CrossMark

Symbolic magnitude processing in elementary school children: A group administered paper-and-pencil measure (SYMP Test)

Carmen Brankaer[1] · Pol Ghesquière[1] · Bert De Smedt[1]

Abstract The ability to compare symbolic numerical magnitudes correlates with children's concurrent and future mathematics achievement. We developed and evaluated a quick timed paper-and-pencil measure that can easily be used, for example in large-scale research, in which children have to cross out the numerically larger of two Arabic one- and two-digit numbers (SYMP Test). We investigated performance on this test in 1,588 primary school children (Grades 1–6) and examined in each grade its associations with mathematics achievement. The SYMP Test had satisfactory test-retest reliability. The SYMP Test showed significant and stable correlations with mathematics achievement for both one-digit and two-digit comparison, across all grades. This replicates the previously observed association between symbolic numerical magnitude processing and mathematics achievement, but extends it by showing that the association is observed in all grades in primary education and occurs for single- as well as multi-digit processing. Children with mathematical learning difficulties performed significantly lower on one-digit comparison and two-digit comparison in all grades. This all suggests satisfactory construct and criterion-related validity of the SYMP Test, which can be used in research, when performing large-scale (intervention) studies, and by practitioners, as screening measure to identify children at risk for mathematical difficulties or dyscalculia.

Keywords Symbolic magnitude processing · Number sense · Mathematical difficulties · Screening · Reliability and validity · Dyscalculia

✉ Bert De Smedt
Bert.DeSmedt@ppw.kuleuven.be

[1] Faculty of Psychology and Educational Sciences, Parenting and Special Education Research Unit, University of Leuven, Leopold Vanderkelenstraat 32, box 3765, 3000 Leuven, Belgium

Introduction

Number knowledge and mathematical skills are important in everyday life, for example when paying for a purchase, reading timetables to catch a train or bus, or interpreting recipes and measuring ingredients for cooking. Moreover, mathematical proficiency is associated with greater labor market success (Chiswick, Lee, & Miller, 2003), better medical decision making (Reyna, Nelson, Han, & Dieckmann, 2009) and lower mortgage default rates (Gerardi, Goette, & Meier, 2013). It is also a major target in primary schools (Kilpatrick, Swafford, & Findell, 2001). However, around 15–25 % of children and adults experience difficulties with the development of mathematics and 5–7 % of them even have specific mathematical learning disabilities or dyscalculia (American Psychiatric Association, 2013; Butterworth, Varma, & Laurrilard, 2011; Geary, 2011). These mathematical difficulties might have far-reaching consequences for the future school career of children and the quality of their daily life.

Against this background, it is crucial to detect and support children who are at risk for developing mathematical difficulties at an early age. The question arises which competencies should be assessed when designing screening measures to identify these at-risk children. Research has pointed to the importance of numerical magnitude processing, or people's elementary intuitions about numerical magnitudes, for the development of mathematics achievement as children's understanding of numerical magnitudes correlates with (e.g., Holloway & Ansari, 2009) and predicts (e.g., De Smedt, Verschaffel, & Ghesquière, 2009; Halberda, Mazzocco, & Feigenson, 2008; Mazzocco, Feigenson, & Halberda, 2011; Vanbinst, Ghesquière & De Smedt, 2015) individual differences in mathematics achievement (De Smedt, Noël, Gilmore, & Ansari, 2013, for a narrative review; Schneider et al., 2016, for a meta-analysis).

A typical and well-established paradigm to examine numerical magnitude processing is the numerical magnitude comparison task (Sekuler & Mierkiewicz, 1977). In this task, children have to indicate the numerically larger of two numerical magnitudes. These magnitudes can be presented both in a symbolic (Arabic digits or number words) or non-symbolic (dot arrays or sequences of sounds) format (De Smedt et al., 2013). When individuals compare numerical magnitudes, the so-called distance effect occurs (Moyer & Landauer, 1967): They are faster and more accurate at judging which of two magnitudes is numerically larger when the numerical distance between both magnitudes is relatively large (e.g., 1 vs. 9) than when the distance is small (e.g., 8 vs. 9). This distance effect is assumed to originate from the approximate nature of numerical magnitude representations, with more representational overlap for magnitudes that are closer to each other than for magnitudes that are further apart (see Noël, Rousselle, & Mussolin, 2005). Additionally, individuals' performance on the numerical magnitude comparison task is also influenced by the size of the magnitudes that are presented (Moyer & Landauer, 1967): They will respond faster and more accurate when magnitude pairs with a smaller magnitude are used than when pairs with a larger magnitude are presented, even when the numerical distance in both magnitude pairs is held constant (e.g. 4 vs. 3 is solved faster and more accurate than 9 vs. 8, even though both number pairs have a distance of 1).

It has been widely documented that children's performance on numerical magnitude comparison tasks is associated with their performance on mathematics achievement tests (De Smedt et al., 2013 for a narrative review). Moreover, children with mathematical learning difficulties or dyscalculia are impaired in their ability to compare numerical magnitudes (De Smedt & Gilmore, 2011; Landerl, Bevan, Butterworth, 2004; Noël & Rousselle, 2011; Vanbinst, Ghesquière, & De Smedt, 2014). This association between numerical magnitude comparison deficits and mathematical learning difficulties seems to be independent of intellectual ability, as Brankaer, Ghesquière, and De Smedt (2014) found that children with discrepant (low math scores, average IQ) and non-discrepant (low math scores, low IQ) mathematical difficulties have highly similar impairments in numerical magnitude processing, despite differences in intellectual ability.

The majority of studies examined children's performance on non-symbolic magnitude comparison tasks and its association to mathematics achievement. Inconsistent results have been reported as some studies found significant associations between children's ability to compare dot arrays and mathematics achievement (Halberda et al., 2008; Mazzocco et al., 2011), while others did not (e.g., De Smedt & Gilmore, 2011; Holloway & Ansari, 2009; Vanbinst, Ghesquière, & De Smedt, 2012). Recently, two meta-analyses have demonstrated that non-symbolic magnitude processing was significantly associated with mathematics achievement, although the

correlations tended to be weak ($r = .20$ in Chen & Li, 2014; $r = .22$ in Fazio, Bailey, Thompson, & Siegler, 2014), but unfortunately these analyses did not consider measures of symbolic magnitude processing.

Findings on the symbolic magnitude comparison task have been, in contrast, more robust as most studies showed that the better children are in determining which of two Arabic digits is the largest, the higher their concurrent and future scores on mathematics achievement tests (e.g., De Smedt et al., 2009; Kolkman, Kroesbergen, & Leseman, 2013; Vanbinst et al., 2012). Summarizing the available evidence, De Smedt et al. (2013) therefore argued that symbolic magnitude processing might be a more robust predictor of individual differences in mathematics achievement. Such descriptive comparisons could be, however, misleading as they do not take into account the effect sizes or sample sizes under investigation. In an attempt to resolve this issue, Schneider et al. (2016) conducted a meta-analysis and statistically contrasted the effect sizes of non-symbolic as well as symbolic numerical magnitude processing as predictors of mathematics achievement. Their data revealed that the association between symbolic numerical magnitude processing and mathematics achievement ($r = 0.302$, 95 % CI = [0.243, 0.361]), was significantly larger than the association with non-symbolic numerical magnitude processing ($r = 0.241$, 95 % CI = [0.198, 0.284]). We therefore focused in the design of our paper-and-pencil measure on symbolic rather than non-symbolic magnitude processing.

