Learning multiplication: An integrated analysis of the multiplication ability of primary school children and the difficulty of single digit and multidigit multiplication problems

van der Ven, S.H.G.; Straatemeier, M.; Jansen, B.R.J.; Klinkenberg, S.; van der Maas, H.L.J.

DOI
10.1016/j.lindif.2015.08.013

Publication date
2015

Document Version
Final published version

Published in
Learning and Individual Differences

Learning and Individual Differences 43 (2015) 48–62

Contents lists available at ScienceDirect

Learning and Individual Differences

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

Learning multiplication: An integrated analysis of the multiplication ability of primary school children and the difficulty of single digit and multidigit multiplication problems☆

Sanne H.G. van der Ven [a,*], Marthe Straatemeier [b], Brenda R.J. Jansen [c], Sharon Klinkenberg [b], Han L.J. van der Maas [b]

[a] Department of Pedagogical and Educational Sciences, Utrecht University, PO Box 80140, 3508 TC Utrecht, The Netherlands
[b] Department of Psychology, Psychological Methods, University of Amsterdam, Weesperplein 4, 1018 XA Amsterdam, the Netherlands
[c] Department of Psychology, Developmental Psychology, University of Amsterdam, Weesperplein 4, 1018 XA Amsterdam, the Netherlands

ARTICLE INFO

Article history:
Received 10 June 2014
Received in revised form 24 April 2015
Accepted 15 August 2015

Keywords:
Mathematics
Multiplication
Difficulty of multiplication problems
Computer adaptive testing
Item response theory

ABSTRACT

In this study the mental multiplication ability of primary school children and the difficulty structure of all single digit and 469 multidigit multiplication problems, each solved tens of thousands of times in a web-based practice program, were investigated.
Child analyses indicated three groups: single digit problem solvers, multidigit problem solvers, and high performers. Within-grade ability differences were very large.
In the item analyses, previously identified effects in single digit multiplication, such as the *problem size effect* and the *tie effect*, were replicated in one integrated analysis. Data from two tasks were contrasted: one in which children used predominantly computational strategies, and one in which they are expected to rely mostly on retrieval. In both tasks we found most support for the computational efficiency model. Finally, exploratory analyses on the difficulty of multidigit problems suggest that children rely on the base 10 structure of our number system.

1. Introduction

As one of the four basic mathematical operations, multiplication is an important skill learned in primary education. Especially mastery of the single-digit tables of multiplication is an important aim of primary education, as it forms the basis for other operations such as division and multidigit multiplication. With the present paper we provide a comprehensive view of the development of multiplication skills. The aims of this paper are twofold: first we show what multiplication problems children in different grades are capable of solving and we elucidate the nature of individual differences within grades in multiplication ability. Second, we investigate the difficulty structure of single digit and multidigit multiplication problems. We focus on the development of mental calculation, i.e., calculation in the head, without external aids such as paper and pencil. Especially in single digit multiplication mental calculation is considered important, as in most countries children learn to memorize the multiplication tables. Moreover, mental multiplication is considered an important skill in mathematics, as mental multiplication seems to evoke the use of insightful procedures rather than the application of rote-memorized procedures, and countries that stress the use of mental calculation in education tend to do well in international comparisons (Dowker, 2005).

In schools, mathematics curricula are based on intuitive notions of the order of problem difficulty: children learn to solve easy problems first, and then continue with more difficult problems. When applying this to one of the basic mathematical operations, multiplication, children start with single-digit multiplication, and then continue with multidigit multiplication. For instance, the Dutch aims for arithmetic and mathematics state that children should understand the concept of multiplication and start memorizing the single-digit tables of multiplication in grade 2 (7–8 years) and continue this memorization process in grades 3 and 4 (SLO, 2009). Then children learn to solve multidigit multiplication problems by means of written column-wise algorithms and analogical strategies (e.g., $20 \times 5 = 100$ because $2 \times 5 = 10$). However, it is unknown to what degree this order in the curriculum accurately reflects the order in which children master the ability to solve these problems. Moreover, while it is known that children vary greatly in multiplication skills, most studies on individual differences focus on explaining these differences (e.g., by age differences or differences in cognitive ability) rather than identifying the size of these differences (e.g., Lemaire & Siegler, 1995; Mabbott & Bisanz, 2003; Van der Ven, Boom, Kroesbergen, & Leseman,

☆ This research was supported by a grant from the Netherlands Organization for Scientific Research (NWO).
* Corresponding author.
E-mail addresses: s.vanderven@uu.nl (S.H.G. van der Ven), m.straatemeier@uva.nl (M. Straatemeier), b.r.j.jansen@uva.nl (B.R.J. Jansen), s.klinkenberg@uva.nl (S. Klinkenberg), h.l.j.vandermaas@uva.nl (H.L.J. van der Maas).

http://dx.doi.org/10.1016/j.lindif.2015.08.013
1041-6080/

2012). To design an optimal curriculum, it is important to determine the order in which problems are mastered, the problems that are typical for each grade, and the extent of the within-grade individual differences in multiplication skills. Moreover, if individual differences are qualitative, different types of instruction may be recommendable.

1.1. Solving single-digit multiplication problems: effects, theories and mechanisms

Differences exist not only at the level of the child, but also at the level of math problems; not all single digit multiplication problems are solved equally fast and accurately, even by adults. For example, 8×7 is more difficult than 2×2. Different effects have been identified in the literature, explaining systematic differences in response time and accuracy between problems: the *problem size effect*, the *tie effect*, the *order effect*, the *parity effect* and effects related to *specific operands*. The *problem size effect* (for a review, see Ashcraft & Guillaume, 2009) is the effect that perhaps has been demonstrated most often: problems with small operands are solved faster and more accurately than problems with large operands (Butterworth, Marchesini, & Girelli, 2003; Campbell & Graham, 1985; Imbo, Vandierendonck, & Rosseel, 2007; Mabbott & Bisanz, 2003). Despite the robustness of this finding, little attention has been paid to the operationalization of problem size, and various seemingly arbitrary definitions have been used: the sum of the operands (Parkman, 1972), the product (Butterworth et al., 2003; Imbo, Vandierendonck, & Rosseel, 2007; Mabbott & Bisanz, 2003; Siegler, 1988), the minimum (Parkman, 1972), the maximum (Parkman, 1972), or problems are divided into a small-sized and a large-sized subset (Campbell & Graham, 1985). Other studies have posited an alternative for the *problem size effect*: *consistency effects* (Campbell, Dowd, Frick, McCallum, & Metcalfe, 2011; Domahs, Delazer, & Nuerk, 2006; Verguts & Fias, 2005a, 2005b). The answer to each problem is compared to the answers to its eight neighboring problems: problems in which one of the operands is 1 or 2 higher or lower, e.g., for problem 6×3 the neighboring problems are 4×3, 5×3, 7×3, 8×3, 6×1, 6×2, 6×4, and 6×5. If the decades of the answers match, the neighbors are decade-consistent and thus anchored more firmly in a network. For example, 6×3 is decade-consistent with three of its neighboring problems: 4×3, 5×3, and 6×2 but not with all other neighbors. Similarly, problems with the same unit are said to be unit-consistent, but there are only few unit-consistent neighbors. Problem 6×3 has none. Problems with a high degree of consistency have been found to be easier; as smaller problems have a higher decade consistency, this may explain the problem size effect (Domahs et al., 2006).

The *tie effect* describes the phenomenon that problems with equal operands, such as 6×6, are easier than problems with different operands (Campbell & Graham, 1985; Campbell & Gunter, 2002; De Brauwer & Fias, 2009; De Brauwer, Verguts, & Fias, 2006). The *order effect* is the phenomenon that problems presented with the larger operand first may be easier than problems in the reverse format (Butterworth et al., 2003), although this effect was not confirmed in another study (Robert & Campbell, 2008). In most curricula problems are first introduced with the larger operand first, as part of the smaller times table: 9×2 is part of the two times table and presented earlier than 2×9, which is part of the nine times table.

The *parity effect* refers to the even/odd status of the answer. In some studies with verification tasks (Krueger, 1986; Lemaire & Fayol, 1995), participants rejected incorrect products more quickly if the parity of the incorrect answer was incongruent with the parity of the correct answer. Alternative explanations, however, are that regardless of the odd/even status of the answer, participants have a bias favoring even answers, because three quarters of all multiplication results are even (Lochy, Seron, Delazer, & Butterworth, 2000), or, that children have a bias for answers with the same parity as the answer to the addition problem with the same operands. Finally, *operand effects* mean that specific operands make a problem relatively easy because of regularities in the times tables: especially the operands 0, 1, 2, 5, and 9 (Butterworth et al., 2003; Campbell & Graham, 1985; De Brauwer et al., 2006; LeFevre et al., 1996; Mabbott & Bisanz, 2003).

There are different models that account for these differences in problem difficulty. These models are based on computational efficiency, memory strength, and network interference (Ashcraft & Guillaume, 2009).

In computational efficiency models it is assumed that various computational strategies are used to solve multiplication problems (Imbo, Duverne, & Lemaire, 2007; Imbo & Vandierendonck, 2008; Lemaire & Siegler, 1995; Mabbott & Bisanz, 2003; Siegler, 1988; Van der Ven et al., 2012; Wu et al., 2008). Some problems require more and more difficult steps than others (LeFevre et al., 1996). For example, when repeated addition or skip counting (counting while skipping numbers, e.g., saying only every third number, thus effectively reciting a times table) is used, this requires only three small steps for the problem $3 \times 2 = 2 + 2 + 2$, but six larger steps for $6 \times 7 = 7 + 7 + 7 + 7 + 7 + 7$, leading to a longer latency and a higher probability of procedural errors for the latter problem.

In both memory strength and network interference models, the emphasis is placed on retrieval as the dominant strategy. According to memory strength models, during the learning process each problem is increasingly strongly associated to its answer in long term memory. However, problems can also be associated with incorrect answers. The more strongly a problem is associated with the correct answer, and the less it is associated with competing incorrect answers (in other words, the more peaked the distribution of answers is), the more easily it can be retrieved (Siegler, 1988). Frequency of exposure might underlie these differences (Campbell & Graham, 1985). Network interference models state that all math facts are stored together in a rich network. Each problem activates different candidate answers from which the correct alternative must be selected (Campbell, 1987; Domahs et al., 2006; Verguts & Fias, 2005a, 2005b). Problems associated with many other candidate answers may be more difficult than problems with few competing answers. Problems and their commutative counterparts (e.g., 7×3 and 3×7) are stored together in one network node (Rickard, Healy, & Bourne, 1994), possibly with the larger operand first (Butterworth et al., 2003). There are some correspondences between memory strength and network interference models: in both, different answers compete to be retrieved. However, only network interferences models emphasize the interrelations between different math problems.

The different models differ partially in the predictions they make about the existence of the different effects in multiplication problems as described in the previous section, but specific predictions are not always clear. It should also be noted that different research groups are working on these models and even within each type of model full consensus about the specific details has not been reached. Table 1 lists the predictions for each effect as we derived them from each of the three types of models. To contrast the models as much as possible we attempt to be strict in the derivation of the predictions. The predictions from these models show several similarities. This is not surprising for two reasons: (1) frequency of exposure is supposed to underlie differences in memory strength, and those problems that are computationally easier are presented earlier and more often in math curricula (Ashcraft & Christy, 1995), and (2) a shorter computational procedure may by itself lead to a stronger memory anchoring of the problem. Yet, as Table 1 shows, there are sufficient differences between the models to allow for empirical tests.

