Fast and slow strategies in multiplication

Hofman, A.D.; Visser, I.; Jansen, B.R.J.; Marsman, M.; van der Maas, H.L.J.

DOI
10.1016/j.lindif.2018.09.007
Publication date
2018
Document Version
Final published version
Published in
Learning and Individual Differences
License
Article 25fa Dutch Copyright Act

Learning and Individual Differences 68 (2018) 30–40

Contents lists available at ScienceDirect

Learning and Individual Differences

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

Fast and slow strategies in multiplication☆

Abe D. Hofman*, Ingmar Visser, Brenda R.J. Jansen, Maarten Marsman, Han L.J. van der Maas

University of Amsterdam, Netherlands

ARTICLE INFO

Keywords:
Item response theory
Response times
Multiplication
Strategies

ABSTRACT

In solving multiplication problems, children use both fast retrieval-based processes and slower computational processes. In the current study, we explore the possibility of disentangling these strategies using information contained in the observed response latencies using a method that is applicable in large data sets.

We used a tree-based item response-modeling framework (De Boeck & Partchev, 2012) to investigate whether the proposed qualitative distinctions in fast and slow strategies can be detected. This so-called fast-slow model was applied to responses to a set of multiplication items, totalling more than 180,000 responses, collected in an online computer-adaptive training environment for mathematics.

Parameters describing person characteristics (ability) and item characteristics (easiness) are estimated with the model. Both item and person characteristics differed between fast and slow processes and match predictions from substantive models of multiplication. Moreover, the parameters allowed us to describe the fast and slow strategies in more detail. Results emphasize the utility of the fast-slow model in the detection of strategies in multiplication but also in other areas of cognition and learning where strategies are expected.

1. Introduction

The concept of strategy is central in the study of human problem solving. Important aspects of problem solving behavior such as accuracy, duration, and type of errors, are due to the choice of the solution strategy. For instance, in solving arithmetic items, people may use either retrieval from memory or a computational strategy (Ashcraft & Guillaume, 2009; Dowker, 2005; LeFevre et al., 1996), where the former typically requires less time than the latter. In the case of basic multiplication (for example single-digit problems), detailed models for the retrieval process exist (Geary, Widaman, & Little, 1986; Verguts & Fias, 2005), and several models for computational strategies have been developed as well (Imbo, Vandierendonck, & Rosseel, 2007; Lemaire & Siegler, 1995). These models make different predictions about item difficulty and solution time (van der Ven, Straatemeier, Jansen, Klinkenberg, & van der Maas, 2015).

When measuring arithmetic ability by using psychometric tests, such as in IQ tests, individual differences in strategy choice are usually not taken into account. Arithmetic ability is ultimately tested by counting the number of correct items that participants solve in any particular test (e.g., Aunola, Leskinen, Lerkkanen, & Nurmi, 2004; Liu, Wilson, & Paek, 2008). Different patterns of response times and errors are hence ignored when the aim is to compare individuals on a scale of arithmetic ability. Using the number of correct responses may be warranted when testing and comparing test takers, but may be inappropriate when concerned with studying development and understanding ability differences. In the latter case, different qualitative processes or strategies should be considered.

In spite of the importance of the strategy concept, detecting strategies is still a major challenge in many areas of cognitive science. Verbal reports and neural imaging features are both correlated with strategy choice (Jost, Beinhoff, Hennighausen, & Rösler, 2004; Price, Mazzocco, & Ansari, 2013; Tenison, Fincham, & Anderson, 2014), but both also have pitfalls as strategy indicators. Verbal reporting, the most commonly accepted method of strategy detection, may interfere with the solution process and bias strategy choice (Kirk & Ashcraft, 2001; Reed, Stevenson, Broens-Paffen, Kirschner, & Jolles, 2015). Another important problem with using verbal reports for detecting strategy choice is that it is time-consuming and thus not feasible in combination with large scale automatic assessment of arithmetic abilities, which is very common nowadays. The latter problem also applies when using neural patterns to identify strategy choice. A third approach, whereby strategies are assessed through a combination of latencies and accuracy, is more promising. The utility of response times, obtained with large-scale computer-based assessment, has already been demonstrated for the detection of individual differences in reading literacy and problem

