another set of items still used only non-symbolical quantities but used different objects for the stimulus than for the answer choices. While the task still measures counting and one-to-one correspondence, it does so at a higher level of abstraction than the previous set. In another set, I mixed non-symbolic quantities with Arabic digits. The task is similar yet different to the previous one: Count the objects and point to the same numerosity. However, the numerosity was now represented as a digit in the answer section and required an additional cognitive function, that of transcoding the numerosity from one format into the other. That is a different task than the previous iteration in which no transcoding or symbolical knowledge was required. Moreover, no measure of the principle of one-to-one correspondence was present in this variant. The various item sets of this task could thus not be measuring the same skill. For this reason, and because cuts had to be made to the second version of NUMTEST due to time

constraints, the task was put aside for the second version of NUMTEST but should be revisited and optimized in future versions.

The high variability of stimuli in NUMTEST's pilot tasks lead to many difficulties in transmitting a clear instruction through the video-method in other tasks. One task of note in this context is the comparison task. In the first version of the task, children were shown ten repetitions of a hand pushing on the larger of two numerosities before being offered the practice items. However, the ten repetitions they were shown figured collections of very different objects or even a single collection of objects on one side of the display and an Arabic numeral on the other display. The idea was again to show that no matter the nature of the quantity, you should touch the side with the greater numerosity. This did not work as intended as the high variability in the examples shown lead to participating missing the primary intent of the task. As such, considerably more participants repeated the practice session of this task in the video-instruction group than in the verbal group. Implicitly identifying the instruction for this task was simply not as effective as being told to push on the display that contains *more*. Another consequence of the high variability of the items presented in this first version of the tasks was that their reliability was questionable. The items seemed to measure something different in each participant and could thus hardly be considered for screening purposes in their current form. I didn't consider this to be problematic, since the purpose of this first version was to evaluate if tasks of the type I designed could in principle be understood by providing only implicit video instructions, without emphasizing the psychometric aspects of the items themselves. My observations on these two tasks nevertheless led me to the conclusion that the variability and colorfulness of the items were causing too many problems and so I decided to streamline the design of not only the problematic tasks, but of all tasks.

Indeed, in the second and most recent version of NUMTEST, all colorful images were removed and only two different formats of each task are presented: A symbolic response format (Arabic digits) and a non-symbolic response format, represented by black dots or, depending on the task, by objects identical to the ones used in the question. Almost no modifications were applied to the non-symbolic addition / subtraction tasks or the crossmodal addition task after the pilot studies, as their formats were less variable to begin with. I slightly adapted the length of the animations and the size of the answer images after the pilot studies revealed that for some children, the animations played too quickly and that the objects in the answer fields were difficult to see from a reasonable distance. The ordering task was similarly redesigned, with many image variants being removed and replaced by non-symbolic representations (black dots) and Arabic digits. Finally, the first version of NUMTEST included a symbolic arithmetic task. While it worked well and was correctly solved by most participants, it provided no visible added value over a paper and pencil arithmetic task (such as the Tempo Test Rekenen) and when cuts had to be made for time constraints, it was removed from the battery.

### **NUMTEST as a screener for early numerical competence**

The two pilot studies lead me to conclude that the video instruction method works very well with all but a few children and that it constitutes a solid basis for building an assessment battery. In the next step, I then completed a preliminary validation study in order to evaluate the psychometric structure of the tasks once the teachings from the pilot studies had been implemented into the second version of NUMTEST. The results of this study were reported in the third and final study report in the present thesis. In summary, the tasks and the battery provide for good to excellent inter-subject reliability, correlate strongly and significantly with standardized measures of arithmetic performance (TTR, (De Vos, 1992) ) and deliver similar performance classification

than existing screeners for math learning difficulties (Numeracy Screener, (Nosworthy, Bugden, Archibald, Evans, & Ansari, 2013)). Additionally, the NUMTEST battery provides a more diverse picture of a child's basic numerical abilities than existing screeners. Preliminary factor analysis showed that the tasks in the NUMTEST battery seem to measure four distinct competencies that overlap with the four different tasks in the battery. However, while these preliminary results are promising, this study was only the first step towards complete psychometric validation. On one hand, the sample I could measure for this study was rather small and does not permit the drawing of robust conclusions on the validity aspects of the battery. Moreover, while I could provide encouraging data on the concurrent validity of NUMTEST, external predictive validity could not be established yet. Measuring and establishing predictive validity is of essence in order to use performance on NUMTEST as a valid predictor for later mathematics achievement. However, establishing predictive validity is not the only future step in my vision of NUMTEST's further development, which leads me to final part of the hereby presented thesis: What is next?

### **General conclusions and the future of NUMTEST**

Before I conclude, I want to discuss the many possible ways in which NUMTEST should be further developed and prepared for being used as a language-neutral screener for basic math competence in multilingual contexts. First, the second version of the test battery takes longer to complete than it should, considering the still relatively short attention span of our target population. One way to cut down on completion time would be the elimination of certain items from the pool. Data provided by the validation study showed that many items were correctly solved by almost all participants while others were completed by only a few. While setting an exact criterion for necessary and sufficient item difficulty would have to be determined according to the needs of the context in which NUMTEST would be deployed, many items could be removed from the battery

without significantly compromising the battery's reliability. Other time savings would be possible by switching to a dynamic branching system for the instruction and practice phases of NUMTEST as described earlier in the discussion of NUMTEST methodology. While the proposed solution would lengthen the duration of the procedure for some participants, I am confident that most of them would be faced with a shorter instruction and practice phase than in the current implementation. After all, a large majority of children don't have math learning difficulties and immediately understand tasks like the ones presented in NUMTEST.