Although the three types of models have sometimes been treated as mutually exclusive (e.g., Domahs et al., 2006; LeFevre & Liu, 1997), it is perhaps more likely that they hold under different circumstances. Indeed, the use of only one of these models seems to conflict with the existence of individual and developmental differences in addressing multiplication problems.

Adults rely predominantly (albeit not exclusively) on memory retrieval for single digit multiplication (Campbell & Xue, 2001; LeFevre

Table 1
Predictions and explanations of each type of theory for several effects identified in single-digit multiplication problems.

Effect		Model		
		Computational efficiency	Memory strength	Network interference
Problem size	Prediction	Larger problems are more difficult	Larger problems are more difficult	Larger problems are more difficult
	Explanation	Larger problems contain more and more difficult steps (LeFevre et al., 1996; Van der Ven et al., 2012)	Larger problems are presented less often during the learning process (Ashcraft & Christy, 1995)	Larger problems are less distinct from each other, relative to their own size (Dehaene, 1997; Siegler & Opfer, 2003), and/or neighboring problems show less consistency (Campbell et al., 2011; Domahs et al., 2006)
Tie	Prediction	No effect	Ties are easier	Ties are easier
	Explanation	Computation is not different for ties than for non-ties	Ties may be presented more frequently (Siegler, 1988) and ease storage (Thevenot, Barrouillet, & Fayol, 2001)	Ties have fewer neighbors and, on average, neighbors are more often inconsistent than consistent (Campbell et al., 2011; Verguts & Fias, 2005b)
Order	Prediction	No effect	m > n easier[1]	Unclear: no effect, or m > n easier.
	Explanation	Problems are equally complex, regardless of the format: either there are more steps, or the steps are bigger.	In the Dutch curriculum, problems in m > n format are on average presented earlier in education (SLO, 2009).	Single node for both problem types (Rickard et al., 1994); possibly with larger operand first (Butterworth et al., 2003)
Parity	Prediction	No effect	No effect	Even problems are perhaps easier
	Explanation	No overall difference in difficulty of computational steps, apart from special operands.	Even and odd problems are not presented in a particular order.	Most multiplication answers are even (Lochy et al., 2000)
Special operands (1,2,5,9)	Prediction	These numbers are easier	No effects	Effects of 1 and 5
	Explanation	Computational steps are easier because of regularities (Krueger, 1986; Lemaire & Siegler, 1995; Mabbott & Bisanz, 2003; Sherin & Fuson, 2005)	No effects in addition to problems size effect	Problems with a 1 have fewer neighbors; problems with a 5 have high unit consistency (Domahs et al., 2006)

[1] m and n refer to the first and second operand of a multiplication problem (m × n), respectively.

et al., 1996). Children, on the other hand, at least in many western countries such as the Netherlands, the setting of the present study, are in the beginning of their learning process encouraged to learn from using their own computational strategies (Freudenthal, 1991) that become increasingly sophisticated (Lemaire & Siegler, 1995; Van der Ven et al., 2012) and that contribute to the creation of a rich network of meaningful associations between numbers (Baroody, 1985), a process taking years to develop (Campbell & Graham, 1985; De Brauwer et al., 2006). Only once the network is strong enough to retrieve an answer with sufficient confidence, retrieval will be used frequently (Siegler, 1988). This is expected to lead to computational efficiency models dominating during the learning process, and network interference models and/or memory strength models during adulthood.

Indeed, a decrease in strength with age has been demonstrated for the *problem size effect* (Campbell & Graham, 1985; De Brauwer et al., 2006; Koshmider & Ashcraft, 1991), and for the *tie effect* (De Brauwer et al., 2006). Another study showed that the *tie effect* and *five effect* disappeared when corrected for retrieval use (LeFevre et al., 1996). It should, however, be noted that when a combination of retrieval and computational strategies is used, relative differences in response time within the group of problems solved with retrieval are small compared to the differences between problems solved with computational strategies. This may obscure differences within the group of problems predominantly solved with retrieval. Correcting for retrieval also implies an unwanted correction for difficulty of the problem (since easy problems are solved with retrieval more often). It is therefore valuable to gather data in which similar strategies (computational or retrieval) are used on all problems, regardless of the difficulty of the problem, and then test for the presence of the previously identified effects. In the present study we created such data sets: by using computer-adaptive algorithms (explained in the next section), we ensured that regardless of difficulty of the problem, in one task problems were predominantly solved with computational strategies, while in another task children relied mostly on retrieval.

Moreover, in most previous studies, effects were tested in isolation, and a comprehensive analysis including all effects, including interaction effects, is still lacking. Interactions between effects are conceivable: the *problem size effect* has been shown to be smaller for tie problems than for non-ties (De Brauwer & Fias, 2009; De Brauwer et al., 2006), and a smaller *parity effect* has been shown for small problems than for large problems (Lemaire & Fayol, 1995). Though not investigated to our knowledge, one may also imagine an interaction between problem size and special operands, because the regularities in the counting strings are relatively more advantageous for large problems. In the present study, therefore, we included a comprehensive analysis of all previously described effects together: problem size, decade consistency, tie, parity (with two operationalizations: whether the problem is even, or whether the parity matches the parity of addition problems), order, and the special operands one, two, five, and nine, and abovementioned interactions.

1.2. Difficulty of multidigit multiplication problems

In contrast to the number of studies exploring the difficulty of single digit multiplication problems, there are very few studies in which the difficulty structure of multidigit problems was investigated systematically. In one study it was found that the number of carries was related to response latency and (inversely) to accuracy (Imbo, Vandierendonck, & Vergauwe, 2007). Other studies investigating multidigit multiplication were done, but mostly focused on other aspects of multiplication, such as brain activation (Grabner et al., 2009), the role of working memory (Imbo, Duverne, et al., 2007; Imbo, Vandierendonck, & Rosseel, 2007; Imbo, Vandierendonck, & Vergauwe, 2007), and the influence of problem format (Hickendorff, 2013). More studies on the difficulty structure of multidigit problems are needed, especially studies in which multiple effects are integrated (Imbo, Vandierendonck, & Vergauwe, 2007). Therefore we formulated and tested two hypotheses, which we call the single digit analogy and the decomposition hypothesis.

If the single digit analogy hypothesis holds, then the same effects as in single digit multiplication, discussed in the previous section, affect the difficulty of multidigit problems: effects should be found of *problem size*, *tie*, *order*, *parity*, and *special operands*. Note that in multidigit problems there are even more possible operationalizations of problem size,

such as the average size of each digit, or the number of digits in the multiplicands.

The second hypothesis is that children multiply numbers in a decomposed way. When two-digit numbers are encoded, units and decades have been shown to be processed separately (Moeller, Huber, Nuerk, & Willmes, 2011). If an analogous process takes place while doing mental calculation, each combination of two digits (one digit from each number) is multiplied and the results are added. This is somewhat similar to the algorithms that children use to solve written problems. This means that in each multiplication problem $m \times n$, if N_m is the number of digits in m and N_n the number of digits in n, the number of necessary multiplication operations equals $N_m \times N_n$ (e.g., 26×749 requires $2 \times 3 = 6$ intermediate multiplication operations). This is then followed by a series of additions: one fewer than the number of multiplications. The total number of operations is thus $2 * (N_m \times N_n) - 1$. If the result of an intermediate operation is a multidigit number, this further complicates the procedure, since it requires a carry in the addition. In the example above, the multiplication of the units is $6 \times 9 = 54$; the 5 must be carried from the units to the tens. The digits 0 and 1 do not require a carry, and require only a very easy multiplication operation: especially zeros in the end can merely be added to the result. The digit 2 only requires a carry when multiplied with 5 or higher. These numbers are therefore expected to be easier. Finally, if one of the multiplicands is a single digit number, this reduces the complexity of the required operations. Therefore, if children solve multidigit multiplication problems in this way, the number of operations, number of carries, and the number of zeros, ones, and twos, and presence of a single-digit operand, should give a good account of problem difficulty. If children also apply the analogical strategy of adding zeros in the end, the number of operations excluding these end-zeros should give a better account than the total number of operations. If one of the multiplicands is a power of ten (10, 100, 1000…), the problem is reduced to the addition of the same number of zeros to the right of the other multiplicand (e.g., 46×100 can be solved by adding two zeros to the 46), so powers of ten serve as special operands.

In the present study we compare and contrast these two hypotheses. We also fit a model combining the best predictors from each model, with interactions, as it can be imagined that powers of ten, and ones or twos are less prone to effects of problem size and number of operations.

1.3. Aims of the study

The aims of the present study are twofold. The first aim is to provide an overview of children's multiplication ability and the structure of within-grade differences in this ability in different grades of primary education. The second aim is to investigate and explain the difficulty structure of single digit and multidigit multiplication problems at different stages during the learning process: the period in which children predominantly rely on computational strategies and the later period when they rely most on retrieval. The second aim is complicated, because children gradually learn to solve problems with retrieval, a process dependent on the difficulty of the problem (Lemaire & Siegler, 1995; Van der Ven et al., 2012). It is therefore impossible to identify a population or sensitive age during which children solve all multiplication problems with computational strategies: children who apply computational strategies to easy problems are not capable of solving the difficult problems at all. Moreover, even children of the same age and in the same grade differ substantially from each other in multiplication ability.

This problem can be overcome with Math Garden, an online computer-adaptive math program that administers problems of such difficulty that children are reasonably capable of solving them. A brief description follows here; more detailed information follows in the Method section. Problem selection in Math Garden is adaptive: child ability and problem difficulty estimates ensure a predetermined constant success rate for every child. We contrasted data from two Math Garden tasks: the multiplication task (for the sake of contrast from now on referred to as 'regular multiplication task') and the speed task. In the regular multiplication task, children are presented with multiplication problems of such a difficulty that they are predominantly capable of solving the problems, but not overly fast. Retrieval strategies are therefore expected to be rare. In the speed task, children have to answer very fast and are thus presented with problems that they can solve with retrieval in most cases. As a result, we can see to what degree the previously described effects are present in different stages and how well these fit the predictions based on the different theories.

2. Method

2.1. Participants

Math Garden is available for both schools and private users. Most (but not all) Math Garden users are children attending primary schools in the Netherlands that use Math Garden in their curriculum: 91% of Math Garden users are primary school children, and approximately 5% of all primary school children in the Netherlands have a Math Garden account. Participating schools and families gave permission for the use of the data of their pupils for research purposes; the schools accepted the responsibility of informing the parents about the research and voluntary participation. Parents were given the opportunity to refuse the use of their children's data; data from these children were not distributed to the researchers.

The first section of the results is based on the regular multiplication task, and aimed at identifying differences between children. For these analyses we extracted data from those primary school children who had attempted at least 30 problems. Math Garden generates a continuous data flow and data are thus not restricted to a certain time frame, necessitating further restrictions to prevent a contamination of within-grade individual differences with within-grade developmental effects. Without such restriction we would be comparing children that played at many different time points during the school year, depending on when they started (and possibly quit) playing in Math Garden. Therefore we restricted the grade analysis to the data obtained within a limited time frame, halfway through the school year: January 2013–February 2013. During this time frame, 27,495 children (47% females; mean age = 10.1 years, SD = 5.1 years) played the regular multiplication task and were thus included in this analysis.

Some children had been registered by their teachers with extremely unlikely combinations of grade and date of birth (e.g., a twelve-year-old in grade 1). In such large samples it is impossible to verify this information but it is likely that mistakes have been made, although it is impossible to tell whether the grade or the date of birth is incorrect. Therefore we decided to remove the data from these children and accept that this inevitably leads to the removal of some correct data too. We excluded data from the children whose reported age deviated more than one year from the trimmed mean age (trimming the 10% extremes on each end), and from the children with reported dates of birth that had been reported improbably frequently (mainly January 1 of any year). This led to the removal of data from 2.7% of the children. Descriptive statistics for the remaining 26,753 children are given in Table 2.