☆ Funding by NWO (The Netherlands Organisation for Scientific Research), grant number 406-11-163.
* Corresponding author at: University of Amsterdam, Department of Psychological Methods, Nieuwe Achtergracht 129-B, Amsterdam 1018 WT, Netherlands.
E-mail address: a.d.hofman@uva.nl (A.D. Hofman).

https://doi.org/10.1016/j.lindif.2018.09.007
Received 28 February 2017; Received in revised form 20 September 2018; Accepted 24 September 2018
1041-6080/

solving (Goldhammer et al., 2014). For the detection of strategies we use a different approach. This approach will be applied to developmental data on multiplication skills but applications to other cognitive skills are possible as well (see Coomans, Hofman, Brinkhuis, van der Maas, & Maris, 2016), as long as strategies are associated with diverging patterns of accuracy and response times.

An important developmental trend in learning multiplication can be described by changes in strategy choice. Initially children will apply various slower computational strategies (Freudenthal, 1991). Over time, these computations become more sophisticated (Lemaire & Siegler, 1995). Through practicing multiplication, children will build up a network of associations between numbers. When this network is sufficiently strong, children will be able to confidently retrieve answers to items, and will tend to use faster retrieval from this network instead of a slower computational strategy (Siegler, 1988). This development from computation to automaticity is in line with the more general theory on skill acquisition (Ackerman, 1988; Ackerman & Cianciolo, 2000). Children with learning difficulties do not show this typical transition from computational to retrieval strategies (De Smedt, Holloway, & Ansari, 2011; De Visscher & Noël, 2014). After years of practice, adults will rely predominantly on memory retrieval for single digit multiplication (LeFevre et al., 1996). Hence, the largest divide in strategy choice is whether children and adults use a retrieval strategy or a computational strategy.

In this paper, we investigate whether the fast-slow model (DiTrapani, Jeon, De Boeck, & Partchev, 2016; Partchev & De Boeck, 2012) allows for automatic analyses of strategy use. The fast-slow model is based on splitting the data into fast and slow responses and estimating separate parameters for each of the processes. A third process, based on the response latencies, indicates choice for the fast or slow process. This approach is intermediate between the purely psychometric approach of fitting IRT models to capture multiplication ability on a single latent trait (e.g., Aunola et al., 2004; Liu et al., 2008) and the purely cognitive approach of using computational models to predict response accuracy based on problem characteristics and strategies (partial abilities) (e.g., de la Torre & Douglas, 2008).

We will first introduce the fast-slow model, derive predictions for the case of multiplication, and then apply the model to a data set. This data set includes a large set of responses, both accuracy data and response times, collected with a popular Dutch online adaptive learning environment for mathematics; the Math Garden (Klinkenberg, Straatemeier, & van der Maas, 2011; Straatemeier, 2014).

1.1. The fast-slow model

The fast-slow model is a tree-based item response theory (IRT) model (De Boeck & Partchev, 2012). The rationale of this model is that responses are governed by one of two processes, one fast and one slow, that can be separated by an additional observed variable, in this case the (recoded) response times. The response times are recoded to either fast (1) or slow (0), which serves as an approximation of the underlying process and is modeled as a latent speed dimension. This tree model can be formulated as follows, assuming that a (unidimensional) Rasch model (Rasch, 1960) holds in dimension d, where $d = 1,2$, or 3 and denotes the speed-, fast- and slow dimensions, respectively. In these dimensions, the probabilities of respectively a fast response, a fast and correct response, and a slow and correct response are modeled using a Rasch model. In the Rasch model, the probability of a correct (or for the speed dimension a fast) response of a person p on an item i in dimension d is given by the logistic function:

$$P(x_{pid} = 1|\theta_{pd}, \beta_{id}) = \frac{exp(\theta_{pd} + \beta_{id})}{1 + exp(\theta_{pd} + \beta_{id})}, \tag{1}$$