After implementing the proposed modifications, a new version of NUMTEST would then have to be submitted to a new, large sample for evaluating different facets of validity. As was explained before, predictive validity for primary school mathematics achievement could not be measured during this project. This is an essential requisite before using NUMTEST as a screener in practice and should be completed by linking participant's performance on NUMTEST with a standardized measure of mathematical achievement, both at the time of collection (concurrent validity) and at a later time point during the curriculum (predictive validity). Another important endeavor in such a future study would be to measure the battery's test-retest reliability over time in order to complement inter-subject reliability as measured in my first validation study. Additionally, reverse validation should be considered in the future. For such an undertaking, children that were previously properly diagnosed as dyscalculic with an existing standardized battery would be asked to complete the NUMTEST battery in order to evaluate if it would draw a similar conclusion.

After these studies, NUMTEST would then be technically fit the minimum criteria for being employed as a screener for math learning difficulties in Luxembourg. However, a claim of this project is to provide a language-neutral tool for universally assessing a child's basic math competence, independently from its language background. Anecdotally, the application was

presented on a science fair in Luxembourg in 2018. During the fair, a visiting Russian family with their 7-year-old daughter stopped at our booth. Without any incentive or instruction, the girl sat in front of a NUMTEST tablet and went on to flawlessly complete the entire battery on her own. Beyond this anecdote, the data on participant's language background that was analyzed in the third study report provides first evidence to the claim of NUMTEST's universality. Nonetheless, further substantiating this claim would require the administration of the NUMTEST battery to different samples of children at the beginning of first grade in different countries. As shown in the introduction, a standardized, linguistically unbiased measure of basic number competence than can be deployed in any country would significantly contribute to establishing an objective, universal definition of dyscalculia or less severe math learning disorders.

Going further, as previously noted, task design was not only driven by theoretical considerations about the predictive value of each measured competence, but also by what was and was not possible to implement given the technology I had at my disposal. Ideally, NUMTEST should be professionally redesigned from scratch as a standalone web application while using the existing product as a blueprint. This would provide for higher customizability of the proposed tasks, improved graphics and animation quality and generally more freedom when it comes to developing dynamic tasks that go beyond the question-answer format I used so far. For example, for quantity correspondence, one could imagine a task displaying an animated balance with a given quantity on one side while the participant has to put *as much* (be it as many discrete objects, for a non-symbolic form, or an Arabic digit symbolizing the equivalent quantity for a symbolic version) on the other side in order to maintain the balance's equilibrium. Another task that was a promising candidate for inclusion in the NUMTEST battery in the beginning stages of development was the number-line estimation task, of which certain variants constitute solid predictors for later

mathematics achievement (Schneider et al., 2018). However, the technical limitations of the OASYS framework made such a task impossible to implement and it was thus discarded. On another note, the focus of the current NUMTEST battery lays on assessing children's answer *accuracy*, but offers no measure of answer *time*. Indeed, most available batteries share this focus although research has shown that children with mathematical learning disorders often do manage to solve the problems we present to them, but it takes them significantly longer to do so (Kaufmann & Aster, 2012). I did not delve into this aspect during NUMTEST's development, again because the technical framework I used so far could not provide for reliable time measurement without significant additional programming work.

Finally, one could consider adapting the method and the battery to older populations by extending the numerical range of its items. Although the research I presented clearly points to preschool-level basic number competence as the root of all evil, in practice it is not uncommon to observe that although children with mathematics learning disorders have problems in understanding basic concepts of magnitude and its symbolizations, they will nevertheless acquire factual and procedural knowledge in the lower ranges through repeated exposure and training over time. In this case, a child's or even an adult's difficulties would only become apparent when confronted with a higher numerical range and problems that have not been learned by heart. These are just a few examples of how NUMTEST could evolve, but many other ideas could be implemented using NUMTEST's methodology for task instruction, with the only limiting factor being the amount of dedicated resources.

In conclusion, the research presented in this thesis provides important foundations for the future development of language-neutral assessment paradigms. As multilingual societies tend to become the norm rather than the exception in a globalized world, dependence on competence in multiple



languages for many aspects of life will become an increasingly widespread issue. My research shows that for assessing basic numerical concepts in children, it is possible to do without language and instead use a quasi-natural method of conveying information. After all, what do you do when you want to communicate with someone that doesn't speak your language? Instead of talking, you will try to use signing and symbols. You will try to decouple meaning from its symbol, or, to use Ferdinand de Saussure's jargon, you will try to return to the immediate meaning of the *signified* while bypassing the *signifier*. I operationalized and formalized this approach by replacing words with videos and animations that *show* rather than *explain* what to do. Never would I have thought that it works so well.