It is important to note that previous studies on symbolic magnitude processing showed that particularly reaction time measures were related to mathematics achievement and that differences between typically developing children and children with mathematical difficulties were most prominent in reaction times rather than accuracy (see De Smedt et al., 2013; Schneider et al., 2016). A possible explanation for these findings is that reaction time data can reveal subtle yet important differences that cannot be uncovered by looking at accuracy data alone (Berch, 2005).

Nearly all of the existing studies used computerized tasks to measure numerical magnitude comparison and these tasks are time-consuming as they often depend on one-on-one administration (De Smedt et al., 2013). In view of the association between numerical magnitude comparison and mathematics achievement, the question arises whether large-scale measures, which also take into account children's speed in solving number comparison, could be designed to evaluate children's symbolic comparison skills in a quick and classroom friendly manner.

To the best of our knowledge, only two studies have been conducted with group-administered numerical magnitude processing tasks. Durand, Hulme, Larkin, and Snowling (2005) tested 162 7- to 10-year-olds with a paper-and-pencil test in which children had to judge which of two Arabic digits was numerically larger. Digits varied from 3 to 9 and the numerical

distance between the digit pairs was small (1 or 2). Twenty-eight items were presented and children were given 30 s to select the larger magnitude per pair. Children's performance on this paper-and-pencil measure was associated with arithmetic ability: children who correctly solved more magnitude comparison items had better arithmetic skills. Nosworthy, Bugden, Archibald, Evans, and Ansari (2013) recently developed a paper-and-pencil tool to assess children's ability to compare symbolic and non-symbolic numerical magnitudes. In this study, 160 children aged 6–9 years completed a symbolic (Arabic digits) and non-symbolic version (dot arrays) of the magnitude comparison task, with magnitudes ranging from 1 to 9. The numerical distance between the magnitudes varied from 1 to 8. Fifty-six items were presented for both the symbolic and non-symbolic version of the comparison task and children had 1 min per task to cross out the larger of the two magnitudes. Children's performance on the symbolic, but not on the non-symbolic comparison task, was uniquely correlated with their arithmetic skills, even when individual differences in working memory and intelligence were additionally controlled for, which strengthens the notion that especially symbolic magnitude processing is associated with individual differences in mathematics achievement.

One limitation of the studies of Durand et al. (2005) and Nosworthy et al. (2013) is that the authors did not report on the psychometric aspects of their paper-and-pencil measure. It is, however, important to establish whether a measure is reliable and valid before implementing it in large-scale research and educational practice. Another limitation of Durand et al. (2005) and Nosworthy et al. (2013) is that they only focused on one-digit comparison. Indeed, children become more accurate and faster in numerical magnitude comparison throughout development (Holloway & Ansari, 2009; Landerl & Kölle, 2009) and ceiling effects are often observed for older children in one-digit comparison tasks (Holloway & Ansari, 2009). This suggests that the variability in numerical magnitude comparison cannot be adequately captured when only using single-digit numbers, particularly in older children. Therefore, magnitude comparison tasks with multi-digit numerals should be included when assessing children's numerical magnitude processing.

Studies have examined numerical magnitude processing for two-digit numbers and found that children were slower and less accurate when solving two-digit magnitude comparison tasks than when solving one-digit magnitude comparison tasks (e.g., Landerl, Fussenegger, Moll, & Willburger, 2009). Participants' ability to process two-digit numbers was also related to individual differences in mathematics achievement, as children with mathematical learning difficulties performed significantly lower on two-digit magnitude comparison tasks than their typically developing peers (Andersson & Östergren, 2012; Landerl et al., 2009; Landerl & Kölle, 2009), yet it should be pointed out that the number of studies that have

investigated the association between multi-digit comparison and mathematics achievement is very small compared to the flurry of studies with one-digit comparison.

It is important to emphasize that the processing of multi-digit numbers might be different from single-digit numbers. Two processing models for multi-digit numbers have been proposed. According to the holistic model, two-digit numbers are processed holistically, i.e., as a uniform unit (Dehaene, Dupoux, & Mehler, 1990; Reynvoet & Brysbaert, 1999). The compositional model, in contrast, states that the decade-digit (tens place value) and unit-digit (one's place value) of a number are processed separately (Nuerk, Weger, & Willmes, 2001). Support for this compositional view comes from the compatibility effect or the observation that children are faster when comparing compatible number pairs (the decade-magnitude comparison and the unit-magnitude comparison lead to the same decision, for example when comparing 32 to 45) than when comparing incompatible number pairs (the decade-magnitude comparison and the unit-magnitude comparison lead to a different decision, for example when comparing 38 to 45). This all illustrates that different effects might come into play when children need to process one-digit vs. two-digit magnitudes.

The present study

The main goal of the present study was to develop and evaluate a measure to investigate children's numerical magnitude processing skills that can be used for screening at-risk children or in large-scale research. In view of the meta-analytic data by Schneider et al. (2016), who showed that symbolic magnitude comparison was a more robust predictor of mathematics achievement, we decided to only use magnitude comparison tasks with Arabic digits. The decision to use a paper-and-pencil measure was based on several reasons. First, such a test can be easily and quickly assessed in large groups. Further, the costs for paper-and-pencil tools are much lower compared to computerized screening measures and, from a practical point of view, less instruction is required for teachers to administer and score this type of test (see also Nosworthy et al., 2013).

Extending the studies of Nosworthy et al. (2013) and Durand et al. (2005), the present study included not only one- but also two-digit magnitude comparison tasks in the paper-and-pencil test of SYmbolic Magnitude Processing (SYMP Test). We also investigated symbolic number comparison, and its association with mathematics achievement in a much larger sample compared to previous studies, including children from all grades of elementary school (1–6) with a considerable number of children per grade, which also allowed us to investigate associations within each grade and compare associations across grades of primary school. We predicted that children would improve in their symbolic number comparison abilities across primary school and that the

association between symbolic number comparison and mathematics achievement would remain significant across all grades.

Reliability and validity of the SYMP Test were investigated in Grades 1–6 of primary school. Test-retest reliability was investigated by calculating Pearson correlations coefficients between children's performance on the SYMP Test at two different time points. Test-retest correlations of at least .70 are needed to indicate adequate or satisfactory test reliability (Hunsley & Mash, 2008).

Construct validity was investigated in several ways. Pearson correlation coefficients between (1) the one-digit and two-digit subtest of the SYMP Test (*convergent validity*), (2) the SYMP Test and a computerized version of this test (*convergent validity*), and (3) the SYMP Test and standardized achievement tests for mathematics (*convergent validity*) and spelling (*discriminant validity*) were calculated. These standardized tests were curriculum-based achievement tests that covered the various skills that were taught according to the mathematics and spelling curriculum. We expected high and significant correlations between (1) the one-digit and two-digit subtest, (2) the SYMP Test and the computerized version of this test, and (3) the SYMP Test and the standardized achievement test for mathematics. Additionally, we expected lower correlations between the SYMP Test and the standardized achievement test for spelling because both measures are supposed to measure different constructs. Following Cohen (1988), correlation coefficients of .10 were considered as low, coefficients of .30 as moderate, and coefficients of .50 as high.

Criterion-related validity was examined by comparing the performance of children with mathematical learning difficulties (MLD) and typically developing children on the SYMP Test. In accordance with contemporary research on children with MLD (e.g., Geary, 2011; Geary et al., 2007; Mazzocco et al., 2011), children with MLD had to perform below the 10th percentile on a standardized mathematics achievement test in order to be classified as having MLD; and typically achieving (TA) children had to perform above the 35th percentile. We expected that children with MLD would perform significantly lower on the test than their typically developing peers, which would indicate that the test has satisfactory criterion-related validity.