where θ_{pd} denotes the ability of person p and β_{id} denotes the easiness of item i on dimension d. Hence, the full model has three sets of person parameters, and three sets of item parameters: θ_{p1} reflects the overall probability of a person to generate a fast response, θ_{p2} reflects the ability to give a fast and correct response, and θ_{p3} reflects the ability to give a slow and correct response. Likewise, item easiness parameters correspond to the probability that items are answered fast versus slow (β_{i1}, with a high β_{i1} indicating a high probability of a fast response), the probability of a correct response given that the response was fast (β_{i2}), and the probability of a correct response given that the response was slow (β_{i3}).[1] In line with De Boeck (2008), both $\boldsymbol{\theta_p} = (\theta_{p1},\theta_{p2},\theta_{p3})$ and $\boldsymbol{\beta_i} = (\beta_{i1},\beta_{i2},\beta_{i3})$ are treated as random variables with $\boldsymbol{\theta_p} \sim \mathcal{N}(\boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta})$ and $\boldsymbol{\beta_i} \sim \mathcal{N}(\boldsymbol{\mu_\beta}, \boldsymbol{\Sigma_\beta})$, constraining $\boldsymbol{\mu_\theta}$ to zero to identify the model (see Appendix A for a description of the model estimation procedure).

1.2. Qualitative differences in the fast-slow model

Within the fast-slow model, qualitative differences between fast and slow processes would be reflected by a different ordering of the item parameters, person parameters or both, in the fast compared to the slow component of the model. Hence, to test the hypothesis that the fast and slow processes differ qualitatively, the full fast-slow model with different item parameters for the fast and the slow process as well as different person parameters for the two processes is compared against three constrained versions of the model. This resulted in four different models: (1) the full fast-slow model, (2) constrained item parameters: i.e., $\beta_{fast} = \beta_{slow}$, (3) constrained person parameters: i.e., $\theta_{fast} = \theta_{slow}$, and (4) a baseline model in which both item and person parameters are constrained. If one, or both, constraints resulted in a worse model fit (in terms of prediction; see Section 2.2), this would support the notion that indeed qualitatively different processes were involved in the fast and the slow responses. However, from a measurement perspective different item parameters do not necessarily suggest that the person parameters are different, since these abilities could be highly correlated (the same holds for item parameters if person parameters are different).

Whenever a constraint was imposed we allowed for a difference in the overall mean and in the variances of the fast and slow item and/or person parameters. This reflects the idea that only a correlation between the fast and slow parameters that is significantly lower than one truly reflects a qualitatively different process. For example, if fast responses are more often correct than slow responses it does not necessarily suggest that slow and fast responses have distinct response processes. It may be that for slower responses, retrieval is simply more difficult. However, if for some persons or items the slow responses are more often (in)correct than the fast responses, thereby influencing the correlations of these parameters, this would indeed suggest that different response processes are involved.

1.3. Empirical predictions for a fast-slow model of multiplication processes

Given the observed qualitative differences between fast and slow strategies in multiplication (LeFevre et al., 1996), the full fast-slow model is expected to describe the data best. In this model, both item and person parameters have different estimates in the fast compared to the slow process. It is expected that the fast process will more often match fact retrieval and that the slow process will more often match computational strategies. If this is the case, some parameter estimates of the processes should relate differently to item and person characteristics. Finding that these relations match common findings in the multiplication literature would support the claim that the fast-slow model is a useful method to identify strategies in multiplication at the individual level.

[1] For the readability of the remainder of the paper, we refer to θ_{p1}, θ_{p2} and θ_{p3} as θ_{speed}, θ_{fast}, and θ_{slow}, and to β_{i1}, β_{i2} and β_{i3} as β_{speed}, β_{fast}, and β_{slow}.

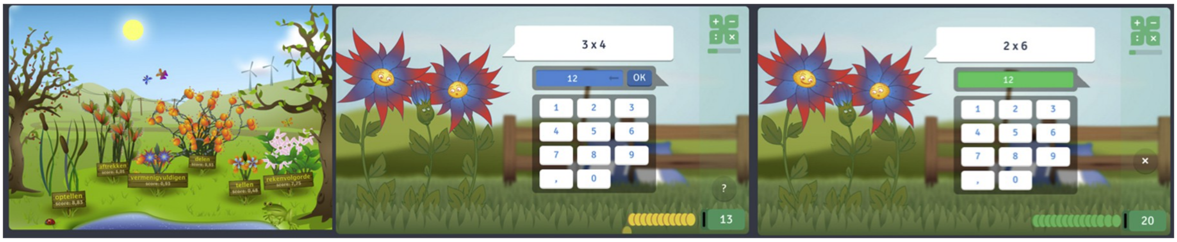

Fig. 1. Three screen shots of the Math Garden. The left panel shows the garden page where each plant represents a game measuring a different aspect of mathematics. The middle panel shows an open format multiplication item, where children use a numeric keypad to provide a response. The coins at the bottom represent points, and players lose one for each second that they do not provide a response. The coins turn green (or red) in case of a correct (incorrect) response, and are added to (subtracted from) the total (right-panel). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

1.3.1. *Item effects*

We focus on three prominent effects of item characteristics that are reported in the literature on simple multiplication; the problem-size effect, the tie-effect and effects of special operands. These characteristics are associated with systematic differences in accuracy and response times between items. Models of retrieval and computation strategies in simple multiplication have coined different explanations for these differences.