Table 2
Characteristics of the children in January and February 2013.

Grade	N (% girls)	Age in years			
		Mean	SD	Min	Max
1	1325 (33%)	6.84	0.41	5.89	7.86
2	4556 (43%)	7.80	0.40	6.85	8.84
3	6142 (49%)	8.83	0.41	7.90	9.89
4	5717 (50%)	9.86	0.42	8.91	10.91
5	5095 (50%)	10.89	0.42	9.92	11.92
6	3918 (49%)	11.91	0.42	10.92	12.92
Total	26,753 (47%)	9.62	1.53	5.89	12.92

The second part of the results section concerns analyses of problem difficulties of both the speed task and the regular multiplication task. All Math Garden users, i.e. everyone who used Math Garden (91% primary school children), contributed to the difficulty estimates of the items. Children who were excluded from the child analyses also influenced item difficulty, and thus contributed to the data used for the item analyses. These analyses were based on a total of 92,092 users that had played the regular multiplication task, and the 59,955 users that had played the speed task within the time frame of the study: 2011, week 42 up to 2013, week 20 (81 weeks).

2.2. Instruments

2.2.1. Math Garden: computer interface

Math Garden is a web interface with plants, each representing a task (see left panel of Fig. 1). The higher the child's ability, the bigger the plant. In the present paper, the regular multiplication task and the speed task were used.

2.2.1.1. Regular multiplication task. The regular multiplication task becomes available once children have the basic addition or subtraction ability that children are expected to have when multiplication is introduced in schools: i.e., when they can solve problems like 4 + 18 or 57–5. Most children reach this ability in grade 1 or 2.

After selecting the multiplication plant, the regular multiplication task starts, in which children receive a problem that must be solved within 20 s. The answer is given by clicking on a virtual keypad (see middle panel of Fig. 1). The remaining time is reflected as a row of coins, from which a coin disappears with each passing second. Upon answering, the correct answer is shown and the child receives the number of remaining coins if the answer was correct, but loses this number of coins if it was incorrect. This implements the High Speed High Stakes scoring rules (see section Math Garden: computer adaptive technology). With the coins, prizes for a virtual trophy cabinet can be bought. The child is thus motivated to answer as quickly as possible once (s)he knows the answer, but to refrain from answering otherwise. There is a question mark that the child can click if (s)he does not know how to compute the answer. In this case, and also if the child did not provide an answer within the time limit, no coins are won or lost and the correct answer is shown. The task ends after 15 problems, but children can quit earlier or play several times.

2.2.1.2. Speed task. The speed task (see right panel of Fig. 1) is similar to the regular multiplication task, with a few differences: the problems are mixed (addition, subtraction, multiplication, division); the answer is given in a 6-choice multiple choice format and children have only 8 s instead of 20 to solve the problem. This shorter time limit means that the task is more difficult, and children are thus presented with easier problems to establish the same constant success rate of 75%. The task contains single digit as well as multidigit problems, but for the scope of this paper we only analyzed the single digit problems: multidigit problems were not presented very frequently, as for most children these are too difficult to be solved correctly in the short time allowed in this task.

2.2.2. Math garden: computer adaptive technology

Person abilities and item difficulties are central in Math Garden. Every instance of a child solving an item is used to update the estimates of this ability of the player and the difficulty of the item by means of a computer adaptive algorithm based on Item Response Theory (IRT) modeling. The continually updated estimates of player abilities and item difficulties are used to select appropriate math problems for each child. The updating procedure is based on a procedure invented by Elo (1978), still often used to rank chess players. This method is briefly outlined here and in more detail in Appendix A. For a full description, we refer to Klinkenberg, Straatemeier, and Van der Maas (2011), and for the mathematical foundations to Maris and Van der Maas (2012).

A child's score on an item depends on both accuracy and response time. The score equals the proportion of remaining time: positive if the answer was correct and negative if it was incorrect. That means that the score is 1 for an immediate correct answer and −1 for an immediate incorrect answer. The score linearly decreases/increases to 0 as the response time approaches the time limit. E.g., a response after 8 s in the regular multiplication task, with a time limit of 20 s, would result in a score of 0.6 if the answer is correct and −0.6 if the answer is incorrect, because 60% of the time remains. This combination of accuracy and response time leads to a system with strong psychometric properties (Klinkenberg et al., 2011; Maris & Van der Maas, 2012).

Prior to each problem, the program determines a child's expected score. Based on previous performance, each child has an ability estimate and each item has a difficulty estimate. These estimates are both on the same scale. This scale is arbitrary, but because both are on the same scale, the child's expected score on a particular item can be derived from the difference between child ability and item difficulty, in a way very similar to IRT modeling (Klinkenberg et al., 2011). This expected score is close to 1 if the child's ability is far higher than the problem difficulty, and close to −1 if the reverse is the case; if both are equal, the expected score is 0.

After the child solves the item, the expected score and the actual, obtained score are compared. The child's ability is adjusted upward if the child scored higher than expected and downward if the child scored lower than expected. In a similar fashion, the item difficulty estimate is adjusted downward if the child scored higher than expected and otherwise upward. The larger the discrepancy between observed and expected score, the larger the adjustment. This means that the rank order of the players (from high to low ability) and the problems (from high to low difficulty) are determined empirically, on the fly, without the necessity to administer every problem to every child. In a developmental context, the order of problem difficulties shows the average order in which problems are acquired by children. New children and new items enter the system with an initial ability estimate based on their age (child) and problem size (item). These estimates are updated after every attempted problem, and they tend to approach their true value rather quickly (Klinkenberg et al., 2011).

Fig. 1. Math Garden. Left panel: Math Garden computer interface. Each plant represents a task; the third plant from the left is the regular multiplication task. Middle panel: example problem in the regular multiplication task. 'wis' means 'erase'. Right panel: example problem in the speed task. The coins in the bottom of the screen represent the score; these gradually disappear as the time passes.

The item selection procedure is also based on item difficulty and child ability estimates. Based on a child's current ability estimate, the difficulty is determined at which the child has a probability of .75 of answering correctly. The next item is sampled around this difficulty level (M = .75,[1] SD = .10). This results in selecting items with a difficulty estimate that is approximately 1 point lower than the child's ability estimate; then the probability of success is around .75. This probability of .75 correct is the default success rate, but children can also choose to solve easier items, with an average probability of .90 correct, or more difficult items, with a probability of .60.

The presence of an experimenter in the room is not required, which enables vast sample sizes: tens of thousands of children from different backgrounds. The problems were solved tens of thousand times each, leading to highly reliable estimates of difficulty. The lack of physical presence of an experimenter means that 'cheating' is possible, as someone other than the child may solve the problems on the child's behalf. However, the system is self-corrective: massive amounts of data prevent these occasions from having a large influence on the problem difficulties, and once the child starts playing by him- or herself again, the adaptive algorithms ensure that the ability estimate also quickly reaches its true value again. Indeed, the Math Garden math tasks have good criterion validity: they correlate strongly, $r = .78$–$.84$ (Klinkenberg et al., 2011), with norm-referenced math tests (Janssen, Scheltens, & Kraemer, 2005), and recent, unpublished data from two samples show a correlation of .55 and .71 between the regular multiplication task and the multiplication subscale on the Tempo Toets Automatiseren (TTA; De Vos, 2010), a speeded math fact retrieval task. The latter correlation is somewhat lower, despite the greater similarity in the test, but this is likely due to a ceiling effect on the multiplication subscale of the TTA. For the speed task these data are not available.

User ability estimates and problem difficulty estimates enable two types of analyses: user and item analyses. User analyses, in which the sample consists of the children playing the task, aim at identifying ability differences between children. Item analyses, in which the sample consists of a set of items, are aimed at explaining item difficulty differences using various item characteristics. Both types of analyses were carried out in the present paper, although the focus lies on the item analyses.

It must be noted that the IRT algorithms work with the assumption that there is one, unidimensional, scale underlying all math problems. Since every item is different from the other items, there may be qualitative differences, for instance if a child has been practicing one specific item really hard. Although this issue may be most visible when IRT procedures are used, every type of study that uses a sum score on a test for further analyses faces this issue. Nevertheless, the items in this specific test are more alike than in most other multiplication tests. It would be surprising if multidigit multiplication is very different from single digit multiplication, as multidigit multiplication is essentially a combination of multiple single digit multiplication problems. Moreover, the item difficulty estimates have proven to be highly stable, whereas violations of the unidimensionality assumption lead to instabilities. Therefore we assume sufficient unidimensionality for the data to be used in this study.

2.3. Item inclusion

We analyzed the difficulty estimates of all 81 single digit multiplication problems (from 1×1 to 9×9), obtained under two conditions: in the regular multiplication task (relatively long time limit, predominantly computational strategies are expected) and in the speed task (short time limit, predominantly retrieval strategies are expected). We extracted these difficulty estimates on a weekly basis for 81 weeks: 2011, week 42 until 2013, week 20. During this period, the problems had been solved on average 48,830 times each (min = 30,495, max = 62,801) in the regular multiplication task and 9212 times (min = 0, max = 12,227) in the speed task. There were five items (1×1, 1×3, 1×4, 1×6, and 1×7) that as an unfortunate and unnoticed side-effect to a minor change in the selection algorithms had not been played during these 81 weeks in the speed task; other, similar problems were always selected instead. However, these items had all been played very often (around 28,000 times) before the time frame of the study, and therefore their difficulty estimates were considered sufficiently accurate to be used in the analyses.

For the multidigit problems, the problems with a 0 or 1 as one of the operands (e.g., 0×34 or 1×34) were excluded, as these problems do not require computational steps on multiple digits. Problems containing decimal numbers were also excluded. This yielded a total of 467 multidigit problems included in the analyses, that on average had been played 29,795 times each (min = 218, max = 83,649) in the regular multiplication task. To enhance ecological validity, total play frequency of the multidigit problems was entered as a weight in the regression analyses. This ensured that problems that were so difficult that they had hardly been presented would not affect the results too much. It must be noted that the problems were often relatively easy multidigit problems: when correcting for play frequency, 40% of all solved problems contained one single digit operand, and 41% contained an operand that was a power of 10 (10, 100, or 1000).

2.4. Analyses

The results section consists of two parts: an analysis of children's ability and an analysis of the difficulty estimates of the items. In the child ability section the distribution of the children's ability estimates in the regular multiplication task was examined. This analysis shows the extent of between-grade and within-grade ability differences. Since the number of children in these analyses is very large, we used an alpha of .001. We also investigated the possibility of multimodality: if different groups can be discerned, the data do not show a unimodal distribution with one peak but multiple peaks instead. With mixture modeling using the R package Mixtools (Benaglia, Chauveau, Hunter, & Young, 2009) this possibility was investigated. The purpose of mixture modeling is to find clusters of individuals with a similar ability. Multimodality of distributions (i.e., the presence of multiple normal distributions) indicates the existence of such clusters. Outcome measures of the analysis are the relative size (the proportion of individuals in the cluster), mean, and standard deviation of each cluster.

In the item analyses, we compared the relative difficulty of single digit and multidigit problems in the regular multiplication task. Furthermore, we performed separate regression analyses to predict the difficulty of single digit and multidigit problems. The Akaike Information Criterion (AIC) was used to compare different models: it is a measure favoring parsimony that can be used to compare non-nested models. A lower value indicates a better model (Akaike, 1974). Since the number of items is not so high, a correction of the alpha level was not necessary in these analyses and thus an alpha of .05 was used.