### Additional references (Introduction & Discussion)

- Abedi, J. (2002). Standardized Achievement Tests and English Language Learners: Psychometrics Issues. *Educational Assessment*, 8(3), 231–257. [https://doi.org/10.1207/S15326977EA0803\\_02](https://doi.org/10.1207/S15326977EA0803_02)
- Abedi, J., & Lord, C. (2001). The Language Factor in Mathematics Tests. *Applied Measurement in Education*, 14(3), 219–234. [https://doi.org/10.1207/S15324818AME1403\\_2](https://doi.org/10.1207/S15324818AME1403_2)
- Agrillo, C., Dadda, M., Serena, G., & Bisazza, A. (2009). Use of number by fish. *PloS One*, 4(3), e4786.
- Alloway, T. P., & Passolunghi, M. C. (2011). The relationship between working memory, IQ, and mathematical skills in children. *Learning and Individual Differences*, 21(1), 133–137. <https://doi.org/10.1016/j.lindif.2010.09.013>
- American Psychiatric Association (Ed.). (1998). *Diagnostic and statistical manual of mental disorders: DSM-IV ; includes ICD-9-CM codes effective 1. Oct. 96* (4. ed., 7. print). Washington, DC.
- American Psychiatric Association (Ed.). (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5. ed). Washington, DC: American Psychiatric Publishing.
- Aunio, P., & Niemivirta, M. (2010). Predicting children’s mathematical performance in grade one by early numeracy. *Learning and Individual Differences*, 20(5), 427–435. <https://doi.org/10.1016/j.lindif.2010.06.003>
- Aunola, K., Leskinen, E., Lerkkanen, M.-K., & Nurmi, J.-E. (2004). Developmental Dynamics of Math Performance From Preschool to Grade 2. *Journal of Educational Psychology*, 96(4), 699–713. <https://doi.org/10.1037/0022-0663.96.4.699>
- Austin, J. L., & Howson, A. G. (1979). Language and mathematical education. *Educational Studies in Mathematics*, 10(2), 161–197. <https://doi.org/10.1007/BF00230986>
- Baddeley, A. D., & Hitch, G. (1974). Working memory. In *Psychology of learning and motivation* (Vol. 8, pp. 47–89). Elsevier.
- Bartelet, D., Vaessen, A., Blomert, L., & Ansari, D. (2014). What basic number processing measures in kindergarten explain unique variability in first-grade arithmetic proficiency? *Journal of Experimental Child Psychology*, 117, 12–28. <https://doi.org/10.1016/j.jecp.2013.08.010>
- Benson-Amram, S., Gilfillan, G., & McComb, K. (2018). Numerical assessment in the wild: Insights from social carnivores. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1740), 20160508.
- Bernardo, A. B. I. (1999). Overcoming Obstacles to Understanding and Solving Word Problems in Mathematics. *Educational Psychology*, 19(2), 149–163. <https://doi.org/10.1080/0144341990190203>
- Bull, R., & Lee, K. (2014). Executive Functioning and Mathematics Achievement. *Child Development Perspectives*, 8(1), 36–41. <https://doi.org/10.1111/cdep.12059>
- Casey, B. M., Pezaris, E., Fineman, B., Pollock, A., Demers, L., & Dearing, E. (2015). A longitudinal analysis of early spatial skills compared to arithmetic and verbal skills as predictors of fifth-grade girls’ math reasoning. *Learning and Individual Differences*, 40, 90–100. <https://doi.org/10.1016/j.lindif.2015.03.028>
- Caviola, S., Mammarella, I. C., Lucangeli, D., & Cornoldi, C. (2014). Working memory and domain-specific precursors predicting success in learning written subtraction problems.

- Learning and Individual Differences*, 36, 92–100.  
<https://doi.org/10.1016/j.lindif.2014.10.010>
- Chen, Q., & Li, J. (2014). Association between individual differences in non-symbolic number acuity and math performance: A meta-analysis. *Acta Psychologica*, 148, 163–172.  
<https://doi.org/10.1016/j.actpsy.2014.01.016>
- Chiswick, B. R., Lee, Y. L., & Miller, P. W. (2003). Schooling, Literacy, Numeracy and Labour Market Success. *Economic Record*, 79(245), 165–181. <https://doi.org/10.1111/1475-4932.t01-1-00096>
- Clements, D. H., Sarama, J., & Germeroth, C. (2016). Learning executive function and early mathematics: Directions of causal relations. *Early Childhood Research Quarterly*, 36, 79–90. <https://doi.org/10.1016/j.ecresq.2015.12.009>
- Cormier, D. C., Bulut, O., McGrew, K. S., & Singh, D. (2017). Exploring the Relations between Cattell–Horn–Carroll (CHC) Cognitive Abilities and Mathematics Achievement. *Applied Cognitive Psychology*, 31(5), 530–538. <https://doi.org/10.1002/acp.3350>
- 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>
- Dehaene, S. (2011). *The number sense: How the mind creates mathematics* (Rev. and updated ed). New York, NY: Oxford University Press.
- Devine, A., Soltész, F., Nobes, A., Goswami, U., & Szűcs, D. (2013). Gender differences in developmental dyscalculia depend on diagnostic criteria. *Learning and Instruction*, 27, 31–39. <https://doi.org/10.1016/j.learninstruc.2013.02.004>
- De Vos, T. (1992). Tempo-Test-Rekenen. *Handleiding.[Tempo Test Arithmetic. Manual]*. Nijmegen: Berkhout.
- Dowker, A., & Nuerk, H.-C. (2016). Editorial: Linguistic Influences on Mathematics. *Frontiers in Psychology*, 7, 1035. <https://doi.org/10.3389/fpsyg.2016.01035>
- Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., ... Duckworth, K. (2007). *School Readiness and Later Achievement*. 20.
- Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, 8(7), 307–314. <https://doi.org/10.1016/j.tics.2004.05.002>
- Friso-van den Bos, I., van der Ven, S. H. G., Kroesbergen, E. H., & van Luit, J. E. H. (2013). Working memory and mathematics in primary school children: A meta-analysis. *Educational Research Review*, 10, 29–44. <https://doi.org/10.1016/j.edurev.2013.05.003>
- Geary, D. C. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin*, 114(2), 345–362. <https://doi.org/10.1037/0033-2909.114.2.345>
- Geary, D. C. (2010). Mathematical disabilities: Reflections on cognitive, neuropsychological, and genetic components. *Learning and Individual Differences*, 20(2), 130–133.  
<https://doi.org/10.1016/j.lindif.2009.10.008>
- Gelman, R., & Butterworth, B. (2005). Number and language: How are they related? *Trends in Cognitive Sciences*, 9(1), 6–10. <https://doi.org/10.1016/j.tics.2004.11.004>
- Gilmore, C., Göbel, S. M., & Inglis, M. (2018). *An introduction to mathematical cognition*. Routledge.
- Gjicali, K., Astuto, J., & Lipnevich, A. A. (2019). Relations among language comprehension, oral counting, and numeral knowledge of ethnic and racial minority young children from