1) The problem size effect (Ashcraft & Guillaume, 2009) refers to the fact that items with larger problem sizes are more difficult than items with smaller problem sizes. In retrieval based models, this effect is explained by less frequent practice with items with large operands and therefore weak connections in the memory network (Ashcraft, 1995). According to models of computational strategies this effect is due to the additional steps necessary for computing the answer (LeFevre et al., 1996; van der Ven et al., 2015). Thus, both the probability of a correct, fast response (β_{fast}) and the probability of a correct, slow response (β_{slow}) is expected to relate negatively with problem size. Concerning overall speed, it is expected that items with larger problem size are more often solved by means of slow, computational processes than by retrieval. Hence, it is expected that the probability of a fast response versus slow response (β_{speed}) is negatively related to problem size.
2) The tie-effect (De Brauwer, Verguts, & Fias, 2006; Miller, Perlmutter, & Keating, 1984) implies that ties (items with an equal operand; e.g., 7 × 7) are easier than other items. This effect is explained by more practice and easier storage in retrieval based models. Hence, a tie-effect is expected in the fast process. More specifically, the probability of a correct response given that the response was fast (β_{fast}) is expected to be higher for tie problems than for non-tie problems. Models of computational strategies do not predict a tie-effect since the computations involved in ties are the same as in non-tie items. Hence, no tie-effect is expected in the slow process, which is expected to be associated with computational strategies. So, the probability of a correct response given that the response was slow (β_{slow}) is not expected to differ between tie problems and non-tie problems. Concerning overall speed, it is expected that tie problems are more often solved by retrieval than by computational processes. Hence, the tie effect is expected to be positively related to the probability of a fast versus slow response (β_{speed}).
3) The special operands effect refers to the finding that items with 1, 2, 5 or 9 as operands are easier than other items (Lemaire & Siegler, 1995). This effect follows from easier computations according to computational accounts, but is not predicted in models of retrieval. Hence, the probability of a correct response given that the response was fast (β_{fast}) is not expected to differ between multiplication problems with and without special operands, whereas the probability of a correct response given that the response was slow (β_{slow}) is expected to be higher for multiplication problems that include special operands than for problems that do not include such operands. Finally, it is expected that problems with special operands are more often solved with retrieval than with computational strategies, implying a higher probability of a fast versus slow response (β_{speed}) for problems with special operands than for problems without such operands.

1.3.2. *Person effects*

As explained in the introduction the development of simple multiplication ability involves a shift from computational strategies to retrieval. This shift is expected to be reflected in a higher number of fast (and correct) responses for older compared to younger children. The increase in faster responses would result in a positive correlation between age and θ_{speed}. The increase in correct responses would result in a positive relation between age and the probability of a correct fast response (θ_{fast}) as well as a correct slow response (θ_{slow}). A gender effect on speed is expected as well, due to individual differences in response styles. In addition and subtraction problems, boys provided more retrieval responses than girls, while girls were more likely to count with their fingers (Carr & Jessup, 1997). Although addition and subtraction are different domains, we assume that these response styles may be active in multiplication as well. Hence, it is expected that boys have a higher probability to respond fast (i.e., a higher θ_{speed}) than girls. No gender effects are expected for the accuracy in executing either the fast or slow process. In other words, no gender effects are expected for θ_{fast} and θ_{slow}.

2. Methods

2.1. *Data sets: items and participants*

Data are collected with the website Math Garden (see Fig. 1). Math Garden is an online adaptive learning environment for learning basic arithmetic, that is currently used by more than 200,000 children involving more than 1500 schools in the Netherlands. Math Garden provides a valuable data set, including accuracies and response times of a large group of children, on a large set of multiplication items. Children (or their parents) who indicated that they did not want to participate in scientific research conducted with Math Garden were excluded from the analyses. This research study was approved by the local Ethics Committee.