For the single digit analyses, the effects of all known characteristics and interactions between these characteristics were analyzed, both on the data from the regular multiplication task in which children predominantly used computational strategies, and on the speed task in which children predominantly used retrieval. All characteristics described in the introduction were included, i.e., *problem size*, the special numbers *one*, *two*, *five*, *and nine*, *the order effect*, two different operationalizations of *parity* (parity of the answer (odd or even) and same or different parity compared to addition problems with same operands), and *tie*. The exact operationalizations of these effects are presented in Table 3. It was also investigated whether different operationalizations of problem

[1] Normally in adaptive testing, a probability of .50 is maximally informative. This is, however, experienced as discouragingly low by test takers. Since response latency is included in the scoring method as well, it is possible to present easier problems, yet obtain sufficient and reliable information to determine players' abilities (see section 3.1 of Klinkenberg et al., 2011, for a more in-depth discussion).

Table 3
Operationalization of the different predictors in the regression analyses to predict problem difficulties.

Predictor	Operationalization
Single digit multiplication analyses	
Problem size	m + n (centered)
One	1 if at least one operand = 1, otherwise 0
Two	1 if at least one operand = 2, otherwise 0
Five	1 if at least one operand = 5, otherwise 0
Nine	1 if at least one operand = 9, otherwise 0
Order	1 if m > n, otherwise 0
Parity answer	1 if the answer is even, otherwise 0
Parity addition	1 if the outcome has the same parity as the outcome of the addition of the two problems (i.e., if both operands are even), otherwise 0
Tie	1 if m = n, otherwise 0
Relative decade consistency	Decade-consistent neighbors/number of neighbors[1]
Relative unit consistency	Unit-consistent neighbors/number of neighbors[1]
Multidigit multiplication analyses	
Problem size	Number of digits (centered)
Zeros	Number of zeros in both operands together
Ones	Number of ones in both operands together
Twos	Number of twos in both operands together
Fives	Number of fives in both operands together
Nines	Number of nines in both operands together
Order	1 if m > n, otherwise 0
Tie	1 if m = n, otherwise 0
Parity answer	1 if the answer is even, otherwise 0
Parity addition	1 if the outcome has the same parity as the outcome of the addition of the two problems (i.e., if both operands are even), otherwise 0
Operations	Number of digits in m * number of digits in n
Single digit	1 if one of the operands is a single digit, otherwise 0
Operations without zeros	Number of non-zero digits in m * number of non-zero digits in n
Ten	1 if at least one operand = 10, otherwise 0
Hundred	1 if at least one operand = 100, otherwise 0
Thousand	1 if at least one operand = 1000, otherwise 0
Five	1 if at least one operand = 5, otherwise 0
Fifty	1 if at least one operand = 50, otherwise 0
Five hundred	1 if at least one operand = 500, otherwise 0

[1] Neighbors consist of the two adjacent cells in all four direction with either the upper or lower triangle of the multiplication matrix following the definition of Verguts and Fias (2005a). 1 × 1 therefor only has 2 neighbors (2 × 1 and 3 × 1), whereas 3 × 1 has 6 neighbors (1 × 1, 2 × 1, 4 × 1, 5 × 1, 3 × 2, 3 × 3). The maximum number of neighbors is 8. Because answers differ in the number of neighbors we use relative measures of decade and unit consistency.

size (sum, product, minimum or maximum of the operands), all used in previous research, led to different results.

We started with specific predictors for each of the three models presented in Table 1. We then included predictors from all three models. These estimates of this model for the two tasks (regular and speed) are compared to investigate whether task demands influence preference for one of the theoretical models. For this analysis we applied a multigroup comparison using the R package Lavaan (Rosseel, 2012). For each predictor we tested with a chi-square test whether it can be constrained to be equal in both tasks. If not, the difference in estimates can be interpreted.

For the multidigit problems, data were analyzed from the regular multiplication task only. For these problems, a comparison with the speed task was not possible because these problems were so difficult that they had hardly ever been presented in the speed task.

The operationalization of all multidigit predictors is also presented in Table 3. Problem size was centered to prevent multicollinearity (Tabachnick & Fidell, 2001). In Model 1, predictors were as closely matched as possible to the single digit analysis. Number of zeros was added as a special number, because many multidigit problem operands end on zeros (e.g., 10 × 34). Model 2a also contains predictors related to the decomposition approach: predictors that were expected to be significant if children solve multidigit problems in a decomposed way, by multiplying all subcomponents and adding these results. These were number of operations, single digit as one of the operands, number of carries, number of zeros, ones and twos, and special numbers ten, hundred and thousand. In Model 2b two predictors, the number of operations and zeros, were replaced by one predictor: the number of operations excluding end zero(s) because zeros in the end are exceptional, as they do not require any operation but can merely be added to the final result. In Model 3a, an exploratory, combined model was created, including all predictors from the previous models that proved to be significant. In Model 3b interactions were added between the special numbers (from Model 1) and number of operations (from Model 2), resembling the interactions in the single digit analysis.

3. Results

3.1. Descriptive statistics

The upper part of Fig. 2 shows the distribution of child ability estimates in the regular multiplication task. The horizontal axis of Fig. 2 shows the ability of the children: the further towards the right, the higher the ability. As in all latent variable models, the scale of the item difficulty estimates is the same as the children's ability estimates. The ability estimate thus yields information about the difficulty of the items that children are capable of solving. When the child ability and problem difficulty are equal, the child has a probability of .50 of answering the problem correctly. To illustrate this, in Fig. 2 a number of example problems of various difficulty levels are shown. The upper row in gray shows some example single digit problems, and the lower two rows display multidigit problems. The problems are a representative selection of all difficulty levels — all single digit problems had a difficulty estimate below −1.

For instance, the most common ability of children in grade 4 is around −1, and children performing at this level have a .50 probability of solving problems of this difficulty level correctly: e.g., 8 × 7 and 100 × 500. When children play in Math Garden with the default medium difficulty setting of a probability of .75, the difficulty estimates of the presented problems are on average 1.1 point lower than their own ability.[2]

3.2. Child analyses: children's multiplication ability

Fig. 2 clearly shows quantitative differences related to grade. Children in higher grades had a higher mean ability: the curves shifted towards the right for each consecutive grade. A one-way ANOVA confirmed that this difference was significant and large, $F(5; 26{,}747) = 3581$, $p < .001$, $\eta^2 = .401$. Tukey HSD post-hoc tests showed that differences between all grades were significant. Nevertheless, within each grade, the ability level of the children differed greatly: whereas on average the mean ability estimates of consecutive grades differed by 1.7 point, the standard deviation within each grade was far higher: 2.3 points in grade 1, 2.2 in grade 2, 2.3 in grade 3, 2.9 in grade 4, 3.7 in grade 5 and even 4.9 points in grade 6. The best performing children in grade 1 even outperformed the poorest performers in grade 6.

Interestingly, the distributions per grade also suggest qualitative differences. Especially the distributions of the highest grades demonstrate multimodality: multiple peaks are visible. To test for the presence of multiple groups rather than a single unimodal distribution, we fitted mixtures of normal distributions for each grade using the R-package MixTools (Benaglia et al., 2009). Per grade we compared mixtures of 1, 2, 3, and 4 clusters. Visual inspection of the data in Fig. 2 suggested that a two-cluster solution would be sufficient, but according to the fit statistics we needed either 3 or 4 clusters to fit the data. Given the

[2] If a child chooses to solve easy problems (90% correct), the problem difficulty is on average 2.2 points lower than the child's ability, and if the difficult setting (60%) is chosen, the problem difficulty is 0.4 points lower.

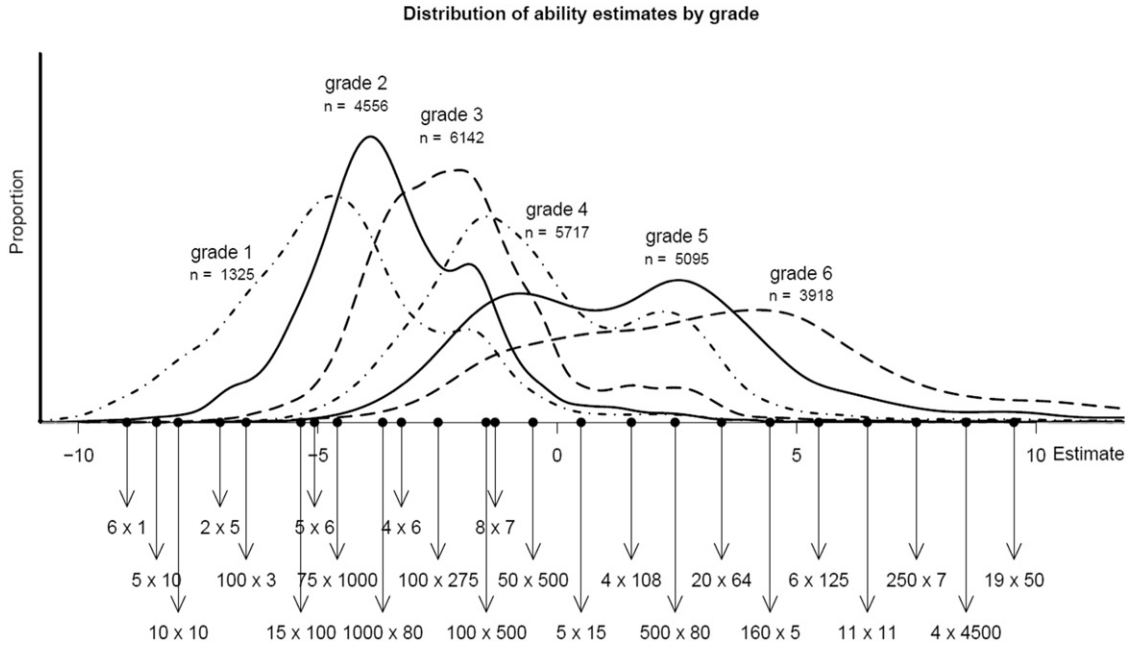

Fig. 2. Problem difficulty and child ability estimates in the regular multiplication task. The horizontal axis displays the obtained estimates: both on the same scale. The further towards the right, the more able the child. Example problems illustrate the difficulty of the problems along this scale. When playing in Math Garden, children are presented problems with a difficulty estimate that is around 1 point lower than their own ability estimate; this ascertains that their success rate is around the preset level of .75 correct.

very large sample size, the power of these analyses is very high, meaning that there is a statistical bias in favor of (too) many clusters. Inspection of the four-cluster solution showed that the fourth cluster was always very small and strongly overlapped with one of the other clusters. The three-cluster models, on the other hand, proved very well interpretable. Therefore we present the results of the three-cluster models. Fig. 3 presents for each cluster the relative size (left panel), the mean ability (middle panel) and standard deviation (right panel) in each grade.