Method

Participants

Participants were 1,588 children in Grades 1–6 from 10 elementary schools in Flanders, Belgium. Parental consent was obtained for all children and they all completed the first assessment of the SYMP Test. Participants came from a variety of socio-economic backgrounds. Table 1 shows the descriptive statistics of these children. Grades did not differ in the number of boys and girls, $\chi^2(5, N = 1,588) = 4.30$, $p = .51$.

Three to 5 weeks after the first assessment of the SYMP Test, retesting took place in a large group of children ($n = 1,425$) to evaluate test-retest reliability. At this second assessment, 59 students were missing because they were absent at school (e.g., due to illness) and this pattern was not systematic across classes with on average 1.40 children missing per class ($SD = 1.41$, range 0–5 students); 100 students were missing because their teacher did not want to participate anymore at the second time point (e.g., not interested, not fitting the class schedule). This comprised six classes: one class in grade 1, two classes in grade 2, one class in grade 4, one class in grade 5, and one class in grade 6. Students who missed at the second time point were compared to those who did not miss at the second time point on the SYMP Test that was administered at the first time point. This analysis revealed that there was no difference between these two groups on the one-digit subtest ($F(1,1586) = 0.81$, $p = .37$) and two-digit subtest ($F(1,1322) = 0.05$, $p = .83$). The two groups also did not differ in their mathematics ($F(1,1166) = 0.03$, $p = .86$) and spelling achievement ($F(1,1046) = 1.38$, $p = .24$) and in the number of children with MLD (Fisher's exact test: $p = .87$).

Additionally, from the initial sample of participants, 355 children were randomly selected for the individual assessment of the computerized symbolic magnitude comparison tasks. These computer tasks were administered to examine convergent validity. Because all participating schools used the standardized Flemish Student Monitoring System (Dudal, 2000a, b) to evaluate children's academic achievement, we also obtained the scores of 1,168 children on a standardized achievement test for mathematics and of 1,048 children on a standardized achievement test for spelling to assess convergent and discriminant validity.

Table 1 Descriptive statistics of the sample (n = 1588) and of children's performance on the one-digit and two-digit subtests of the SYMP Test

Grade	Sex		Age (years)			one-digit [a]		two-digit[a]	
	Boys	Girls	M	SD	Range	M	SD	M	SD
1	138	126	6.63	0.56	5.7-8.6	15.02	5.66	-	-
2	125	150	7.62	0.53	6.7-9.2	22.00	4.37	11.67	4.59
3	133	115	8.63	0.53	7.5-10.6	25.56	4.78	16.81	3.96
4	140	133	9.69	0.57	8.7-12.0	28.31	5.03	19.85	4.39
5	131	125	10.67	0.54	9.0-12.2	32.30	4.90	23.02	4.17
6	135	137	11.70	0.56	10.6-14.0	34.50	5.20	25.83	4.36

[a] Number of correctly solved items. The two-digit subtest was not administered in Grade 1. Grade 1 is the first formal year of schooling

Measures

Paper-and-pencil tests

Paper-and-pencil test of SYmbolic Magnitude Processing (SYMP Test) The SYMP Test consisted of two numerical magnitude comparison tasks: a one-digit subtest with digits between 1 and 9 and a two-digit subtest with digits ranging from 11 to 99 (Fig. 1). This paper-and-pencil test was constructed in three phases and evaluated in several pilot studies to decide on the content of the test items and the optimal test duration.

The final versions of the one-digit and two-digit subtest each consisted of 60 digit pairs, presented in four columns of 15 pairs (Verdana font, size 12). For the one-digit subtest, the distance between both digits was 1 on half of the items and 3 or 4 on the other half of the trials. All possible combinations with these distances were included. The number pairs were randomly presented, while controlling for several factors: (1) the side of the correct answer and small vs. large distances were counterbalanced in each column, (2) different numbers were used in subsequent or neighboring number pairs, (3) no more than three consecutive correct answers on the same side (left/right) were presented and (4) no identical or inverse number pairs (e.g., 3–1 vs. 1–3) were presented in the same column or row. In the two-digit subtest, the distance between both digits varied between the small distances 2 to 6 on one half of the trials, to larger distances ranging from 12 to 16 on the other half of the trials. Each distance was used six times, resulting in 60 test items. Each distance was equally represented in each column and the position of correct answers (left/right) was counterbalanced. No more than three consecutive correct answers on the same side were presented and different decade-magnitudes were used in subsequent or neighboring number pairs.

During the test, participants were asked to cross out the larger of the two digits. They were given 30 s to solve as many items as possible. To ensure that all children understood the task, four practice trials were included in both subtests. The time limit was included in order to be able to assess children's fluency of numerical magnitude comparison and to avoid ceiling effects. The decision for a time constraint of 30 s was based on various pilot studies. In the first pilot version in Grades 1, 3, and 6 ($n = 16$), we used no time limit but recorded the time that children needed to solve all comparison items. Based on children's response times on the one-digit subtest (Grade 1: 170 s; Grade 3: 62 s; Grade 6: 51 s) and two-digit subtest (Grade 3: 78 s and Grade 6: 64 s), a time limit of 45 s was implemented. In the second pilot study, this time constraint was evaluated in 93 children from Grades 1 to 6. There was a ceiling effect for the one-digit subtest in Grades 5 and 6, as 57 % of the children in Grade 5 and 14 % of the children in Grade 6 could solve all 60 items within 45 s. Therefore, the time limit of both subtests was set at 30 s. In the third and final pilot study, the paper-and-pencil task was administered in 61 children from Grades 2, 4, and 5 to evaluate this new time limit. No ceiling effects for the one-digit and two-digit subtest were observed, suggesting that 30 s was an adequate time restraint.

Paper-and-pencil motor speed control task Because children's performance on the SYMP Test could be influenced by general response speed, we also developed a paper-and-pencil task to control for motor speed in our analyses (see Fig. 2). In this task, children were presented with 60 pairs of figures (circle, square, triangle, heart, star, or moon). One figure of the pair was colored black and the other figure was colored white. Children were instructed to cross out the black figures as fast as possible. Figure pairs were randomly presented in four columns of 15 pairs, while controlling for a number of factors: (1) the side of the correct answer was counterbalanced in each column, (2) maximum three consecutive correct answers on the same side were presented, (3) no identical or inverse figure pairs were presented in the same
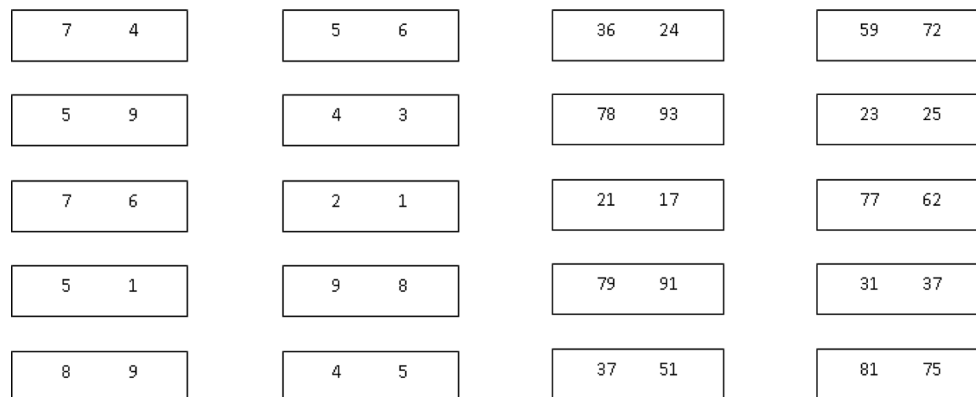