In the multiplication game, children are given 15 items that must be solved using the virtual numerical keypad, each with a time limit of 20 s. The time is visualized by disappearing coins (one is lost each second that they do not provide a response). If a correct response is given the coins are added to a total, whereas the coins are subtracted if the response is incorrect. This explicit High Speed High Stakes (HSHS) scoring-rule informs the users how to weigh speed and accuracy (Klinkenberg et al., 2011; Maris & Van der Maas, 2012). This ensures that children perceive some time-pressure since they are motivated to provide fast responses, but they are discouraged from guessing due to the penalty of a fast but incorrect response. When a child does not know the answer (s)he can best wait the full 20 s. To prevent such waiting times the child can also use the question-mark button, in which case (s) he does not win or lose any coins. In our analyses, the question-mark responses were labeled as incorrect, and the percentage of question-mark responses was included as a person characteristic in secondary analyses.

With the HSHS scoring rule in the Math Garden the person and item estimates can be updated after each response.[2] Based on these estimates relevant items were selected for the child at each time point, such that children were expected to provide 60%, 75% or 90% of correct responses when playing at the hard, medium or easy difficulty level (for more details see Jansen, Hofman, Savi, Visser, & van der Maas, 2016).

For this study, we selected responses collected between June 1, 2011 and June 1, 2015 to items belonging to the multiplication tables from two up to nine (64 items in total). We discarded the first 90 responses that each child made to allow children to become acquainted with the task. Furthermore, because data were collected longitudinally and abilities tend to change over time we selected a time frame for a single assessment of a child's ability. This time-frame must contain sufficient data but should also be small enough to ensure a relatively stable ability, and was fixed to one week. Additionally, in order to set a minimum number of responses for this time frame, we selected data of children who completed at least 30 items within one week.[3] Only the child's first response to an item was selected (multiple responses for the same item within the time frame are possible). In total 180,651 responses of 3551 children were analyzed. The adaptive item selection resulted in 21% of missing responses. These missing responses are missing by design, and can be seen as missing at random (MAR) since the missingness is conditional on the estimated ability (Eggen & Verhelst, 2011; Mislevy & Wu, 1996; Rubin, 1976).

In order to apply the model, the response times needed to be dichotomized into fast or slow categories. In our analyses, we used three different approaches based on a median split: (1) a split on the overall response times distribution; (2) a within person split allocating 50% of the responses of each person to either fast or slow and (3) a within item split allocating 50% of the responses to each item to either fast or slow. The first split captures both person and item differences in speed, whereas the person (item) split only captures differences between items (persons) in speed respectively. A comparison of the results of each of these split-methods provides information on the robustness of the results (see Appendix B).

2.2. Model comparison

To compare the fit of the four models we used the DIC statistic (a Bayesian version of the AIC statistic; Gelman, Carlin, Stern, & Rubin, 2014, p. 172) and three statistics based on a cross-validation procedure. We performed this additional cross validation procedure since it enables a straightforward model comparison. The model with the best prediction of the responses in the test set wins. For each person, data from one response were selected for the test data. The remainder of the data were used to estimate (train) the model parameters, and the estimated models were subsequently used to predict the test data. This approach naturally prevents over-fitting the data with overly-complex models. The test data formed between 1.4% and 3.0% of the total data in the different data sets but was still fairly large as, despite including one response per person, a large number of persons was included. Model predictions were based only on accuracy as the models did not differ in their analyses of response times.

Three cross-validation statistics were used, all three based on the deviation between the observed and the predicted response: the prediction accuracy (ACC), the root mean squared error (RMSE) and the log-likelihood (LL; Pelánek, 2015; see Appendix A for a detailed description). In both RMSE and LL, the continuous prediction of the probability of a correct response is analyzed. This results in a more detailed model comparison than the ACC, while the ACC provides a simpler interpretation of the goodness-of-fit. When interpreting the ACC and the LL, higher (less negative) values indicate better fit, while for the RMSE and the DIC lower values indicate better fit.

3. Results

Since the results of the model comparisons were similar across the various dichotomizations, we limit the results section to the analyses from the data set where fast or slow was defined by the overall medium split (see Appendix B for the results based on the other split methods).