It must be noted that models in which the means and standard deviations of the clusters were constrained to be equal across grades did not fit the data. The clusters are therefore not the same in every grade, as the middle and right panel of Fig. 3 illustrate. Nevertheless, there are great similarities. The mean ability estimate of cluster 1 was below 0 in every grade, as the middle panel of Fig. 3 shows, although it slightly increased with grade. When looking at the horizontal axis of Fig. 2 for the interpretation of this value it becomes clear that this cluster consisted of children of such ability that they were still working on single digit

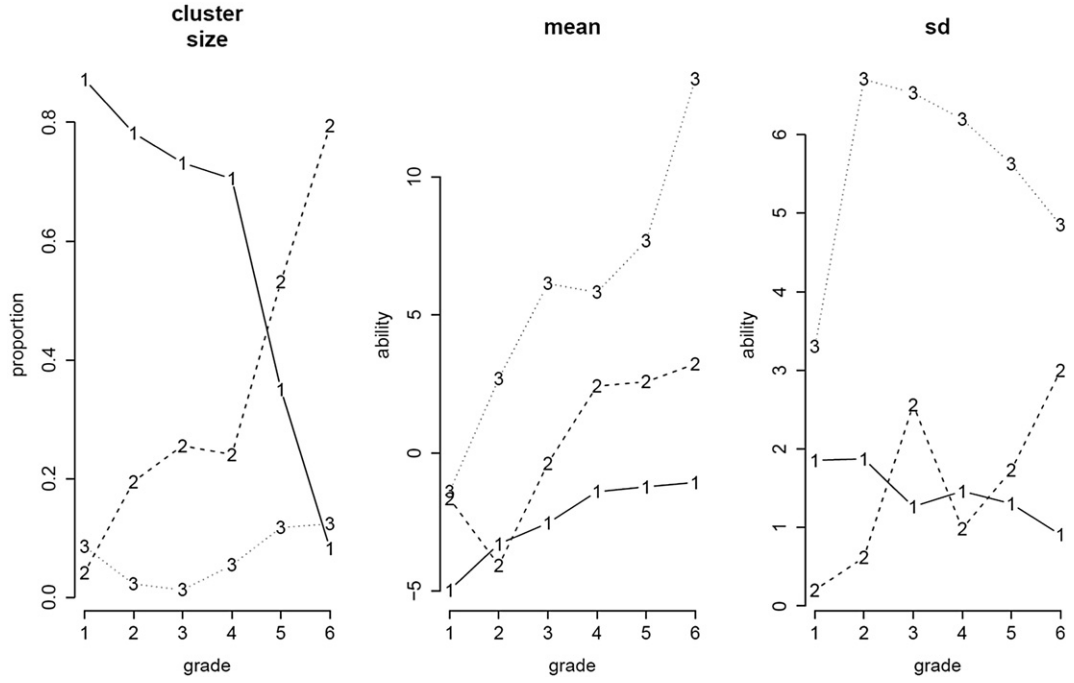

Fig. 3. Size, mean and sd of the three clusters in each grade. Clusters are indicated by their respective numbers in the graphs. Cluster 1 = single digit problem solvers, cluster 2 = multidigit problem solvers, and cluster 3 = high performers.

multiplication and simple multidigit problems (multiplying by a power of 10, such as 75 × 1000). This single digit problem solving group was dominant in size in grades 1–4, where 70 to 90% of the children were single digit problem solvers (see left panel of Fig. 3). The small increase with grade in mean illustrates that in higher grades these children performed slightly better: in the lower grades children in cluster 1 were working on easier problems than children in cluster 1 in the higher grades.

The second cluster, dominant in grades 5 and 6, had a higher mean. Children in this cluster no longer worked on single digit problem solving; they solved the more difficult multidigit problems. Cluster 3, finally, was very small in size, with a high mean that increased with grade, and a high standard deviation. Cluster 3 thus consisted of a small and diverse group of high to extremely high performers.

In grades 1 and 2, clusters 2 and 3 show a slightly different pattern that fits with a lower developed ability in these lower grades. In grade 1, the mean ability of clusters 2 and 3 was still rather low; groups of multidigit problem solvers and high performers could not be distinguished in this grade. In grade 2, cluster 2 was approximately equal to cluster 1, and cluster 3 was comparable to cluster 2 in other grades, meaning that there were essentially two groups in this grade: a group of single digit problem solvers and a group of multidigit problem solvers, and no high performers.

3.3. Item analyses

3.3.1. The difficulty of single digit vs multidigit problems: regular multiplication task

Fig. 2 already illustrated that the single-digit multiplication problems were found in the lower part of the item difficulty distribution, and indeed, a t-test showed that in the regular multiplication task, the single digit problems were significantly easier than the multidigit problems, $t(458.28) = 12.06$, $p < .001$. Nevertheless, as Fig. 2 shows, some multi-digit multiplication problems, for instance 100 × 25, were easier than even the majority of single-digit multiplication problems, and they were predominantly solved by children in the lower grades. A regression analysis showed that the type of problem (single digit or multidigit) explained only 16% of the variance, $F(1, 546) = 103.3$, $p < .001$.

3.3.2. Single digit multiplication: regular multiplication task and speed task

In the regular multiplication task, the mean response latency of the single digit problems was 7.21 s ($SD = 1.10$) and 81% of all problem solving attempts took more than four seconds. The mean response latency of the multidigit problems was 8.87 s ($SD = 1.20$) and 96% of the attempts took more than four seconds, indicating that computational strategies dominated in this task (Wu et al., 2008). In the speed task, the mean response latency of the single digit problems was 3.63 s ($SD = 0.30$) and only 34% of all problem solving attempts took more than four seconds, indicating that these problems were predominantly solved with retrieval. This means that although the division may not be perfect, computational strategies dominated in the regular multiplication task whereas retrieval dominated in the speed task.

While Fig. 2 only shows the difficulty estimates of a few example problems in the regular multiplication task, Fig. 4 displays the average difficulty estimates of all 81 single digit multiplications problems over the 81 weeks, for both the regular multiplication task (left panel) and the speed task (right panel). The horizontal axis shows the first operand, the second operand is displayed in the graph and the vertical position represents the problem difficulty.

Regression analyses were run, as described in the Method section. First we conducted regression analyses for each of the three theoretical models separately. Table 4 displays the parameter estimates for the predictors associated with the three theoretical models for the data of the regular task. The Computational Efficiency model explained the data very well, with an R^2 of .97 (.90 if interaction terms are excluded). All predictors are significant. The negative interaction terms imply that advantage of the special operands 1, 2, 5, and 9 increases with problem size.

The Memory Strength model is less convincing: the R^2 is only .68, and only one predictor, *problem size* is significant.

The Network Interference model explains .76 (.70 for the speed task) of the variance in ratings. The two *consistency* measures were significant, but the *parity* measures were less convincing – one effect was in the opposite direction: even problems were more difficult. An additional interaction term between *unit* and *decade consistency* did not improve the model. It was also tested whether the decade and unit consistency variables could explain the *special operand effects* from the Computational Efficiency model. Therefore all main effects from the Computational Efficiency model except *problem size* were added to the

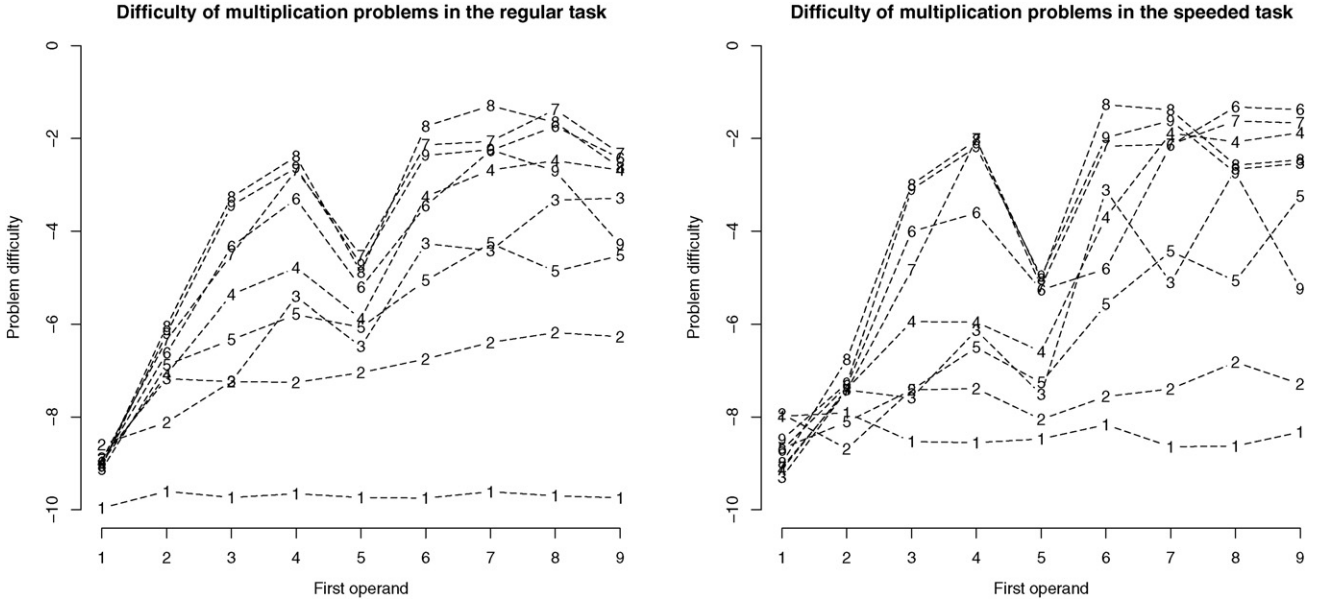

Fig. 4. Single digit problem difficulty estimates (averaged over 81 weeks) in the regular and speed task. The first operand is shown on the horizontal axis; the second operand is represented within the graph.

Table 4
Regression models for the three theoretical models for the data of the regular multiplication task.

	Computational efficiency (CE) R^2 = .97 (.91)[1]			Memory strength (MS) R^2 = .68 (.66)[1]			Network interference (NI) R^2 = .76 (.70)[1]		
Predictors	B	SE	Z-value	B	SE	Z-value	B	SE	Z-value
Intercept	−3.98	0.09	−44.4***	−5.57	0.25	−22.46***	−2.39	0.38	−6.29***
Problem size (PS)[2]	0.55	.03	18.3***	0.65	0.07	8.97***			
One	−5.53	0.25	−22.4***						
Two	−2.64	0.25	−11.1***						
Five	−1.33	0.13	−10.1***						
Nine	0.53	0.24	2.19*						
Order (m > n)				0.18	0.35	0.51			
Parity answer							0.70	0.34	2.07*
Parity addition							0.18	0.38	0.47
Tie (m = n)				0.27	0.55	0.49			
PS * one	−0.57	0.06	−10.4***						
PS * two	−0.34	0.05	−6.8***						
PS * five	−0.24	0.05	−5.2***						
PS * nine	−0.42	0.06	−7.3***						
PS * tie				−0.20	0.12	−1.68			
PS * order				−0.06	0.10	−0.58			
Decade consistency							−8.30	0.53	−15.63***
Unit consistency							−4.32	1.69	−2.55*

[1] R^2 for speed task between parentheses.
* p < .05.
*** p < .001.

Network Interference model. As a result, R^2 became .89, with strong significant *one* and *two effects*. The *one* and *two effects* were thus not explained by *decade* and *unit consistency*, but the *five* and *nine effects* were. When in the next step also *problem size* was entered as a predictor, the *decade* and *unit consistency effects* were no longer significant, whereas the *problem size effect* was, showing that *decade* and *unit consistency* could not explain the *problem size effect*.

It might be argued that some of the effects now included in one model could also apply to other models. Therefore we have also created a hybrid, somewhat more exploratory model with all predictors from the three models together. Table 5 presents an overall model with all predictors from Table 3, for both the regular and the speed task.