Fig. 1 SYmbolic Magnitude Processing (SYMP) Test. Examples of items on the one-digit subtest (left) and two-digit subtest (right). Children were instructed to cross out the numerically larger of the two digits
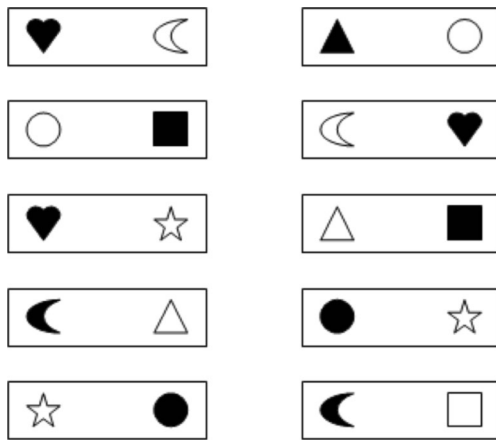
Fig. 2 Examples of items on the motor speed task. Children were instructed to cross out the black figure in the figure pair

column or row, (4) different figures were used in subsequent or neighboring items, and (5) each figure was presented five times in each column. Based on the pilot studies, children were given 20 s to answer as many items as possible. Two practice trials were presented to ensure that all participants understood the task. Test-retest reliability of this test was .86.

Computerized tasks

Two experimental tasks were used to individually assess the different subtests of the SYMP Test: a one-digit computer task and a two-digit computer task. These tasks were presented using the E-prime 2.0 software (Schneider, Eschmann, & Zuccolotto, 2002) and were administered on a laptop with a 15-in. screen. Children had to indicate the numerically larger of two simultaneously presented Arabic digits, one displayed on the left and one displayed on the right side of the computer screen. In both tasks, children were instructed to answer as quickly and accurately as possible. Stimuli were the first 30 items of the one-digit and two-digit subtest of the paper-and-pencil test. Each trial started with a 200 ms fixation cross in the center of the screen. After 1,000 ms the stimuli appeared and remained visible until response. Children had to respond by pressing a key on a computer keyboard that was put in front of the laptop and was connected to it. The left response key, labeled with a blue sticker, was 'D'; the right response key, labeled with a yellow sticker, was 'K'. Both the one-digit and two-digit computer task were preceded by the same four practice trials as were used in the paper-and-pencil subtests. This was done to familiarize the child with the key assignments. There was no time limit, but answers and reaction times were recorded by the laptop. These computerized tasks were highly similar to the ones that were used in previous studies on the association between numerical magnitude comparison and mathematics achievement (see De Smedt et al., 2013 for an overview).

Standardized tests

Mathematics General mathematics achievement was assessed using a curriculum-based standardized achievement test for mathematics from the Flemish Student Monitoring System (Dudal, 2000a). This untimed test consisted of 60 items covering number knowledge, understanding of operations, (simple) arithmetic, word problem solving, measurement and geometry. The content of this standardized test differed between the different grades of primary school and focused on what children should have learned during formal mathematics education in the months preceding to the test. At the start of Grade 1, for example, children's counting skills from 1 to 10 are evaluated, while at the end of Grade 2 children are instructed to solve addition and subtraction problems up to 100. The score on this standardized achievement test for mathematics was the number of correctly solved problems (maximum = 60). Standardization sample data were available for all grades of primary school. Reliability of these tests was between .86 and .91

Spelling Children's spelling skills were administered using a curriculum-based standardized achievement test for spelling from the Flemish Student Monitoring System (Dudal, 2000b). In this test, children are instructed to write letters, words and sentences from dictation. Analogous to the test for mathematics, the content of the spelling test differed between the different grades of primary school – ranging from simple CVC-words in Grade 1 to complex rule-based words and sentences in the upper grades – and focuses on what children should have learned according to their grade curriculum. The test consisted of 60 items, with 1 point for each correct answer. Standardization sample data were available for all grades of primary school. Reliability of these tests was between .85 and .91

Procedure

The paper-and-pencil tests were collectively administered in children's classroom during regular school hours. The one-digit, two-digit, and motor tests were administered on three different sheets of paper (one sheet for each test). The front of the sheet contained space for children to write their name as well as the four practice trials, which were made together with the test administrator. Children were given further instructions about the test, but were not allowed to turn the page until the test administrator told them to do so. The back of each sheet contained the final 60 test items. When the start sign was given by the administrator, children turned the page and started to work on the test. When the stop sign was given, children had to drop their pens. The administrator kept time by means of a stopwatch. This group-based

session took about 10–15 min. The data were collected by three administrators that were carefully trained before data collection. The computerized symbolic magnitude comparison tasks were assessed individually in a quiet room and these computerized versions were always administered after the paper-and-pencil tests. This individual session took approximately 10–15 min. Children's scores on the standardized Flemish Student Monitoring system (Dudal, 2000a, b) for mathematics and spelling were obtained from their school records.

Results

Descriptive analyses

Table 1 shows children's performance on the SYMP Test. The two-digit subtest was not assessed in Grade 1, because children in this grade did not yet receive enough instruction in this number domain. In general, children's maximum score on the one-digit subtest was 52 and on the two-digit subtest 45, which indicates that there were no ceiling effects. There were two first graders that were not able to solve any item of the one-digit subtest. No floor effects for the two-digit subtest were observed, as all children were able to solve at least two items.

To evaluate children's development on the SYMP Test throughout the school years, grade differences were evaluated using one-way ANOVAs. Scores on the one-digit subtest varied significantly between the different grades, $F(5,1582) = 538.55$, $p < .01$, $\eta_p^2 = .63$, and Tukey post-hoc t-tests revealed that all grades differed significantly from each other ($ps < .01$; Grade 1 < Grade 2 < Grade 3 < Grade 4 < Grade 5 < Grade 6). A similar pattern of results was obtained for the two-digit subtest, $F(4,1319) = 441.54$, $p < .01$, $\eta_p^2 = .57$, and the scores from all grades differed significantly from each other ($ps < .01$; Grade 2 < Grade 3 < Grade 4 < Grade 5 < Grade 6).

Reliability

Pearson correlation coefficients between children's test and retest scores on the SYMP Test were calculated to examine test-retest reliability. All these correlations were significant at the .01 level for both subtests ($rs > .62$; Table 2). We also analyzed children's overall score on the SYMP Test by taking the total number of correctly solved items summed for the two subtests. This yielded higher test-retest correlations for these overall scores in all grades (rs between .72 and .86, $ps < .01$), which suggests that both subtests should be administered at the same time when assessing symbolic magnitude processing.

Validity

Firstly, the associations between the one-digit and two-digit subtest of the SYMP Test were examined to assess convergent validity. Both subtests were significantly correlated with each other, with correlation coefficients ranging from .57 to .66 (Grade 2: $r = .64$, Grade 3: $r = .57$, Grade 4: $r = .66$, Grade 5: $r = .63$, Grade 6: $r = .57$, all $ps < .01$) in the different grades. These correlations remained significant after controlling for children's performance on the control task for motor speed (Grade 2: $r = .54$, Grade 3: $r = .42$, Grade 4: $r = .55$, Grade 5: $r = .51$, Grade 6: $r = .43$, all $ps < .01$).