3.1. Data Description

The RT distribution of the data set is presented in the left-panel of Fig. 2. The median response time (RT) was 6.22 s; 59% ($SE = 0.0016$) of the fast responses and 62% ($SE = 0.0016$) of the slow responses were correct. The lower percentage for the fast responses was related to the higher proportion of fast question-mark responses: 33% and 11% respectively for fast and slow responses. This is also shown by the relationship between RT and the probability of a question-mark response, plotted in the right-panel of Fig. 2.

3.2. Model comparison

To estimate the model parameters we used 3000 iterations and a burn-in of 100. Since some high auto-correlations were found we used every third iteration for the MAP estimates of the model parameters. Table 1 shows the fit measures for the estimated models. In line with our hypothesis, the results indicated that the model with separate item parameters and separate person parameters for the fast and slow process – the full fast-slow model – provided a better fit that any of the constrained models (see Table 1). This showed that qualitatively different processes were involved in the fast compared to the slow processes.[4]

These results indicate that the response times (split into fast and slow) distinguished between two qualitatively different response processes, both with respect to item and person parameters. In the following sections, we will further describe the estimated parameters, and thereby test our hypotheses concerning the relation between item characteristics (problem size, ties, special operands) and item parameters as well as the relation between person characteristics (age, gender) and person parameters in order to investigate whether differences between the fast and slow strategies can be explained by retrieval and computational models of multiplication.

3.2.1. Effects of item characteristics on item parameters

After having established the superior fit of the full fast-slow model, the next step consisted of analyses in which item parameters were regressed on different item characteristics. We intended to replicate the effects of problem-size, tie and effects of special operands and it was

[2] See Maris and Van der Maas (2012) for a detailed description of the psychometric properties of the HSHS rule and Klinkenberg et al. (2011) and Straatemeier (2014) for a description of how the parameters are estimated using an Elo (Elo, 1978) algorithm.

[3] It was possible to make different choices for selecting data. However, using different inclusion criteria yielded comparable results, see Appendix B.

[4] To validate our results, we also fitted the four models on a second data set that included a larger subset of items (150). The model comparison again selected the full model as the best model. This supports the validity and the robustness of our results. See Appendix B for the fit statistics of all models and a description of the second data set.

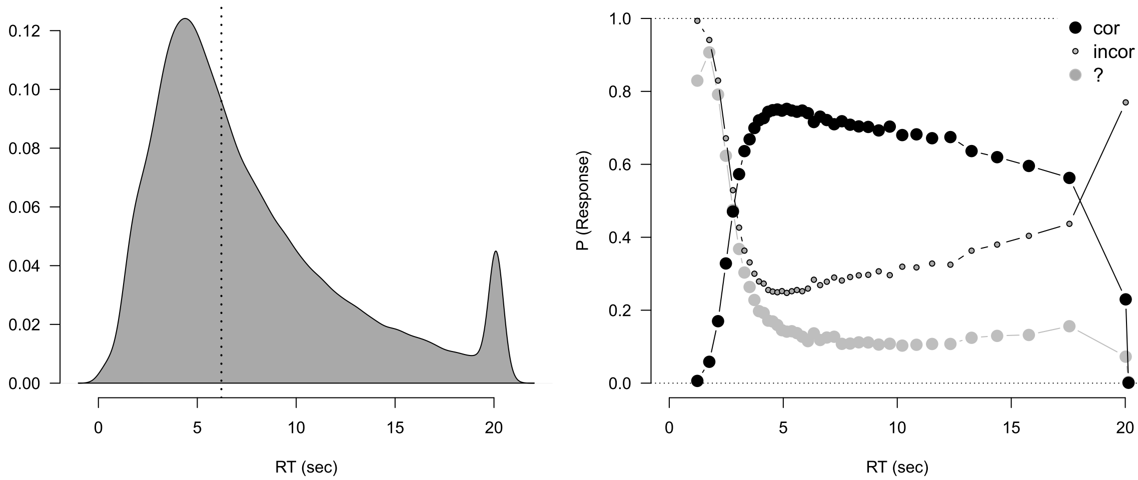

Fig. 2. Data description. The left-panel shows the RT distribution. The vertical line indicates the median. The peak around 20 s is caused by the response deadline. The right-panel describes the proportions of a correct, incorrect and question-mark response for the different observed response times. The question-mark response is counted as incorrect.

Table 1
Fit statistics for each of the four models.