- low-income communities. *Early Childhood Research Quarterly*, 46, 5–19.  
<https://doi.org/10.1016/j.ecresq.2018.07.007>
- Göbel, S. M., Watson, S. E., Lervåg, A., & Hulme, C. (n.d.). *Children's Arithmetic Development: It Is Number Knowledge, Not the Approximate Number Sense, That Counts*. 11.
- Green, C. T., Bunge, S. A., Briones Chiongbian, V., Barrow, M., & Ferrer, E. (2017). Fluid reasoning predicts future mathematical performance among children and adolescents. *Journal of Experimental Child Psychology*, 157, 125–143.  
<https://doi.org/10.1016/j.jecp.2016.12.005>
- Hawes, Z., Nosworthy, N., Archibald, L., & Ansari, D. (2019). Kindergarten children's symbolic number comparison skills relates to 1st grade mathematics achievement: Evidence from a two-minute paper-and-pencil test. *Learning and Instruction*, 59, 21–33.  
<https://doi.org/10.1016/j.learninstruc.2018.09.004>
- Hickendorff, M. (2013). The Language Factor in Elementary Mathematics Assessments: Computational Skills and Applied Problem Solving in a Multidimensional IRT Framework. *Applied Measurement in Education*, 26(4), 253–278.  
<https://doi.org/10.1080/08957347.2013.824451>
- Hornburg, C. B., Schmitt, S. A., & Purpura, D. J. (2018). Relations between preschoolers' mathematical language understanding and specific numeracy skills. *Journal of Experimental Child Psychology*, 176, 84–100. <https://doi.org/10.1016/j.jecp.2018.07.005>
- Hornung, C., Schiltz, C., Brunner, M., & Martin, R. (2014). Predicting first-grade mathematics achievement: The contributions of domain-general cognitive abilities, nonverbal number sense, and early number competence. *Frontiers in Psychology*, 5.  
<https://doi.org/10.3389/fpsyg.2014.00272>
- Howard, S. R., Avarguès-Weber, A., Garcia, J. E., Greentree, A. D., & Dyer, A. G. (2019). Numerical cognition in honeybees enables addition and subtraction. *Science Advances*, 5(2), eaav0961. <https://doi.org/10.1126/sciadv.aav0961>
- Imbo, I., Vanden Bulcke, C., De Brauwer, J., & Fias, W. (2014). Sixty-four or four-and-sixty? The influence of language and working memory on children's number transcoding. *Frontiers in Psychology*, 5, 313.
- Jacobs, C., & Petermann, F. (2014). *Rechenfertigkeiten- und Zahlenverarbeitungs-Diagnostikum für die 2. Bis 6. Klasse (RZD 2-6)* (Hogrefe).
- Jordan, N. C., Glutting, J., & Ramineni, C. (2010). The importance of number sense to mathematics achievement in first and third grades. *Learning and Individual Differences*, 20(2), 82–88. <https://doi.org/10.1016/j.lindif.2009.07.004>
- Kaufmann, L., & Aster, M. von. (2012). The Diagnosis and Management of Dyscalculia. *Deutsches Ärzteblatt Online*. <https://doi.org/10.3238/arztebl.2012.0767>
- Kempert, S., Edele, A., Rauch, D., Wolf, K. M., Paetsch, J., Darsow, A., ... Stanat, P. (2016). Die Rolle der Sprache für zugewanderungsbezogene Ungleichheiten im Bildungserfolg. In *Ethnische Ungleichheiten im Bildungsverlauf* (pp. 157–241). Springer.
- Kempert, S., Saalbach, H., & Hardy, I. (2011). Cognitive benefits and costs of bilingualism in elementary school students: The case of mathematical word problems. *Journal of Educational Psychology*, 103(3), 547–561. <https://doi.org/10.1037/a0023619>
- Korpipää, H., Koponen, T., Aro, M., Tolvanen, A., Aunola, K., Poikkeus, A.-M., ... Nurmi, J.-E. (2017). Covariation between reading and arithmetic skills from Grade 1 to Grade 7.