Secondly, convergent validity was assessed by investigating the relationship between the SYMP Test and the computerized comparison tasks. We only analyzed children's reaction times on these computerized tasks because accuracy was very

Table 2 Test-retest scores and Pearson correlation coefficients for the one-digit and two-digit subtest of the SYMP Test

Grade	N	One-digit subtest[a]							Two-digit subtest[a]					
		Test		Retest		r			Test		Retest		r
		M	SD	M	SD				M	SD	M	SD	
1	238	14.66	5.48	18.64	5.22	.70**			-	-	-	-	-
2	222	21.72	4.28	25.28	5.13	.62**			11.22	4.56	13.09	5.31	.73**
3	242	25.56	4.74	30.48	6.19	.64**			16.80	3.98	18.49	4.66	.71**
4	240	28.03	4.94	33.14	5.46	.66**			19.62	4.20	21.65	4.27	.77**
5	235	32.25	4.95	37.16	5.78	.77**			22.93	4.15	25.44	4.46	.77**
6	248	34.45	5.18	41.13	6.31	.70**			25.73	4.18	28.10	4.75	.65**
Total[b]	1,425	26.21	8.27	31.09	9.38	.71**			19.41	6.52	21.5	7.02	.74**

** $p < .01$

[a] Number of correctly solved items. The two-digit subtest was not administered in Grade 1

[b] Test-retest correlation coefficients across all grades were controlled for children's chronological age

Table 3 Descriptive statistics and Pearson correlation coefficients between the SYMP Test and the computerized version of this test

Grade	N	One-digit						Two-digit					
		Paper-and-pencil[a]		Computer[b]		r[c]		Paper-and-pencil[a]		Computer[b]		r[c]	
		M	SD	M	SD			M	SD	M	SD		
1	60	14.85	5.99	1395.62	373.75	−.66**		–	–	–	–	–	
2	59	23.08	4.53	1015.35	282.62	−.58**		12.47	4.72	2007.75	811.82	−.68**	
3	59	26.39	4.27	827.75	184.61	−.60**		16.97	4.04	1433.94	348.76	−.66**	
4	57	28.21	4.60	710.18	116.33	−.45**		19.88	4.26	1194.00	223.50	−.58**	
5	60	32.28	4.80	627.66	122.55	−.57**		22.75	4.04	1002.01	217.50	−.66**	
6	60	34.00	5.09	592.13	106.58	−.59**		25.27	3.63	936.76	139.39	−.66**	
Total[d]	355	26.46	8.03	862.39	355.29	−.57**		19.49	6.10	1313.37	572.79	−.56**	

** $p < .01$

[a] Number of correctly solved items on the paper-and-pencil test

[b] Mean reaction time (ms) on the computerized subtests, based on correct responses only

[c] Correlation coefficients between the paper-and-pencil test and the computerized version of this test. The two-digit subtest was not administered in Grade 1

[d] Correlation coefficients across all grades were controlled for children's chronological age

high and at ceiling (one-digit subtest: 96 % and two-digit subtest: 91 %). These reaction times were based on correct responses only. All correlation coefficients (Table 3) were negative, demonstrating that the more items children could solve on the paper-and-pencil task, the faster they responded on the computerized task. For both the one-digit and two-digit subtests, correlation coefficients in all grades were significant at the .01 level. These correlations all remained significant when the motor speed task was additionally controlled for.

Thirdly, we investigated the construct validity of the SYMP Test by looking at the associations between this test and the standardized tests for mathematics and spelling. It is important to note that these tests for mathematics and spelling achievement were curriculum-based, implying that the test content and standardization data differed from grade to grade. Therefore, children's raw scores on the standardized tests for mathematics and spelling and the paper-and-pencil test were transformed to standardized z-scores, using the standardization norms for each standardized test, to facilitate the comparison between the different grades. As displayed in Table 4, significant correlations between both subtests of the SYMP Test and the standardized test for mathematics were found in all grades, demonstrating that children who scored higher on the paper-and-pencil test had higher scores on the standardized test for mathematics. These associations all remained significant when the motor speed task was additionally controlled for.

The associations between symbolic comparison and mathematics achievement seemed to decrease when children became older. To test this assumption statistically, correlation coefficients were transformed into Fisher's z statistics and were compared by means of a z test. For the one-digit subtest, the correlation for the Grade 1 children differed significantly from that for Grade 4 children ($z = 2.22$, $p = .03$), Grade 5 children ($z = 2.40$, $p = .02$) and Grade 6 children ($z = 2.38$, $p = .02$). The other correlation coefficients did not differ significantly ($zs < 1.86$, $ps > .05$). For the two-digit subtest, no significant differences between the correlations in the different grades were found ($zs < 1.81$, $ps > .05$).

To evaluate the discriminant validity of the SYMP Test, we examined the associations between this test and the standardized achievement test for spelling. No significant correlations between children's scores on both measures were observed, except for the one-digit subtest in Grade 1 and 2 and the two-digit subtest in Grade 3 (see Table 4). We subsequently tested whether these correlations between the SYMP Test and spelling differed significantly from the correlations between the SYMP Test and the standardized test for mathematics. For this analysis, we only included children for which data on both standardized achievement tests was available ($n = 942$). In this subsample, correlation coefficients between the SYMP Test and the standardized test for mathematics were moderate and significant across all grades (one-digit $r = .28$; two-digit $r = .30$; $ps < .01$). The correlation coefficients between the SYMP Test and the standardized test for spelling were rather small across the different grades (one-digit $r = .10$; two-digit $r = .07$; $ps < .05$). A William-Steiger test showed that the associations between the paper-and-pencil test and mathematics achievement were significantly stronger than the associations between the paper-and-pencil test and spelling achievement for both the one-digit ($z = 3.34$, $p < .01$) and the two-digit ($z = 5.90$, $p < .01$) subtests.

Finally, the criterion-related validity of the SYMP Test was evaluated by comparing children with mathematical learning

Table 4 Pearson correlation coefficients between the z-scores on the SYMP Test and the z-scores on the standardized tests for mathematics and spelling

Grade	Standardized test Mathematics			Standardized test Spelling		
	N	r one-digit	r two-digit	N	r one-digit	r two-digit
1	207	.39**	-	126	.21*	-
2	213	.34**	.38**	211	.18**	.06
3	180	.22**	.40**	167	.05	.16*
4	207	.19**	.25**	193	.12	.09
5	169	.16*	.27**	165	.07	.07
6	192	.17*	.23**	186	.14	.04
Total	1168	.25**	.30**	1048	.10**	.08*

* $p < .05$

** $p < .01$

The two-digit subtest was not administered in Grade 1

difficulties (MLD; performance below the 10th percentile on the standardized achievement test for mathematics) and typically achieving children (TA; performance above the 35th percentile on the standardized achievement test for mathematics). The performance of both groups of children on the SYMP Test is reported in Table 5. In all grades, children in the TA-group performed significantly better on the one-digit subtest than children in the MLD-group (Fs > 5.23, ps < .05), with exception for Grade 6 where no group differences were observed ($F(1,196) = 0.78$, $p = .38$). Similar results were obtained for the two-digit subtest, as children in the TA-group

performed significantly better than children in the MLD-group (Fs > 4.68, ps < .05), except for Grade 6 ($F(1,196) = 0.26$, $p = .61$). Because these differences on the SYMP Test might be explained by group differences in processing speed, the abovementioned analyses were repeated with children's performance on the motor speed task as a covariate. Findings revealed that children in the TA-group still performed significantly better than children in the MLD-group on both the one-digit (Fs > 8.80, ps < .01) and two-digit subtest (Fs >7.93, ps < .01), except for Grade 6 (one-digit: $F(1,195) = 1.84$, $p = .18$; two-digit: $F(1,195) = 0.59$, $p = .44$).