Model	*DIC*	ACC_{CV}	$RMSE_{CV}$	LL_{CV}
Baseline model	168,836	0.763	0.399	−2498
$\theta_{fast} \neq \theta_{slow}$	166,932	0.767	0.399	−2507
$\beta_{fast} \neq \beta_{slow}$	169,443	0.768	0.400	−2535
Full model	163,923	0,779	0.393	−2427

Note. Results of the best fitting model are printed in bold. The baseline model refers to the fully constrained fast-slow model. The full model is the model that allows for both different person and item parameters in the fast and slow part. CV indicates that these fit statistics are based on cross validation.

expected that the effects would differ between the fast and slow process.[5]

In three separate regression models, we predicted the item scores reflecting 1) the probability of a correct response given that the response was fast β_{fast}, 2) the probability of a correct response given that the response was slow (β_{slow}), and 3) the probability of a fast response versus a slow response (β_{speed}). We used the BIC (Schwarz et al., 1978) for model selection, using a backward stepwise procedure.

Columns 2–4 from Table 2 show the regression coefficients that link the item characteristics to the accuracy in the fast process. In line with the predictions, we found a main effect of: (1) problem-size, indicating that items with larger problem size were more difficult than items with smaller problem size; and (2) ties, indicating that ties were easier than non-tie items. Unexpectedly, the effects of problems with special operands were significant as well, indicating that problems with two, five and nine were easier than other problems. These main effects explained in total 89.2% of the variance.

Columns 5–7 from Table 2 show the regression coefficients that link the item characteristics to the accuracy in the slow process. All results were in line with the predictions. Both a main effect of problem size and main effects of special operands were observed and the main effect of ties was not significant. The main effect of ties was even removed in the stepwise procedure. The effects resulted in an explained variance of 88.6%.

Columns 8–10 from Table 2 show the regression coefficients that link the item characteristics to the probability that an item is solved using the fast process as compared to the slow process. For the item speed parameter, a high β_{speed} indicated a high probability of a fast response. Thus, the negative effect of problem size in Table 2 shows that responses to items with large problem sizes were more often slow than responses to problems with smaller problem size, which matches the predictions. Also in line with expectations were the findings that responses to ties and items belonging to the two and five multiplication tables were often fast (see Table 2). The latter results indicate that these items were more often solved by retrieval rather than computational strategies. These effects explained in total 70.6% of the variance in the item speed parameters. Unexpectedly, there was no effect of special operand 9 on the probability of a fast response.

To conclude, the high explained variances indicate that the item parameters could be largely understood by these item characteristics. This supported the reliability of both the data and the model estimation. Results were largely in line with expectations, although an operand effect was unexpectedly observed in the fast process and there was no effect of special operand 9 on the probability of a fast response.

3.2.2. *Effect of person characteristics on person parameters*

In the second set of regression models we investigated whether person characteristics were differentially related to fast and slow abilities. For these analyses, we only included children between 6 and 11 years old (N = 4233; excluded 467), and children whose age matched their grade (excluded 417 children for whom their age deviated more than 1.5 years from the grade average). The average age of the selected children was 7.86 (SD = 1.04), and 33% were girls.

Again, three separate regression analyses were performed to predict the person estimates that were associated with 1) the probability of using a fast versus slow process (θ_{speed}), 2) the probability of being correct using a fast process (θ_{fast}), and 3) the probability of being correct using a slow process (θ_{slow}). All results, based on a stepwise backward procedure using BIC, are presented in Table 3. As expected we found main effects of age, indicating that with increasing age, children were more likely to give a fast response than a slow response and were more able in both the fast and the slow process. A gender effect was only expected for the probability of giving a fast response. Indeed, boys were more likely to give a fast response than girls and there were no gender effects for the probability of responding correctly in both processes. In fact, the effect of gender was excluded in the stepwise procedure of the analyses concerned. Finally, children with more question-mark responses were more likely to respond fast. More question-mark response was also related to a lower ability. This is a trivial results since question-mark responses are labeled as incorrect.

[5] To investigate these effects, we imputed the full original data set by generating a new set of responses based on the estimated model parameters. We analyzed the sum-scores over items for both fast and slow responses. This approach ensured that effects can be directly compared between different nodes.

Table 2
Regression of the item easiness parameters for fast and slow processes and speed (reflecting the probability of a fast response).