- Contemporary Educational Psychology*, 51, 131–140.  
<https://doi.org/10.1016/j.cedpsych.2017.06.005>
- Krajewski, K., & Schneider, W. (2009). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: Findings from a four-year longitudinal study. *Learning and Instruction*, 19(6), 513–526.  
<https://doi.org/10.1016/j.learninstruc.2008.10.002>
- Krinzinger, H., Gregoire, J., Desoete, A., Kaufmann, L., Nuerk, H.-C., & Willmes, K. (2011). Differential Language Effects on Numerical Skills in Second Grade. *Journal of Cross-Cultural Psychology*, 42(4), 614–629. <https://doi.org/10.1177/0022022111406252>
- Laski, E. V., Casey, B. M., Yu, Q., Dulaney, A., Heyman, M., & Dearing, E. (2013). Spatial skills as a predictor of first grade girls' use of higher level arithmetic strategies. *Learning and Individual Differences*, 23, 123–130. <https://doi.org/10.1016/j.lindif.2012.08.001>
- Locuniak, M. N., & Jordan, N. C. (2008). Using kindergarten number sense to predict calculation fluency in second grade. *Journal of Learning Disabilities*, 41(5), 451–459.
- Lyons, I. M., Price, G. R., Vaessen, A., Blomert, L., & Ansari, D. (2014). Numerical predictors of arithmetic success in grades 1–6. *Developmental Science*, 17(5), 714–726.  
<https://doi.org/10.1111/desc.12152>
- Marian, V., & Fausey, C. M. (2006). Language-dependent memory in bilingual learning. *Applied Cognitive Psychology*, 20(8), 1025–1047. <https://doi.org/10.1002/acp.1242>
- Martin, R. B., Cirino, P. T., Sharp, C., & Barnes, M. (2014). Number and counting skills in kindergarten as predictors of grade 1 mathematical skills. *Learning and Individual Differences*, 34, 12–23. <https://doi.org/10.1016/j.lindif.2014.05.006>
- Martin, R., Ugen, S., & Fischbach, A. (2013). *Épreuves Standardisées—Bildungsmonitoring Luxemburg*.
- Matsuzawa, T. (2009). *Primate origins of human cognition and behavior*. Springer Science & Business Media.
- MENFP. (2011). *Plan d'études: Ecole fondamentale*.
- Mix, K. S., Levine, S. C., Cheng, Y.-L., Young, C., Hambrick, D. Z., Ping, R., & Konstantopoulos, S. (2016). Separate but correlated: The latent structure of space and mathematics across development. *Journal of Experimental Psychology: General*, 145(9), 1206–1227. <https://doi.org/10.1037/xge0000182>
- Moeller, K., Shaki, S., Göbel, S. M., & Nuerk, H.-C. (2015). Language influences number processing – A quadrilingual study. *Cognition*, 136, 150–155.  
<https://doi.org/10.1016/j.cognition.2014.11.003>
- Mou, Y., Berteletti, I., & Hyde, D. C. (2018). What counts in preschool number knowledge? A Bayes factor analytic approach toward theoretical model development. *Journal of Experimental Child Psychology*, 166, 116–133.  
<https://doi.org/10.1016/j.jecp.2017.07.016>
- Nguyen, T., Watts, T. W., Duncan, G. J., Clements, D. H., Sarama, J. S., Wolfe, C., & Spitler, M. E. (2016). Which preschool mathematics competencies are most predictive of fifth grade achievement? *Early Childhood Research Quarterly*, 36, 550–560.  
<https://doi.org/10.1016/j.ecresq.2016.02.003>
- Noël, M.-P., Grégoire, J., & Nieuwenhoven, V. (2008). *Test diagnostique des compétences de base en mathématiques* (Editions d).
- Nosworthy, N., Bugden, S., Archibald, L., Evans, B., & Ansari, D. (2013). A Two-Minute Paper-and-Pencil Test of Symbolic and Nonsymbolic Numerical Magnitude Processing

- Explains Variability in Primary School Children's Arithmetic Competence. *PLoS ONE*, 8(7), e67918. <https://doi.org/10.1371/journal.pone.0067918>
- OECD. (2016). *PISA 2015 Results (Volume I)*. Retrieved from <https://www.oecd-ilibrary.org/content/publication/9789264266490-en>
- Paetsch, J., Radmann, S., Felbrich, A., Lehmann, R., & Stanat, P. (2016). Sprachkompetenz als Prädiktor mathematischer Kompetenzentwicklung von Kindern deutscher und nicht-deutscher Familiensprache. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 48(1), 27–41. <https://doi.org/10.1026/0049-8637/a000142>
- Passolunghi, M. C., Lanfranchi, S., Altoè, G., & Sollazzo, N. (2015). Early numerical abilities and cognitive skills in kindergarten children. *Journal of Experimental Child Psychology*, 135, 25–42. <https://doi.org/10.1016/j.jecp.2015.02.001>
- Peng, P., & Lin, X. (2019). The relation between mathematics vocabulary and mathematics performance among fourth graders. *Learning and Individual Differences*, 69, 11–21. <https://doi.org/10.1016/j.lindif.2018.11.006>
- Pixner, S., Moeller, K., Hermanova, V., Nuerk, H.-C., & Kaufmann, L. (2011). Whorf reloaded: Language effects on nonverbal number processing in first grade—A trilingual study. *Journal of Experimental Child Psychology*, 108(2), 371–382. <https://doi.org/10.1016/j.jecp.2010.09.002>
- Pixner, S., Zuber, J., Heřmanová, V., Kaufmann, L., Nuerk, H.-C., & Moeller, K. (2011). One language, two number-word systems and many problems: Numerical cognition in the Czech language. *Research in Developmental Disabilities*, 32(6), 2683–2689. <https://doi.org/10.1016/j.ridd.2011.06.004>
- Poncin, A., Van Rinsveld, A., & Schiltz, C. (2018). L'apprentissage de l'arithmétique chez les individus bilingues. *ANAE. Approche Neuropsychologique Des Apprentissages Chez l'enfant*, 30(156), 586–595.
- Purpura, D. J., Baroody, A. J., & Lonigan, C. J. (2013). The transition from informal to formal mathematical knowledge: Mediation by numeral knowledge. *Journal of Educational Psychology*, 105(2), 453–464. <https://doi.org/10.1037/a0031753>
- Purpura, D. J., Napoli, A. R., & King, Y. (2019). Chapter 7—Development of Mathematical Language in Preschool and Its Role in Learning Numeracy Skills. In D. C. Geary, D. B. Berch, & K. Mann Koepke (Eds.), *Cognitive Foundations for Improving Mathematical Learning* (pp. 175–193). <https://doi.org/10.1016/B978-0-12-815952-1.00007-4>
- Purpura, D. J., & Reid, E. E. (2016). Mathematics and language: Individual and group differences in mathematical language skills in young children. *Early Childhood Research Quarterly*, 36, 259–268. <https://doi.org/10.1016/j.ecresq.2015.12.020>
- 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>
- Riccomini, P. J., Smith, G. W., Hughes, E. M., & Fries, K. M. (2015). The Language of Mathematics: The Importance of Teaching and Learning Mathematical Vocabulary. *Reading & Writing Quarterly*, 31(3), 235–252. <https://doi.org/10.1080/10573569.2015.1030995>
- Rugani, R., Vallortigara, G., Priftis, K., & Regolin, L. (2015). Number-space mapping in the newborn chick resembles humans' mental number line. *Science*, 347(6221), 534. <https://doi.org/10.1126/science.aaa1379>