Discussion

Children's ability to compare symbolic magnitudes has been identified as an important predictor of their mathematical development and children with mathematical learning difficulties or dyscalculia are impaired in this ability (De Smedt et al., 2013; Schneider et al., 2016). The aim of the present study was to develop a reliable and valid paper-and-pencil measure that can be primarily used to detect children who are at risk to develop mathematical difficulties or dyscalculia by assessing their symbolic magnitude comparison skills (SYMP Test). It also allows for the quick assessment of children's symbolic magnitude knowledge in large-scale classroom-based studies on other more complex aspects of children's mathematical knowledge, such as rational number understanding (e.g., Van Hoof, Verschaffel, & Van Dooren, 2015), to which

Table 5 Descriptive statistics of the performance of typically achieving children (> pc 35) and children with mathematical learning difficulties (< pc 10) on the SYMP Test and on the control task for motor speed

Grade	Group	N	One-digit subtest[a]		Two-digit subtest[a]		Motor speed[a]	
			M	SD	M	SD	M	SD
1	TA	163	16.73	5.18	-	-	22.89	7.12
	MLD	20	10.05	6.29	-	-	20.70	8.33
2	TA	181	22.56	4.11	12.46	4.57	29.65	6.20
	MLD	21	19.76	5.51	7.81	3.06	30.67	9.23
3	TA	158	26.01	4.82	17.39	3.61	34.34	7.53
	MLD	22	22.41	5.47	13.18	3.78	31.95	7.96
4	TA	155	28.61	4.80	20.35	4.26	35.95	7.49
	MLD	30	26.33	5.86	18.00	4.04	37.37	8.71
5	TA	142	32.87	4.45	23.57	3.69	41.13	6.32
	MLD	19	29.95	6.21	21.53	5.09	43.11	7.48
6	TA	180	34.61	5.17	26.37	4.20	45.77	7.79
	MLD	18	33.50	3.87	25.83	4.89	46.56	8.28
Total	TA	979	26.81	7.75	19.92	6.48	34.91	10.35
	MLD	130	23.62	9.11	16.98	7.26	34.92	11.56

[a] Number of correctly solved items. The two-digit subtest was not administered in Grade 1

TA Typically Achieving children, MLD children with Mathematical Learning Difficulties

symbolic magnitude knowledge might be related (Siegler & Lortie-Forgues, 2015). This will help us to more fully understand the various developmental trajectories of different mathematical abilities and their individual differences and different profiles (e.g., Reeve et al., 2012). Such research inevitably requires large sample sizes in which short, reliable and valid assessments are required to measure different cognitive variables in an efficient way.

The present study demonstrates that the SYMP test has satisfactory test-retest reliability and satisfactory construct and criterion-related validity. On a broader level, our data are in line with the existing body of evidence that showed an association between symbolic numerical magnitude comparison and mathematics achievement (De Smedt et al. 2013; Schneider et al., 2016). The present study, however, investigated this association in a very large sample and goes beyond the previous studies by showing for the first time that this association exists for both one- and two-digit number processing and that it remains significant across all grades of primary school. Children's performance on the one-digit subtest was more strongly associated with mathematics achievement in Grade 1 than in Grades 4, 5, and 6. This is similar to Holloway and Ansari (2009), who observed a decrease in the association between one-digit magnitude comparison and mathematics achievement from Grade 1 to Grade 3. On the other hand, significant associations between the two-digit subtest and mathematics achievement were found in all grades and no significant grade differences in the size of this latter association were observed. As suggested by Holloway and Ansari (2009), the variability in numerical magnitude comparison in older children is not adequately captured by using one-digit numbers, and two-digit magnitude comparison tasks are therefore better suited for this age group.

Data on the SYMP Test revealed age-related improvements in children's performance from Grade 1 to Grade 6, extending the findings of Nosworthy et al. (2013), who observed similar changes on their paper-and-pencil measure from Grade 1 to Grade 3. Our data are in line with studies that used computerized numerical magnitude comparison tasks in more narrow age-ranges (e.g., Holloway & Ansari, 2009; Landerl & Kölle, 2009; Sasanguie et al., 2013) and extend these earlier studies, by showing that these age-related improvements can be observed across the entire primary school.

The current findings are also consistent with the two earlier investigations that studied the association between paper-and-pencil measures of one-digit symbolic numerical magnitude processing and mathematics performance (Durand et al., 2005; Nosworthy et al., 2013). We extend this data by showing that such association can be observed across the entire primary school and by additionally including two-digit magnitude comparison. The focus on two-digit comparison tasks

might be particularly relevant to detecting individual differences in mathematics in languages where there is an inconsistency between the Arabic numbers and verbal number words in two-digit numbers, such as in Dutch (current sample) or German (e.g., the number word for 34 is "four-and-thirty"), as previous studies have shown that specifically in these languages, the processing of two-digit numbers modulates arithmetic performance (e.g., Göbel et al., 2014).

Crucially, our data go beyond the previous ones by evaluating for the first time the psychometric properties of such group-administered paper-and-pencil measures. Test-retest correlations were higher than .70 across all grades and for children's overall test score (= total number of correctly solved items summed for the two different subtests), even though the associations were not always above .70 for a single subtest or single grade. This suggests that the SYMP Test represents a reliable measure of symbolic magnitude processing (Hunsley & Mash, 2008).

Turning to the validity of the SYMP Test, high associations between the one-digit and two-digit subtest as well as high correlations between the paper-and-pencil test and its computerized version of this test were observed. The associations between the paper-and-pencil test and a standardized achievement test for mathematics were moderate. This all shows that the SYMP Test had sufficient convergent validity. Importantly, all these associations remained significant after controlling for children's performance on the control task for motor speed, which excludes the possibility that they were merely the result of children's general processing speed. As expected, correlations between the paper-and-pencil test and a standardized test for spelling were much lower or not significant and, additionally, these correlations were significantly lower than the correlations between the paper-and-pencil test and the standardized test for mathematics achievement, which all indicates that the SYMP Test has satisfactory discriminant validity.

As expected, typically developing children performed better on the one-digit and two-digit subtest than children with MLD for all grades, except for Grade 6. These group differences also remained significant after controlling for children's motor speed. These findings are in line with previous studies who found that children with MLD are impaired in their ability to compare one-digit and two-digit numbers (De Smedt & Gilmore, 2011; Landerl et al., 2004; Landerl & Kölle, 2009; Rousselle & Noël, 2007), and show that the SYMP Test has sufficient criterion-related validity. It is interesting to point out that no group differences were found in Grade 6. One explanation for this observation might be that the processing of one-digit and two-digit numbers is already highly automatized at this age, even in children with MLD, which reduces the probability to find significant differences between children with MLD and their typically developing peers. From an educational point of view, the current measure allows for an early

detection of children at-risk for MLD, and consequently the quick start of effective remedial interventions that target children's numerical skills (Clements & Sarama, 2011), such as the Number Worlds (Griffin, 2007).