In a multigroup regression analysis it was tested whether the beta coefficients of the predictors could be constrained to be equal in both tasks. Where the p-value in the last column of Table 5 is significant, such equality was not allowed. Only a few differences between the models were found. This result was stable: removing non-significant predictors from the analysis did not change the pattern. In both tasks *problem size* significantly increased problem difficulty, illustrated by lines that increase with problem size in Fig. 4. Special operands *one*, *two*, and *five* significantly decreased problem difficulty, which is illustrated by dents and relatively low lines for problems with these operands in Fig. 4. The significant negative interactions of all these operands with *problem size* indicate that the facilitating effect of these operands is stronger for larger problems. Note that the coefficient of the *nine effect* has a positive sign, but this coefficient should be interpreted together with the corresponding negative interaction effect with *problem size*: from 2 × 9 onwards, the resulting effect of the presence of a nine was that the problem became easier. The effects for *two* (both main and interaction) were stronger in the speed task. Since we found no differences for the other special operands we hesitate to interpret this difference between the two tasks.

Table 5
Exploratory multigroup regression analysis.

	Regular task R^2 = .98			Speed task R^2 = .95			Equality test
Predictors	B	SE	Z-value	B	SE	Z-value	P-value
Intercept	−3.67	0.22	−16.96***	−3.40	0.37	−9.19***	0.532
Problem size[2]	0.53	0.04	15.10***	0.58	0.06	9.68***	0.460
One	−5.46	0.20	−27.33***	−4.89	0.34	−14.31***	0.152
Two	−2.35	0.17	−14.09***	−3.31	0.29	−11.58***	0.004**
Five	−1.29	0.23	−5.73***	−1.41	0.39	−3.65***	0.798
Nine	0.45	0.19	2.40*	0.93	0.32	2.88**	0.202
Order (m > n)	0.18	0.08	2.28*	−0.20	0.13	−1.46	0.017**
Tie (m = n)	−0.77	0.14	−5.38***	−1.54	0.24	−6.33***	0.007*
Parity answer	0.11	0.11	1.05	0.06	0.18	0.34	0.813
Parity addition	0.14	0.12	1.16	0.11	0.20	0.52	0.886
PS * one	−0.61	0.04	−15.20***	−0.75	0.07	−10.88***	0.086
PS * two	−0.35	0.04	−9.65***	−0.56	0.06	−8.94***	0.005**
PS * five	−0.24	0.04	−6.80***	−0.23	0.06	−3.78***	0.869
PS * nine	−0.40	0.04	−9.65***	−0.48	0.07	−6.71***	0.355
PS * tie	−0.08	0.03	−2.65**	−0.18	0.05	−3.55***	0.087
PS * order	−0.06	0.02	−2.57*	−0.01	0.04	−0.15	0.245
Decade cons.	−1.03	0.46	−2.25*	−1.22	0.79	−1.55	0.841
Unit cons.	−0.44	1.07	−0.41	−1.97	1.83	−1.08	0.469

* p < .05.
** p < .01.
*** p < .001.

The *order effect* that we associated with the Memory Strength model was significant in this hybrid multigroup analysis in the regular task, but not in the test of the Memory Strength model itself (Table 4). However, in the regular task the effect was positive, contrary to expectations, and in the speed task it was not significant. Altogether there is no stable evidence for an *order effect*.

The *tie effect* was also significant in this hybrid multigroup analysis, but not in the test of the Memory Strength model. In the multigroup analysis, in both tasks there was a strong significant *tie effect* in the expected direction. It was significantly stronger in the speed task, suggesting that memory strength plays a larger role in this task.

We found no evidence for a *parity effect* in the analyses. This is not due to the correlation between the parity predictors: deleting one of the parity predictors from the analysis did not change the results for parity.

We only found a weak effect of decade consistency, and only in the regular task. Unit consistency did not have a significant effect on the problem difficulties. A possible reason for this could be multicollinearity: *decade consistency* correlates .91 with *problem size*. However, deleting consistency measures from the overall model did not noticeably decrease R^2, while deleting problem size from the overall

model decreased R^2 to a value of .94 (.90 for the speed task), showing that problem size was a better predictor overall than the interference measures.[3] We redid the regression analyses with other operationalizations of problem size: the product, the minimum, the maximum and the squared sum of the two operands. The overall pattern of results was the same, but the model with the sum of the operands proved to be the best model.

Together, the results show a stable pattern in favor of the computational efficiency model for both types of tasks, with effects of *problem size* that were stronger than effects of *decade* and *unit consistency*, and effects of the special operands *one*, *two*, *five*, and *nine*, and interactions between problem size and these operands. Furthermore we found weaker evidence in favor of effects of *tie* and *order*, as the Memory Strength models predicted, but no effects for *parity*.

3.3.3. Multidigit multiplication

Regression analyses were then performed on the difficulty estimates of the multidigit problems. The results of the analyses are shown in Table 6. Model 1, with predictors as similar as possible to the predictors entered in the single digit regression analysis, explained 36% of the variance in problem difficulty. Larger problem size[4] made a problem significantly more difficult, while zeros and ones made a problem easier. *Parity* and *tie effects* were also found: ties made a problem more difficult, while even problems were easier than odd problems. Interaction terms between problem size and special operands were also added, but these were removed because of high multicollinearity. None of the interactions was significant in the initial analysis.

In Model 2a, predictors related to the decomposition approach were analyzed. With 76% explained variance, this model was a far better model than Model 1 and all predictors were significant. However, zeros in the end are exceptional, as they do not require any operation but can merely be added to the final result. Therefore in Model 2b two predictors, the number of operations and zeros, were replaced by one predictor: the number of operations excluding end zero(s). While being more parsimonious, model 2b provided an even better fit than the previous models, with 78% explained variance and a lower AIC than model 2a. This suggests that children are aware that they can add zeros to the end result.

Then the exploratory, hybrid models were run, with significant predictors from all previous models. Model 3a, with only main effects, strongly resembled Model 2b, while the tie and *parity effect*, added from Model 1, lost significance. Model 3b, with interactions between the number of operations and all special operands, fitted the data best of all models, judging from the highest amount of explained variance ($R^2 = .83$) and the lowest AIC value. Four of the five interactions were significantly negative: the effect that special numbers make a problem easier is thus stronger in problems involving more operations. It must be noted, however, that the variance inflation factors in this model were somewhat high (up to 14), suggesting some multicollinearity in this model.

Table 6
Regression analyses explaining mean difficulties of multidigit problems.

	B	SE	t
Model 1. Single digit analogy			
$R^2 = .36$, $F(10, 456) = 25.20$, $p < .001$, AIC = 3144.38			
Intercept	12.01	1.22	9.80***
Problem size	4.19	0.36	11.48***
Zeros	−3.67	0.37	−9.84***
Ones	−2.87	0.29	−9.84***
Twos	−0.42	0.38	−1.11
Fives	0.43	0.34	1.25
Nines	0.60	0.59	1.02
Order	−0.23	0.35	−0.65
Tie	2.35	0.98	2.41*
Parity answer	−2.61	0.78	−3.34***
Parity addition	0.17	0.42	0.41
Model 2a: Decomposition approach			
$R^2 = .76$, $F(9,457) = 158.9$, $p < .001$, AIC = 2685.51			
Intercept	0.27	0.52	0.53
Operations	0.24	0.05	4.99***
Single digit	−2.47	0.35	−7.08***
Carry	4.72	0.25	18.94***
Zeros	−0.58	0.20	−2.96
Ones	−0.60	0.20	−3.03
Twos	−1.13	0.22	−5.05***
Ten	−4.82	0.33	−14.81***
Hundred	−4.26	0.36	−11.88***
Thousand	−4.20	0.45	−9.35***
Model 2b: Adapted decomposition approach			
$R^2 = .78$, $F(8,458) = 198.2$, $p < .001$, AIC = 2647.43			
Intercept	4.98	0.49	10.10***
Operations without zeros	0.73	0.09	8.06***
Single digit	−3.17	0.23	−13.88***
Carry	3.99	0.25	15.97***
Ones	−1.24	0.19	−6.62***
Twos	−1.50	0.20	−7.33***
Ten	−4.70	0.31	−15.05***
Hundred	−3.65	0.32	−11.26***
Thousand	−3.18	0.37	−8.58***
Model 3a: Hybrid model, main effects only			
$R^2 = .78$, $F(11,455) = 149.50$, $p < .001$, AIC = 2634.58			
Intercept	4.69	0.62	7.55***
Problem size (number of digits)	0.58	0.15	3.87***
Ones	−1.07	0.19	−5.52***
Twos	−1.31	0.21	−6.14***
Tie	0.60	0.57	1.06
Parity	−0.08	0.47	−0.18
Operations without zeros	0.71	0.09	7.65***
Single digit	−2.20	0.34	−6.56***
Carry	−4.07	0.25	16.40***
Ten	−4.44	0.31	−13.89***
Hundred	4.04	0.34	−11.87***
Thousand	−4.00	0.43	−9.27***
Model 3b: Hybrid model with interactions			
$R^2 = .83$, $F(17,449) = 126.1$, $p < .001$, AIC = 2544.98			
Intercept	5.51	0.90	6.11***
Operations	1.01	0.18	5.66***
Ones	−0.31	0.34	−0.92
Twos	−1.64	0.38	−4.30***
Ten	−7.41	0.89	−8.31***
Tie	0.50	0.52	0.97
Parity	0.62	0.43	1.45
Hundred	−7.62	0.76	−10.00***
Thousand	−10.22	1.01	−10.08***
Carry	3.36	0.30	11.25***
Single digit	−2.25	0.30	−7.30***
Problem size (number of digits)	0.62	0.14	4.47***
Operations*ones	0.46	0.08	5.46***
Operations*twos	−0.09	0.10	−0.85
Operations*tens	−0.97	0.21	−4.62***
Operations*hundred	−1.17	0.19	−6.11***
Operations*thousand	−1.77	0.24	−7.42***
Operations*carry	−0.24	0.07	−3.64***

* $p < .05$.
*** $p < .001$.

[3] Recently, another network-interference based alternative for problem size has been posited (De Visscher & Noël, 2014), based on the degree of similarity between the operands and answer of each problem and operands and answers of previously acquired problems. This measure proved a promising predictor, as it was a significant predictor when added to the Network Interference model in Table 4, and it increased the R^2 to .80 (.75). This encourages further development of network interference based predictors of item difficulty. However, inclusion of this interference measure did not change our final results. In a regression analysis together with all other predictors from Table 5, interference did not reach significance or change the explained variance, while problem size was still a significant predictor.

[4] In this analysis, problem size was operationalized by means of the total number of digits in m and n. Other operationalizations were again also tried: all operationalizations that were tried in the single digit analysis, as well as relative value (sum of digits/total number of digits). All yielded lower explained variance and a higher AIC value than the total number of digits.

The results suggest that children use the base 10 system of our numerical system in their multidigit calculations, and that special operands still play a role, as in single digit multiplication. However, the importance of these special operands matter less than in single digit multiplication, while the requirement of carries in the addition of partial results is strong.

4. Discussion

In the present study, we analyzed within-grade and between-grade differences in children's multiplication ability, and the structure of the difficulty of single digit and multidigit multiplication items. First, the results showed that differences in ability between children in different grades were present, but differences within the same grade were very large. Within-grade differences were larger than between-grade mean differences. Around 10% of the sixth-graders was even still working on the single digit problems while on the other hand approximately a quarter of the children in grade 1 was already working on multiplication before it had been introduced to them in school, sometimes even outperforming the poorest children in sixth grade. While the observation of differences between children is not new, the present data show and visualize just how large individual differences are. It is important that the educational system acknowledges these vast differences and adapts to the needs of children of all aptitude. The use of Math Garden or other (computerized) adaptive tools may be helpful. These tools can serve as a practice instrument that automatically selects appropriate problems for each child, and as a diagnostic instrument to identify differences between children without posing a high workload on the teacher. For instance Math Garden creates individual progress reports of strengths and weaknesses for the teacher.