- Saalbach, H., Eckstein, D., Andri, N., Hobi, R., & Grabner, R. H. (2013). When language of instruction and language of application differ: Cognitive costs of bilingual mathematics learning. *Learning and Instruction, 26*, 36–44. <https://doi.org/10.1016/j.learninstruc.2013.01.002>
- Salillas, E., & Carreiras, M. (2014). Core number representations are shaped by language. *Cortex, 52*(1), 1–11. <https://doi.org/10.1016/j.cortex.2013.12.009>
- Sasanguie, D., Göbel, S. M., Moll, K., Smets, K., & Reynvoet, B. (2013). Approximate number sense, symbolic number processing, or number–space mappings: What underlies mathematics achievement? *Journal of Experimental Child Psychology, 114*(3), 418–431. <https://doi.org/10.1016/j.jecp.2012.10.012>
- Sasanguie, D., Van den Bussche, E., & Reynvoet, B. (2012). Predictors for Mathematics Achievement? Evidence From a Longitudinal Study. *Mind, Brain, and Education, 6*(3), 119–128. <https://doi.org/10.1111/j.1751-228X.2012.01147.x>
- Schaupp, H., Holzer, N., & Lenart, F. (2007). ERT 1+. Eggenberger Rechentest 1+. *Diagnostikum Für Dyskalkulie Für Das Ende Der, 1*.
- Schneider, M., Beeres, K., Coban, L., Merz, S., Schmidt, S. S., Stricker, J., & Smedt, B. D. (2017). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: A meta-analysis. *Developmental Science, 20*(3), e12372. <https://doi.org/10.1111/desc.12372>
- Schneider, M., Merz, S., Stricker, J., Smedt, B. D., Torbeyns, J., Verschaffel, L., & Luwel, K. (2018). Associations of Number Line Estimation With Mathematical Competence: A Meta-analysis. *Child Development, 89*(5), 1467–1484. <https://doi.org/10.1111/cdev.13068>
- Sepeng, P., & Madzorera, A. (2014). Sources of Difficulty in Comprehending and Solving Mathematical Word Problems. *International Journal of Educational Sciences, 6*(2), 217–225. <https://doi.org/10.1080/09751122.2014.11890134>
- Siegler, R. S., & Mu, Y. (2008). Chinese children excel on novel mathematics problems even before elementary school. *Psychological Science, 19*(8), 759–763.
- Singer, V., & Strasser, K. (2017). The association between arithmetic and reading performance in school: A meta-analytic study. *School Psychology Quarterly, 32*(4), 435–448. <https://doi.org/10.1037/spq0000197>
- Spelke, E. S., & Tsivkin, S. (2001). Language and number: A bilingual training study. *Cognition, 78*(1), 45–88. [https://doi.org/10.1016/S0010-0277\(00\)00108-6](https://doi.org/10.1016/S0010-0277(00)00108-6)
- Stock, P., Desoete, A., & Roeyers, H. (2009). Predicting Arithmetic Abilities: The Role of Preparatory Arithmetic Markers and Intelligence. *Journal of Psychoeducational Assessment, 27*(3), 237–251. <https://doi.org/10.1177/0734282908330587>
- Van de Weijer-Bergsma, E., Kroesbergen, E. H., & Van Luit, J. E. H. (2015). Verbal and visual-spatial working memory and mathematical ability in different domains throughout primary school. *Memory & Cognition, 43*(3), 367–378. <https://doi.org/10.3758/s13421-014-0480-4>
- van Luit, J. E. H., van de Rijt, B. A. M., & Hasemann, K. (2001). *Osnabrücker Test zur Zahlbegriffsentwicklung: OTZ*. Hogrefe, Verlag für Psychologie.
- Van Rinsveld, A., Dricot, L., Guillaume, M., Rossion, B., & Schiltz, C. (2017). Mental arithmetic in the bilingual brain: Language matters. *Neuropsychologia, 101*, 17–29. <https://doi.org/10.1016/j.neuropsychologia.2017.05.009>