There are various theoretical reasons for speculating a connection between symbolic number comparison and mathematical development. For example, children gradually progress in their arithmetic development from counting-all to counting-on-from-larger (Geary et al., 1992). This latter advanced counting strategy requires a determination on the larger number and therefore draws on the comparison of numerical magnitudes. Similar comparison processes might play a role in more advanced mental calculation strategies, such as the flexible use of strategies in multi-digit subtraction (Linsen et al., 2015). The use of more advanced counting strategies further fosters the memorization of problem-answer associations in long-term memory, i.e., the development of arithmetic facts and these appear to be stored in long-term memory in a meaningful way that is according to their magnitude (e.g., Robinson, Menchetti, & Torgesen, 2002). Finally, recent data by Bailey, Siegler and Geary (2014) indicate that whole number magnitude knowledge provides an important scaffold for knowledge of fractions in middle school.

It is important to acknowledge that number comparison is only one of the many facets of children's early numerical competencies that might be relevant for subsequent mathematical development. Indeed, other studies have emphasized the roles of other early numerical competencies, such as counting (Geary et al., 1992; Reeve et al., 2012), subitizing (Schleifer & Landerl, 2011), spontaneous focusing on numbers (Hannula & Lehtinen, 2005) and number line estimation (Booth & Siegler, 2008). Future studies should carefully investigate how these early numerical competencies are related to symbolic number comparison and to each other and how each of these variables uniquely predicts subsequent mathematics achievement in school. The integrated study of these various competencies and their interactions in one sample puts a heavy burden on children to complete the tests as well as on researchers to collect the data. The availability of short, reliable, and valid assessments, as were developed in the current study, offers a nice opportunity to study these competencies in concert rather than in isolated studies.

Although the present study included all grades of primary school, it remained cross-sectional. As a result, the comparisons of the results between different grades should be treated with caution. The current cross-sectional data also do not allow us to establish predictive associations between symbolic number comparison and mathematics achievement. Some studies have indeed shown that symbolic number comparison predicts future mathematics achievement (De Smedt et al., 2009) and math development (Vanbinst et al., 2015), yet these predictive associations need to be studied across larger age-ranges in longitudinal studies, in which also bidirectional associations should be considered. The current paper-and-pencil measure allows one to study these associations on a large-scale level.

The current data also do not allow us to establish a causal association between symbolic number comparison and mathematics achievement. This calls for specific intervention studies that target symbolic number comparison skills and investigate their effect on mathematics achievement. Such intervention studies have been reported, yet the transfer of these intervention effects to mathematics achievement has been limited so far (De Smedt et al., 2013, for a discussion).

Taken together, the present findings suggest that the SYMP Test is a reliable and valid instrument to assess primary school children's ability to process numerical magnitudes. In view of the robust associations between symbolic numerical magnitude comparison and mathematics achievement (De Smedt et al., 2013; Schneider et al., 2016), the test has the potential to be used as a screening measure to identify children who are at risk to develop mathematical difficulties at an early age. This early identification might enable earlier treatment for these at-risk children with intervention programs that have been developed to support numerical magnitude processing and mathematics achievement (Clements & Sarama, 2011; Ramani & Siegler, 2011; Räsänen, Salminen, Wilson, Aunio, & Dehaene, 2009).

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington: Author.

Andersson, U., & Östergren, R. (2012). Number magnitude processing and basic cognitive functions in children with mathematical learning disabilities. *Learning and Individual Differences, 22,* 701–714. doi:10.1016/j.lindif.2012.05.004

Bailey, D. H., Siegler, R. S., & Geary, D. C. (2014). Early predictors of middle school fraction knowledge. *Developmental Science, 17,* 775–785. doi:10.1111/desc.12155

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

Booth, J. L., & Siegler, R. S. (2008). Numerical magnitude representations influence arithmetic learning. *Child Development, 79,* 1016–1031. doi:10.1111/j.1467-8624.2008.01173.x

Brankaer, C., Ghesquière, P., & De Smedt, B. (2014). Numerical magnitude processing deficits in children with mathematical difficulties are independent of intelligence. *Research in Developmental Disabilities, 35,* 2603–2613. doi:10.1016/j.ridd.2014.06.022

Butterworth, B., Varma, S., & Laurillard, D. (2011). Dyscalculia: From brain to education. *Science, 332,* 1049–1053. doi:10.1126/science.1201536

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. doi:10.1016/j.actpsy.2014.01.016

Chiswick, B. R., Lee, Y. L., & Miller, P. W. (2003). Schooling, literacy, numeracy and labour market success. *The Economic Record, 79,* 165–181. doi:10.1111/1475-4932.t01-1-00096

Clements, D. H., & Sarama, J. (2011). Early childhood mathematics intervention. *Science, 333,* 968–970. doi:10.1126/science.1204537

Cohen, J. (1988). The significance of a product moment r_s. In J. Cohen (Ed.), *Statistical power analysis for the behavioral sciences* (pp. 75–107). Hillsdale: Lawrence Erlbaum Associates.

De Smedt, B., & Gilmore, C. K. (2011). Defective number module or impaired access? Numerical magnitude processing in first graders with mathematical difficulties. *Journal of Experimental Child Psychology, 108,* 278–292. doi:10.1016/j.jecp.2010.09.003

De Smedt, B., 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,* 48–55. doi:10.1016/j.tine.2013.06.001

De Smedt, B., Verschaffel, L., & Ghesquiére, P. (2009). The predictive value of numerical magnitude comparison for individual differences in mathematics achievement. *Journal of Experimental Child Psychology, 103,* 469–479. doi:10.1016/j.jecp.2009.01.010

Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology, 16,* 626–641. doi:10.1037/0096-1523.16.3.626

Dudal, P. (2000a). *Leerlingvolgsysteem: Wiskunde-Toetsen 1-6 [Student monitoring system: Mathematics-tests 1-6].* Leuven: Garant.

Dudal, P. (2000b). *Leerlingvolgsysteem: spelling-toetsen 1-6 [Student monitoring system: Spelling-tests 1-6].* Leuven: Garant.

Durand, M., Hulme, C., Larkin, R., & Snowling, M. (2005). The cognitive foundations of reading and arithmetic skills in 7- to 10-year-olds. *Journal of Experimental Child Psychology, 91,* 113–136. doi:10.1016/j.jecp.2005.01.003

Fazio, L. K., Bailey, D. H., Thompson, C. A., & Siegler, R. S. (2014). Relations of different types of numerical magnitude representations to each other and to mathematics achievement. *Journal of Experimental Child Psychology, 123,* 53–72. doi:10.1016/j.jecp.2014.01.013

Geary, D. C. (2011). Consequences, characteristics, and causes of mathematical learning disabilities and persistent low achievement in mathematics. *Journal of Developmental and Behavioral Pediatrics, 32,* 250–263. doi:10.1097/DBP.0b013e318209edef

Geary, D. C., Bow-Thomas, C. C., & Yao, Y. H. (1992). Counting knowledge and skill in cognitive addition - a comparison of normal and mathematically disabled children. *Journal of Experimental Child Psychology, 54,* 372–391. doi:10.1016/0022-0965(92)90026-3

Geary, D. C., Hoard, M. K., Byrd-Craven, J., Nugent, L., & Numtee, C. (2007). Cognitive mechanisms underlying achievement deficits in children with mathematical learning disability. *Child Development, 78,* 1343–1359. doi:10.1111/j.1467-8624.2007.01069.x

Gerardi, K., Goette, L., & Meier, S. (2013). Numerical ability predicts mortgage default. *Proceedings of the National Academy of Sciences, 110,* 11267–11271. doi:10.1073/pnas.1220568110

Göbel, S. M., Moeller, K., Kaufmann, L., Pixner, S., & Nuerk, H. C. (2014). Language affects double-digit addition in Arabic notation: Number word inversion leads to a stronger carry-effect. *Journal of Experimental Child Psychology, 119,* 17–25. doi:10.1016/j.jecp.2013.10.001

Griffin, S. (2007). Early intervention for children at risk of developing mathematical learning difficulties. In D. B. Berch & M. M. Mazzocco (Eds.), *Why is math so hard for some children? The nature and origins of mathematical learning difficulties and disabilities* (pp. 373–396). Baltimore: Brookes Publishing.