A multimodal distribution was found and three groups could be discerned: single digit problem solvers, multidigit problem solvers and high performers. In the lower grades the majority of the children belonged to the single digit problem solving group; only in grades 5 and 6 the multidigit problem solving group was largest, while the group of high performers was always small. The multimodality suggests a qualitative shift: once children have memorized the single digit problems, they make a fast leap in their multiplication ability, as they are suddenly able to solve multidigit problems that are more complex than multiplication with only a power of 10 as one of the operands. The differences in cluster sizes suggest that this shift takes place markedly later than the curriculum would predict: the shift takes place in grade 5 or 6 for most children, even though they should have mastered single digit problems by the end of grade 3 according to the Dutch curriculum.

Second, the results showed that for children learning multiplication, as expected, single digit multiplication problems were relatively easy compared to multidigit problems. But there was also a substantial portion of the multidigit problems that was quite easy, especially when one of the operands was a power of 10. This does not follow the curriculum: multiplication by 100 and 1000 is taught years later than single-digit multiplication. This finding suggests that in an optimal curriculum multiplication with powers of ten may be taught earlier than common practice.

The final aim was an analysis of the difficulty structure of single digit and multidigit problems, with two types of tasks: in the regular task computational strategies dominated while in the speed task retrieval was used in the majority of the cases. A perfect distinction between computational and retrieval strategies was not achieved, but note that this reflects actual, common behavior: most people apply strategies flexibly, alternating between retrieval and computation even for very easy problems (Hecht, 2006; LeFevre et al., 1996; Siegler, 2007; Van der Ven et al., 2012).

The regression models for the two tasks are remarkably similar, despite the different time limits, and despite the fact that the regular task was a multiplication-only task in an open format, while the speed task contained multiplication problems interspersed with other operations (addition, subtraction, division) in a multiple choice format. The similarities in the results suggest that our results are very robust. The only meaningful difference was that we found a stronger tie effect in the speed task, consistent with the idea that memory strength plays a larger role in tasks with a lower time limit. We associated the tie effect with the memory strength model and not with the computational efficiency model because there are few computational advantages of problems containing a tie. They are more consistent with the memory strength model since the fact that the two operands are the same may make them especially easy to store in working memory and to associate with the answer, which may facilitate long term memory storage (Thevenot, Barouillet, & Fayol, 2001). Moreover, later in primary education the squaring operation is introduced, and tie multiplication problems are essentially squares, meaning that these problems are practiced and anchored in memory even stronger.

It was, however, not the case that the memory strength model was favored in the speed task above the computational efficiency model. In fact, in all analyses the results fitted better with the predictions made according to the computational efficiency theory. In line with predictions made by the theory and with previous findings, we found a problem size effect and effects of all special operands. Interestingly, the interaction effects between the problem size and special operands were rather strong and consistent. Operationalization of problem size did not affect the results strongly. In previous studies, the sum, product, minimum and maximum of the operands and decade consistency have been used. In our models, the best fitting model was the model using the sum of the operands, but in all cases the same predictors were significant. This shows that the results are fairly robust against different operationalizations of problem size.

The results also show that contrary to earlier research in adults, decade consistency was not a better predictor than problem size, also not in the speed task, a prediction made by network interference models (Domahs et al., 2006; Verguts & Fias, 2005a). Predictions of the network efficiency theory did not match our data: parity effect and possibly an order effect were both not reliably found. Moreover, the literature on network efficiency did not predict an effect of special operands two and nine, which were detected in the data. It may, however, be argued that problems with two as an operand have fewer neighbors, just as problems with one as an operand (although problems containing a one have even fewer neighbors).

This mismatch between our data and the predictions from Network Interference models may be related to our sample and setup of the tasks. It is possible that a memory network continues to develop until adulthood: even in the present speed task, the problems were still administered to children with an expected failure rate of .25, and problems had thus not been consolidated deeply. A similar, comprehensive analysis with adults may shed more light on other differences in the presence and strength of all effects.

Summarizing the single digit multiplication results, the present data are generally in line with the computational efficiency model for the both regular multiplication task and the speed task. The tie effect, and especially the fact that the tie effect is stronger in the speed task implies a role for memory strength too. We found no support for decade consistency, parity and order effects, which are associated with the network interference model.

The analyses of multidigit multiplication were more exploratory, due to a lack of existing theory of multidigit problem solving in the literature. We contrasted a model analogous to single digit multiplication, to two models based on decomposition of multidigit problems into different single digit multiplication problems, and created two models combining aspects of each different model, with and without interactions. The single digit analogy model did not explain the data well: the explained variance was low and most variables were not significant. The decomposition model explained the data much better, especially under the assumption that children first ignore end-zeros and add

these to the final result. The combined model, including the significant predictors from each model and including interactions, fitted the data best ($R^2 = .83$).

The results from this best model showed that in multidigit multiplication special operands were less important than in single digit multiplication, perhaps because the regularity in the tables of 5 and especially 9 is less pronounced in multidigit problems. Nevertheless, the presence of ones and twos made problems significantly easier. The results also showed that children applied the regularities of the base ten structure. First, the presence of 10, 100 and 1000 led to a lower problem difficulty. Second, number of operations without zeros predicted problem difficulty significantly better than number of operations including zeros, which suggests that children ignored end-zeros and added these to the final result. Fig. 2 illustrates this: many problems with 10 or 100 were easier than the majority of single digit problems. This means that children apply analogical strategies, which is one of the curricular aims, but most children started using this pattern already before its introduction in the curriculum: the majority of children in grade 2 solved some problems with operands 100 and 1000. As pattern recognition is an important feature of mathematics (Mulligan & Mitchelmore, 2009), this could be seen as a positive sign. At the same time, it may be promising for educational methods to introduce these analogical strategies earlier in the curriculum, since most children are already capable of carrying out these strategies, even without formal instruction. On the other hand, single digit problem solving ability is at the base of these strategies, and is therefore also a very important ability to train.

It should also be noted that multidigit problems turn out to be difficult to perform by means of mental computation. The majority of the multidigit problems that children received were relatively easy, because one of the operands is a power of ten, and/or because one of the multiplicands is a single digit number. In other words, the majority of Dutch primary school children are only capable of multidigit multiplication within 20 s, when relatively easy rules can be applied. While one might argue that this limits the results of the study, at the same time these results are ecologically valid, because mental calculation was required. More difficult multidigit problems are not solved with mental calculation in daily life; most people then revert to paper and pencil or a calculator.

Apart from yielding insight in children's multiplication learning, this study also demonstrates how very large data sets can be used. With the rise of the Internet, it has become easy to gather enormous amounts of data about every thinkable subject. While data mining is promising for science (Romero & Ventura, 2007), the question of what to do with these data is complicated. Most scientists, especially psychologists, are accustomed to and trained in the analysis of controlled but small experiments. With the present article we showed successfully how these data can be used to perform comprehensive analyses: on differences between children's ability and on the difficulty structure of problems spanning a very wide difference in difficulty. At the same time it must be acknowledged that this large scale data collection makes it impossible to control the circumstances to the same degree as small-scaled experiments allow. For example, the speed task differs in more than one aspect from the regular multiplication task: it employs a multiple choice format, and multiplication problems are intermixed with problems requiring other operations. Typically, when using data collected with methods not originally designed to meet the research questions of a specific study, certain characteristics in these data will not be optimal. Nevertheless, there are still large benefits of our approach: with the same budget, a vastly larger data set can be obtained. Therefore we see this way of data collection not as a substitute for but rather as an addition to traditional small-scaled ways of data collection: each method has its own strengths and weaknesses.

To summarize, we found differences in average ability between all grades, but individual differences within each grade were larger. A multimodal distribution indicated that children could be distinguished as single digit problem solvers, multidigit problem solvers, and high performers.

Single digit multiplication problems were easier than multidigit problems, but there were large differences: the easiest multidigit problems were easier than even the majority of the single digit problems. The difficulty structure of the single digit problems was very similar for two different multiplication tasks with different time limits. The data of both tasks are generally in line with the computational efficiency model. The tie effect, and especially the fact that the tie effect is stronger in the speed task imply a role for memory strength too. We found no support for the network interference model.

For multidigit problem solving, predictors related to the base ten structure of our numerical system (10, 100, 1000), the number of operations needed to solve the problem, and the presence of ones, twos and the requirement of a carry predicted problem difficulty best. Since multiplication is one of the four basic operations, and multidigit multiplication builds on single digit multiplication but is not the same, it is recommended that future studies on multiplication proficiency also look at multidigit multiplication.

Appendix A

The *Math Garden* computer adaptive technology works in an iterative process consisting of four steps: (1) problem selection based on child ability and problem difficulty estimates, (2) computing the expected score, (3) obtaining the score by the child solving the problem and (4) updating the child ability and problem difficulty estimates based on the difference between the expected and obtained score.

In step 1, a suitable problem is selected for a child. This selection is based on the past performance of both the player and the problems (by other players solving the same problem). This past performance has yielded an ability estimate of the player and difficulty estimates of the problems. A problem is selected for which the child has an expected probability of on average .75 (SD = .1) of answering it correctly within the time limit, according to the Rasch model shown in Eq. (1):

$$P(X_{ij} = 1|\theta_j) = \frac{e^{\theta_j - \beta_i}}{1 + e^{\theta_j - \beta_i}}. \quad (1)$$

In this equation P = probability, X = the accuracy (1 = correct, 0 = incorrect) of the answer of player j with ability estimate θ_j on item i with difficulty β_i; this probability increases with increasing player ability and decreases with increasing problem difficulty.

Then, in step 2, the expected score is estimated. While most scoring systems are based on accuracy only, Math Garden uses a new scoring system that incorporates both accuracy and speed: the High Speed High Stakes principle (HSHS; Maris & Van der Maas, 2012). The HSHS principle entails that the score is 1 for correct answers and −1 for incorrect answers, multiplied by the proportion of remaining time. The score is thus largest, positively or negatively, when an answer is given fast, and linearly decreases/increases to 0 as the time approaches the time limit. The principle is illustrated in Fig. 1. E.g., if an answer is given immediately, the score is +1 (correct answer) or −1 (incorrect answer). If an answer is given with 75% of the time remaining, the score is 0.75 (correct) or −0.75 (incorrect). All values between −1 and 1 are possible. With this scoring rule, fast correct responses are rewarded but fast incorrect guesses are punished.

The child's expected score on the problem is estimated with the extended Rasch model, presented in Maris and Van der Maas (2012). This

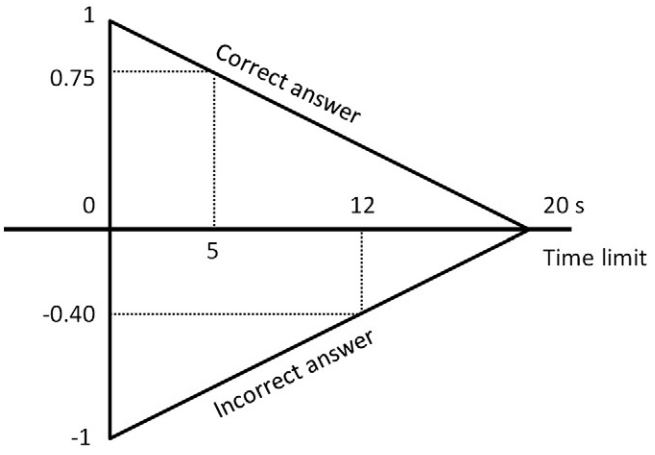

Fig. 1. Illustration of the High Speed, High Stake principle. With a time limit of 20 s, a correct answer after 5 s yields a score of 0.75; an incorrect answer after 12 s yields a score of −0.40.

model is shown in Eq. (2):

$$E(S_{ij}|\theta_j) = \frac{e^{2(\theta_j-\beta_i)}+1}{e^{2(\theta_j-\beta_i)}-1} - \frac{1}{\theta_j-\beta_i}. \quad (2)$$

In this equation, E = expected, S = score, θ_j = ability of player *j*, and β_i = difficulty of problem *i*. The outcome of the Eq. (2) is thus the expected score of player *j* with ability estimate θ_j on item *i* with difficulty estimate β_i. As Eq. (2) shows, this expected score is based on the difference between β_i and θ_j: this expected score ranges from close to −1 (when problem difficulty is far higher than child ability), through 0 (when both are equal), to almost 1 (when child ability is far higher than problem difficulty).