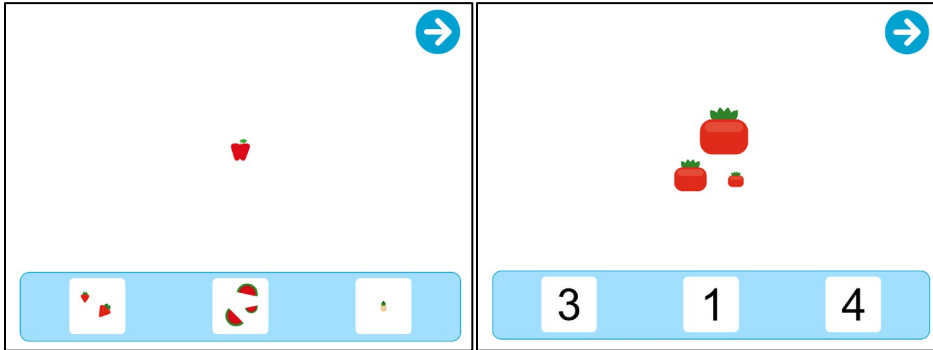
- Van Rinsveld, A., Schiltz, C., Brunner, M., Landerl, K., & Ugen, S. (2016). Solving arithmetic problems in first and second language: Does the language context matter? *Learning and Instruction, 42*, 72–82. <https://doi.org/10.1016/j.learninstruc.2016.01.003>
- Vanbinst, K., Ansari, D., Ghesquière, P., & Smedt, B. D. (2016). Symbolic Numerical Magnitude Processing Is as Important to Arithmetic as Phonological Awareness Is to Reading. *PLOS ONE, 11*(3), e0151045. <https://doi.org/10.1371/journal.pone.0151045>
- Verdine, B. N., Irwin, C. M., Golinkoff, R. M., & Hirsh-Pasek, K. (2014). Contributions of executive function and spatial skills to preschool mathematics achievement. *Journal of Experimental Child Psychology, 126*, 37–51. <https://doi.org/10.1016/j.jecp.2014.02.012>
- Vilenius-Tuohimaa, P. M., Aunola, K., & Nurmi, J.-E. (2008). The association between mathematical word problems and reading comprehension. *Educational Psychology, 28*(4), 409–426. <https://doi.org/10.1080/01443410701708228>
- von Aster, M. G., Bzufka, M. W., & Horn, R. R. (2009). *ZAREKI-K. Neuropsychologische Testbatterie für Zahlenverarbeitung und Rechnen bei Kindern: Kindergartenversion: Manual* (Harcourt).
- von Aster, M. G., & Shalev, R. S. (2007). Number development and developmental dyscalculia. *Developmental Medicine & Child Neurology, 49*(11), 868–873. <https://doi.org/10.1111/j.1469-8749.2007.00868.x>
- Vukovic, R. K., & Lesaux, N. K. (2013). The language of mathematics: Investigating the ways language counts for children’s mathematical development. *Journal of Experimental Child Psychology, 115*(2), 227–244. <https://doi.org/10.1016/j.jecp.2013.02.002>
- Wang, A. Y., Fuchs, L. S., & Fuchs, D. (2016). Cognitive and linguistic predictors of mathematical word problems with and without irrelevant information. *Learning and Individual Differences, 52*, 79–87. <https://doi.org/10.1016/j.lindif.2016.10.015>
- Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What’s Past Is Prologue: Relations Between Early Mathematics Knowledge and High School Achievement. *Educational Researcher, 43*(7), 352–360. <https://doi.org/10.3102/0013189X14553660>
- World Health Organization. (1992). The ICD-10 Classification of Mental and Behavioural Disorders. *International Classification, 10*, 1–267. [https://doi.org/10.1002/1520-6505\(2000\)9:5<201::AID-EVAN2>3.3.CO;2-P](https://doi.org/10.1002/1520-6505(2000)9:5<201::AID-EVAN2>3.3.CO;2-P)
- Zhang, X. (2016). Linking language, visual-spatial, and executive function skills to number competence in very young Chinese children. *Early Childhood Research Quarterly, 36*, 178–189. <https://doi.org/10.1016/j.ecresq.2015.12.010>
- Zuber, J., Pixner, S., Moeller, K., & Nuerk, H.-C. (2009). On the language specificity of basic number processing: Transcoding in a language with inversion and its relation to working memory capacity. *Journal of Experimental Child Psychology, 102*(1), 60–77. <https://doi.org/10.1016/j.jecp.2008.04.003>



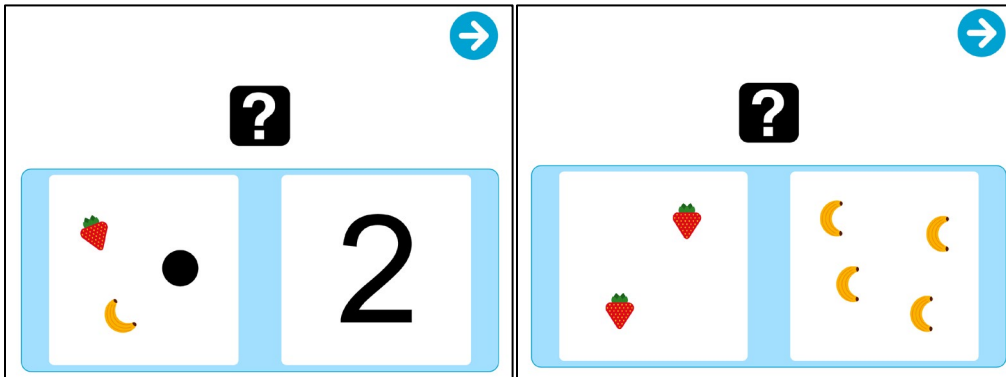
## Appendix

### NUMTEST Item examples

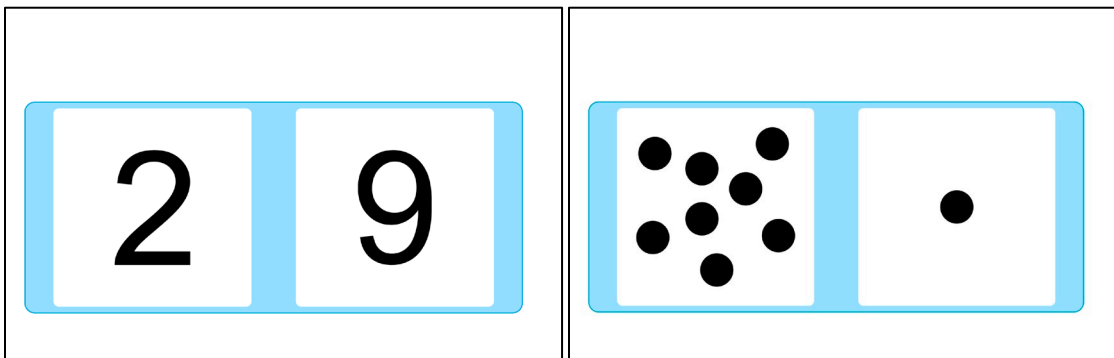
#### 1. Counting & correspondence









#### 2. Comparison (Version 1)



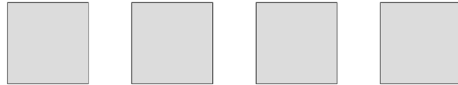
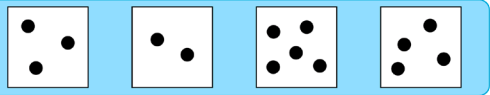