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

Hannula, M., & Lehtinen, E. (2005). Spontaneous focusing on numerosity and mathematical skills of young children. *Learning and Instruction, 15,* 237–256. doi:10.1016/j.learninstruc.2005.04.005

Holloway, I. D., & Ansari, D. (2009). Mapping numerical magnitudes onto symbols: The numerical distance effect and individual differences in children's mathematics achievement. *Journal of Experimental Child Psychology, 103,* 17–29. doi:10.1016/j.jecp.2008.04.001

Hunsley, J., & Mash, E. J. (2008). Developing criteria for evidence-based assessment: An introduction to assessments that work. In J. Hunsley & E. J. Mash (Eds.), *A guide to assessments that work* (pp. 3–14). New York: Oxford University Press.

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up. Helping children learn mathematics.* Washington, DC: National Academy Press.

Kolkman, M. E., Kroesbergen, E. H., & Leseman, P. P. M. (2013). Early numerical development and the role of non-symbolic and symbolic skills. *Learning and Instruction, 25,* 95–103. doi:10.1016/j.learninstruc.2012.12.001

Landerl, K., Bevan, A., & Butterworth, B. (2004). Developmental dyscalculia and basic numerical capacities: A study of 8-9-year-old students. *Cognition, 93,* 99–125. doi:10.1016/j.cognition.2003.11.004

Landerl, K., Fussenegger, B., Moll, K., & Willburger, E. (2009). Dyslexia and dyscalculia: Two learning disorders with different cognitive profiles. *Journal of Experimental Child Psychology, 103,* 309–324. doi:10.1016/j.jecp.2009.03.006

Landerl, K., & Kölle, C. (2009). Typical and atypical development of basic numerical skills in elementary school. *Journal of Experimental Child Psychology, 103,* 546–565. doi:10.1016/j.jecp.2008.12.006

Linsen, S., Verschaffel, L., Reynvoet, B., & De Smedt, B. (2015). The association between numerical magnitude processing and mental versus algorithmic multi-digit subtraction in children. *Learning and Instruction, 35,* 42–50. doi:10.1016/j.learninstruc.2014.09.003

Mazzocco, M. M. M., Feigenson, L., & Halberda, J. (2011). Impaired acuity of the approximate number system underlies mathematical learning disability (dyscalculia). *Child Development, 82,* 1224–1237. doi:10.1111/j.1467-8624.2011.01608.x

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

Noël, M. P., & Rousselle, L. (2011). Developmental changes in the profiles of dyscalculia: An explanation based on a double exact-and-approximate number representation model. *Frontiers in Human Neuroscience, 5,* 165. doi:10.3389/fnhum.2011.00165

Noël, M. P., Rousselle, L., & Mussolin, C. (2005). Magnitude representation in children: Its development and dysfunction. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 179–195). Hove: Psychology Press.

Nosworthy, N., Bugden, S., Archibald, L., Evans, B., & Ansari, D. (2013). A two-minute paper-and-pencil test of symbolic and non-symbolic numerical magnitude processing explains variability in primary school children's arithmetic competence. *PLoS ONE, 8,* e67918. doi:10.1371/journal.pone.0067918

Nuerk, H.-C., Weger, U., & Willmes, K. (2001). Decade breaks in the mental number line? Putting the tens and units back in different bins. *Cognition, 82,* B25–B33. doi:10.1016/s0010-0277(01)00142-1

Ramani, G. B., & Siegler, R. S. (2011). Reducing the gap in numerical knowledge between low- and middle-income preschoolers. *Journal*

of *Applied Developmental Psychology, 32,* 146–159. doi:10.1016/j.appdev.2011.02.005

Räsänen, P., Salminen, J., Wilson, A. J., Aunio, P., & Dehaene, S. (2009). Computer-assisted intervention for children with low numeracy skills. *Cognitive Development, 24,* 450–472. doi:10.1016/j.cogdev.2009.09.003

Reeve, R., Reynolds, F., Humberstone, J., & Butterworth, B. (2012). Stability and change in markers of core numerical competencies. *Journal of Experimental Psychology-General, 141,* 649–666. doi:10.1037/a0027520

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

Reynvoet, B., & Brysbaert, M. (1999). Single-digit and two-digit Arabic numerals address the same semantic number line. *Cognition, 72,* 191–201. doi:10.1016/S0010-0277(99)00048-7

Robinson, C. S., Menchetti, B.M., & Torgesen, J. K. (2002). Toward a two-factor theory of one type of mathematics disabilities. *Learning Disabilities Research & Practice, 17,* 81–89. doi:10.1111/1540-5826.00035

Rousselle, L., & Noël, M. P. (2007). Basic numerical skills in children with mathematics learning disabilities: A comparison of symbolic vs. non-symbolic number magnitude processing. *Cognition, 102,* 361–395. doi:10.1016/j.cognition.2006.01.005

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,* 418–431. doi:10.1016/j.jecp.2012.10.012

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

Schneider, W., Eschmann, A., & Zuccolotto, A. (2002). *E-prime reference guide.* Pittsburg: Psychology Software Tools.

Schneider, M., Beeres, K., Coban, L., Merz, S., Schmidt, S., Stricker, J., & De Smedt, B. (2016). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: a meta-analysis. *Developmental Science.* doi:10.1111/desc.12372

Sekuler, R., & Mierkiewicz, D. (1977). Children's judgments of numerical inequality. *Child Development, 48,* 630–633. doi:10.1111/j.1467-8624.1977.tb01208.x

Siegler, R. S., & Lortie-Forgues, H. (2015). Conceptual knowledge of fraction arithmetic. *Journal of Educational Psychology, 107,* 909–918. doi:10.1037/edu0000025

Van Hoof, J., Verschaffel, L., & Van Dooren, W. (2015). Inappropriately applying natural number properties in rational number tasks: Characterizing the development of the natural number bias through primary and secondary education. *Educational Studies in Mathematics, 89,* 39–56. doi:10.1007/s10649-015-9613-3

Vanbinst, K., Ghesquière, P., & De Smedt, B. (2012). Numerical magnitude representations and individual differences in children's arithmetic strategy use. *Mind, Brain, and Education, 6,* 129–136. doi:10.1111/j.1751-228X.2012.01148.x

Vanbinst, K., Ghesquière, P., & De Smedt, B. (2014). Arithmetic strategy development and its domain-specific and domain-general cognitive correlates: A longitudinal study in children with persistent mathematical learning difficulties. *Research in Developmental Disabilities, 35,* 3001–3013. doi:10.1016/j.ridd.2014.06.023

Vanbinst, K., Ghesquière, P., & De Smedt, B. (2015). Does numerical processing uniquely predict first graders' future development of single-digit arithmetic? *Learning and Individual Differences, 37,* 153–160. doi:10.1016/j.lindif.2014.12.004