In step 3, the child solves the problem and the actual score is obtained, based on the speed and accuracy of the answer using the HSHS principles as displayed in Fig. 1. Then step 4 is executed, a new feature in *Math Garden* technology: an on the fly item calibration system. This means that the child's obtained score on an item is used to update the estimate of the child's ability and the difficulty of the item. The expected score derived from Eq. (2) is compared to the actual score that the child obtained. Then, following a procedure invented by Elo (1978) and used in competitive chess to rank players, the child's ability is adjusted upward if the child scored higher than expected and downward if the child scored lower than expected; the larger the discrepancy between observed and expected score, the larger the adjustment. In a similar fashion, the problem difficulty is adjusted downward if the child scored higher than expected (apparently the problem was easier than previously thought) and upward if the child scored lower than expected. Then the cycle starts again: the updated estimates are used to select the next problem. This way, estimates of child ability and problem difficulty are obtained reliably with an on the fly algorithm that enables adaptive problem selection such that children only solve problems that are suitable for their level of development.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Ashcraft, M. H., & Christy, K. S. (1995). The frequency of arithmetic facts in elementary texts: Addition and multiplication in grades 1–6. *Journal for Research in Mathematics Education*, *26*, 396–421. http://dx.doi.org/10.2307/749430.

Ashcraft, M. H., & Guillaume, M. M. (2009). Mathematical cognition and the problem size effect. In H. R. Brian (Ed.), *Psychology of Learning and Motivation*, *51*. (pp. 121–151). Academic Press.

Baroody, A. J. (1985). Mastery of basic number combinations: Internalization of relationships or facts? *Journal for Research in Mathematics Education*, *16*, 83–98. http://dx.doi.org/10.2307/748366.

Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D. (2009). Mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, *32*, 1–29.

Butterworth, B., Marchesini, N., & Girelli, L. (2003). Multiplication facts: Passive storage or dynamic reorganization? In A. J. Baroody, & A. Dowker (Eds.), *The development of arithmetical concepts and skills* (pp. 187–202). Mahwah, NJ: Lawrence Erlbaum.

Campbell, J. D., Dowd, R., Frick, J., McCallum, K., & Metcalfe, A. S. (2011). Neighborhood consistency and memory for number facts. *Memory & Cognition*, *39*, 884–893. http://dx.doi.org/10.3758/s13421-010-0064-x.

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

Campbell, J. I. D. (1987). Network interference and mental multiplication. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 109–123.

Campbell, J. I. D., & Graham, D. J. (1985). Mental multiplication skill: Structure, process, and acquisition. *Canadian Journal of Psychology*, *39*, 338–366. http://dx.doi.org/10.1037/h0080065.

Campbell, J. I. D., & Gunter, R. (2002). Calculation, culture, and the repeated operand effect. [Article]. *Cognition*, *86*, 71–96. http://dx.doi.org/10.1016/s0010-0277(02)00138-5.

De Brauwer, J., & Fias, W. (2009). A longitudinal study of children's performance on simple multiplication and division problems. *Developmental Psychology*, *45*, 1480–1496. http://dx.doi.org/10.1037/a0015465.

De Brauwer, J., Verguts, T., & Fias, W. (2006). The representation of multiplication facts: developmental changes in the problem size, five, and tie effects. *Journal of Experimental Child Psychology*, *94*, 43–56. http://dx.doi.org/10.1016/j.jecp.2005.11.004.

De Vos, T. (2010). *Manual tempotoets automatiseren.* Amsterdam: Boom Testuitgevers.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics.* New York: Oxford University Press.

De Visscher, A., & Noël, M. P. (2014). The detrimental effect of interference in multiplication facts storing: Typical development and individual differences. *Journal of Experimental Psychology: General*, *143*, 2380. http://dx.doi.org/10.1037/xge0000029.

Domahs, F., Delazer, M., & Nuerk, H. C. (2006). What makes multiplication facts difficult. *Experimental Psychology (formerly Zeitschrift für Experimentelle Psychologie)*, *53*, 275–282. http://dx.doi.org/10.1027/1618-3169.53.4.275.

Dowker, A. (2005). *Individual differences in arithmetic: Implications for psychology, neuroscience, and education.* New York: Psychology Press.

Elo, A. (1978). *The rating of chess players, past and present.* New York: Arco Publishers.

Freudenthal, H. (1991). *Revisiting mathematics education.* Dordrecht, the Netherlands: Kluwer.

Grabner, R. H., Ischebeck, A., Reishofer, G., Koschutnig, K., Delazer, M., Ebner, F., & Neuper, C. (2009). Fact learning in complex arithmetic and figural-spatial tasks: The role of the angular gyrus and its relation to mathematical competence. *Human Brain Mapping*, *30*, 2936–2952. http://dx.doi.org/10.1002/hbm.20720.

Hecht, S. A. (2006). Group differences in adult simple arithmetic: Good retrievers, not-so-good retrievers, and perfectionists. *Memory and Cognition*, *34*, 207–216. http://dx.doi.org/10.3758/BF03193399.

Hickendorff, M. (2013). The effects of presenting multidigit mathematics problems in a realistic context on sixth graders' problem solving. *Cognition and Instruction*, *31*, 314–344. http://dx.doi.org/10.1080/07370008.2013.799167.

Imbo, I., Duverne, S., & Lemaire, P. (2007a). Working memory, strategy execution, and strategy selection in mental arithmetic. *The Quarterly Journal of Experimental Psychology*, *60*, 1246–1264. http://dx.doi.org/10.1080/17470210600943419.

Imbo, I., & Vandierendonck, A. (2008). Practice effects on strategy selection and strategy efficiency in simple mental arithmetic. *Psychological Research*, *72*, 528–541. http://dx.doi.org/10.1007/s00426-007-0128-0.

Imbo, I., Vandierendonck, A., & Rosseel, Y. (2007b). The influence of problem features and individual differences on strategic performance in simple arithmetic. *Memory & Cognition*, *35*, 454–463. http://dx.doi.org/10.3758/bf03193285.

Imbo, I., Vandierendonck, A., & Vergauwe, E. (2007c). The role of working memory in carrying and borrowing. *Psychological Research*, *71*, 467–483. http://dx.doi.org/10.1007/s00426-006-0044-8.

Janssen, J., Scheltens, F., & Kraemer, J. -M. (2005). *Leerling- en onderwijsvolgsysteem rekenen-wiskunde [Student and education tracking system arithmetic-mathematics].* Arnhem, the Netherlands: Cito.

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

Koshmider, J. W., & Ashcraft, M. H. (1991). The development of children's mental multiplication skills. *Journal of Experimental Child Psychology*, *51*, 53–89. http://dx.doi.org/10.1016/0022-0965(91)90077-6.

Krueger, L. E. (1986). Why $2 \times 2 = 5$ looks so wrong: On the odd-even rule in product verification. *Memory & Cognition*, *14*, 141–149. http://dx.doi.org/10.3758/bf03198374.

LeFevre, J. -A., & Liu, J. (1997). The role of experience in numerical skill: Multiplication performance in adults from Canada and China. *Mathematical Cognition*, *3*, 31–62.

LeFevre, J. A., Bisanz, J., Daley, K. E., Buffone, L., Greenham, S. L., & Sadesky, G. S. (1996). Multiple routes to solution of single-digit multiplication problems. *Journal of Experimental Psychology-General*, *125*, 284–306. http://dx.doi.org/10.1037/0096-3445.125.3.284.

Lemaire, P., & Fayol, M. (1995). When plausibility judgments supersede fact retrieval: The example of the odd-even effect on product verification. *Memory & Cognition*, *23*, 34–48. http://dx.doi.org/10.3758/bf03210555.

Lemaire, P., & Siegler, R. S. (1995). Four aspects of strategic change: Contributions to children's learning of multiplication. *Journal of Experimental Psychology-General*, *124*, 83–97.

Lochy, A., Seron, X., Delazer, M., & Butterworth, B. (2000). The odd-even effect in multiplication: Parity rule or familiarity with even numbers? *Memory & Cognition*, *28*, 358–365. http://dx.doi.org/10.3758/bf03198551.

Mabbott, D. J., & Bisanz, J. (2003). Developmental change and individual differences in children's multiplication. *Child Development*, *74*, 1091–1107. http://dx.doi.org/10.1111/1467-8624.00594.

Maris, G., & Van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*, 615–633. http://dx.doi.org/10.1007/s11336-012-9288-y.

Moeller, K., Huber, S., Nuerk, H. C., & Willmes, K. (2011). Two-digit number processing: Holistic, decomposed or hybrid? A computational modelling approach. *Psychological Research*, *75*, 290–306. http://dx.doi.org/10.1007/s00426-010-0307-2.

Mulligan, J., & Mitchelmore, M. (2009). Awareness of pattern and structure in early mathematical development. *Mathematics Education Research Journal*, *21*, 33–49. http://dx.doi.org/10.1007/bf03217544.

Parkman, J. M. (1972). Temporal aspects of simple multiplication and comparison. *Journal of Experimental Psychology*, *95*, 437.

Rickard, T. C., Healy, A. F., & Bourne, L. E. (1994). On the cognitive structure of basic arithmetic skills: Operation, order, and symbol transfer effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1139.

Robert, N. D., & Campbell, J. I. D. (2008). Simple addition and multiplication: No comparison. *European Journal of Cognitive Psychology*, *20*, 123–138. http://dx.doi.org/10.1080/09541440701275823.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*, 135–146.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

Sherin, B., & Fuson, K. (2005). Multiplication strategies and the appropriation of computational resources. *Journal for Research in Mathematics Education*, *36*, 347–395. http://dx.doi.org/10.2307/30035044.

Siegler, R. S. (1988). Strategy choice procedures and the development of multiplication skills. *Journal of Experimental Psychology: General*, *117*, 258–275.

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science*, *14*, 237–250. http://dx.doi.org/10.1111/1467-9280.02438.

Siegler, R. S. (2007). Cognitive variability. *Developmental Science*, *10*, 104–109.

SLO (2009). Leerlijnen rekenen/wiskunde [learning program arithmetic/mathematics]. Retrieved from www.tule.slo.nl

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (5th ed.). Boston: Pearson Education.

Thevenot, C., Barrouillet, P., & Fayol, M. (2001). Algorithmetic solution of arithmetic problems and operands-answer associations in long-term memory. *The Quarterly Journal of Experimental Psychology*, *54A*, 599–611.

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

Verguts, T., & Fias, W. (2005a). Interacting neighbors: A connectionist model of retrieval in single-digit multiplication. *Memory & Cognition*, *33*, 1–16. http://dx.doi.org/10.3758/bf03195293.

Verguts, T., & Fias, W. (2005b). Neighbourhood effects in mental arithmetic. *Psychology Science*, *47*, 132.

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