#### 3. Comparison (Version 2)






4. Ordering (Version 1)

|   |  |
|---|--|
| <br><br> | <br><br> |
|---|--|

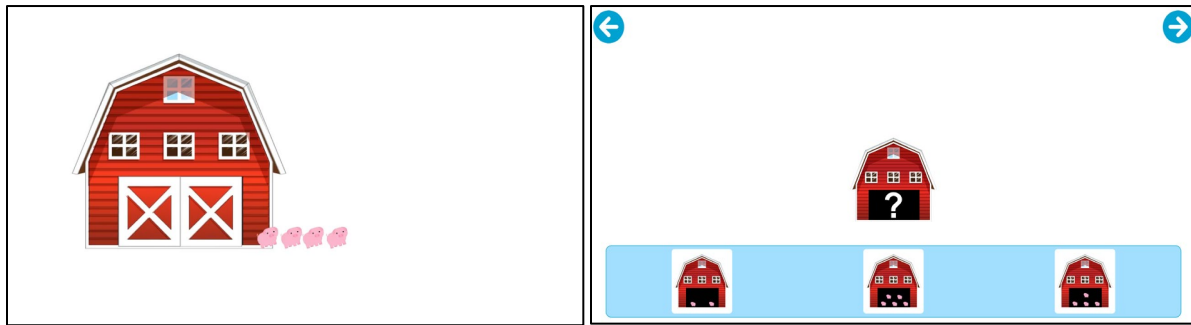
5. Ordering (Version 2)

|   |   |
|---|---|
| <br> | <br> |
|---|---|

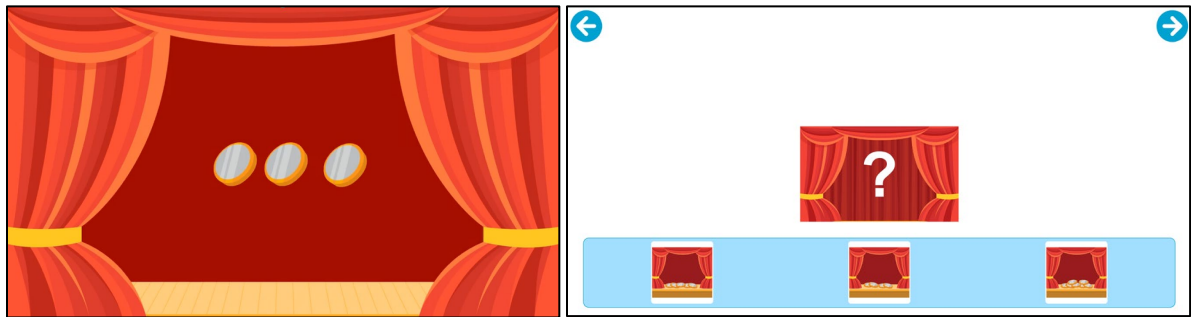
6. Non-symbolic addition

|   |   |
|---|---|
|  | <br> |
|---|---|

7. Non-symbolic subtraction



8. Crossmodal addition



## Acknowledgements

There are so many people to thank for their help in this project that I won't even try to list them all. So, here's a heartfelt thank you to all the people from COSA, LUCET and beyond, thank you for all you've done for me over the past five years. However, I have a few words for several people specifically as without them, there would be nobody reading this right now.

Alex, my dearest colleague, rival, friend, my brother from another mother, the Chap to my Chip. Where do I start... From the countless hours spent smoking and discussing politics, videogames and science in front of the MSH, to the many times you managed to dampen my frustrations with patience and humour. Never did disagreement feel as enjoyable as with you and your absence throughout the last year of my PhD was felt dearly. You are one of the major reasons that I will remember these years as some of the best in my life and for that I will forever be grateful. Thank you for existing, you truly are the very best there ever was! Also, Trump got impeached just in time, so you owe me 100 bucks in my book!

Carrie, thank you for being the best co-tenant of an office that one could dream of. Most people make me nervous by their mere presence, yours on the other hand always felt like an enrichment of the space we shared for so many years. If I could, I would put you in my pocket and carry (hehehe...) you with me to whatever office expects me in the future. As I can't, saying goodbye to you is the hardest thing. Also, I used your toilet paper many times without asking!

Claire, thank you for all your expertise and input over the years, especially during times in which I asked for a new model or a new visualisation seemingly every other day... Thank you also for your words of encouragement during the more difficult times of this endeavour and for the many times you showed me that the glass is indeed half-full, ideally with Gin Tonic!

Tanja, thank you so much for your guidance and expertise when it comes to developing psychometric tools. Your impact on the quality of my work deserves special praise and without it, I would probably still be erring around trying to write up my pilot studies.

Ricky, thank you for the many times you saved the day. Without you, the project would likely still be only a concept. You've repeatedly impressed me with analytical prowess, problem-solving skills and determination. I'm also suspecting that you're slowly poisoning the entire 3<sup>rd</sup> floor, but that's ok because your chocolate recipes sure are delicious!

Caroline, thank you for your continued support over the years, from my first initiation to child assessment that I completed with you, to the many times you gave crucial feedback both on the development of NUMTEST tasks and items as well as the resulting papers. Thank you so much for all you've done!

And finally, Christine. I am by now convinced that there exists exactly one supervisor in the world that could've managed to make me complete a PhD and that is you. They say that the difference between a boss and a leader is that the boss will order you to jump off the cliff while a leader will have people jumping without him asking. By that definition you are a true leader, one that manages to be respected by friends and opponents alike and for which people will happily jump as they know that they will be fine in your hands. Thank you for putting so much trust in me from beginning to end and throughout all headaches I might have caused. Your relentless optimism can pierce through the thickest layers of negativity and I will ask myself many times "*What would Christine do now?*". Thank you for everything.