	Fast			Slow			Speed		
Predictor	B	SE	T-value	B	SE	T-value	B	SE	T-value
Intercept	2.851	0.374	7.632**	2.277	0.225	10.139**	0.034	0.167	0.204
Problem-size	−0.351	0.032	−11.044**	−0.218	0.019	−11.289**	−0.039	0.013	−3.002*
Tie	1.432	0.212	6.765**	ex	ex	ex	0.491	0.104	4.713*
Times 2	2.408	0.207	11.607**	1.152	0.125	9.224**	0.757	0.100	7.577**
Times 5	1.224	0.168	7.282**	0.762	0.101	7.543**	0.211	0.083	2.538*
Times 9	0.817	0.203	4.016**	0.476	0.123	3.876**	ex	ex	ex

Note. ex = excluded in the stepwise procedure.
* $p < 0.05$.
** $p < 0.001$.

Table 3
Regression person ability parameters for fast and slow processes and speed (reflecting the probability of a fast response).

	Fast			Slow			Speed		
Predictor	B	SE	T-value	B	SE	T-value	B	SE	T-value
Intercept	−0.038	0.019	−1.971	−0.023	0.012	−1.860	−0.173	0.021	−8.074*
Age	0.432	0.020	21.508*	0.267	0.013	20.631*	0.221	0.013	17.077*
Gender	ex	ex	ex	ex	ex	ex	0.198	0.026	7.478*
% ?	−1.184	0.020	−58.910*	−0.476	0.013	−36.817*	0.487	0.013	37.786*

Note. Boys are coded as 1 and girls as 0; %? = percentage of question-mark responses; ex = excluded in the stepwise procedure.
* $p < 0.001$.

However, this latter effect was smaller with slow compared to fast abilities. This highlights that the differences between the abilities measured by fast and slow responses can partly be explained by differences in how children relate to the question-mark answer option. These effects explain 54.9%, 36.7% and 27.3% of the variance for fast abilities, slow abilities and the speed dimension, respectively.

3.3. Exploring correlations between fast accuracy, slow accuracy and speed

In this last section, we explore the relations between speed and accuracy from an item and a person perspective.[6] The model comparison indicated that fast and slow item and person parameters are qualitatively different. However, the correlation between β_{fast} and β_{slow} was high ($r = 0.969$). The correlation between θ_{fast} and θ_{slow} was lower ($r = 0.778$). The lower correlation between person parameters might be explained by the smaller number of observations for the person parameters compared to the item parameters (which may have created more measurement error). Furthermore, higher variances in β_{fast} compared to β_{slow} were found ($\sigma_{\beta,fast} = 1.943$ and $\sigma_{\beta,slow} = 1.085$; Levene's test of equality of variance: $F(1, 62) = 30.07$, $p < 0.001$). The lower estimated variance in the slow process correspond to the results of Partchev and De Boeck (2012) and DiTrapani et al. (2016), and could suggest that there is more random variation (error), compared to structural variance (responses explained by the model parameters), in the slow responses. This might be caused by a mixture of different strategies within the slow responses, which could not be captured by the model.

Item speed and accuracy correlated positively, with correlations between β_{speed} and β_{fast} and β_{slow} of 0.837 and 0.739, see Fig. 3.

We observed two interesting results. First, the relationship between speed and accuracy showed an interesting pattern. A regression model with a breakpoint resulted in an explained variance of 83.8%, an increase of 13.5% compared to the explained variance of 70.4% of the linear regression model. Furthermore, the breakpoint could be confidently estimated at zero, indicated by a clear peak in explained variance compared to models with differently located non-zero breakpoints. This strongly suggested that, for items that were solved quickly ($\beta_{speed} > 0$), there was a strong relationship between speed and accuracy, whereas for items that were more often solved slowly this relation was absent. In line with results of the model comparison, this result signifies that fast strategies are qualitatively different from slow strategies.

Second, for persons, a different pattern was found. Negative correlations between person overall speed and fast and slow abilities were found: −0.125 and −0.033 respectively. Thus, in contrast to expectations, children that were faster were more often incorrect. To test whether the negative correlation was related to differences between children in question-mark usage we calculated separate correlations for children who provided less or more than 13% question-marks (median split). We found a correlation of 0.306 for children who used fewer question-marks, indicating that for these children, the faster children were more able than the slower children, see the blue line in the second panel of Fig. 3. This suggested that the negative relation is related to question-mark uses. Furthermore, all correlations were positive (min = 0.191 and max = 0.373) when children were grouped by question mark use from 0 to 90% in increments of 10%. To conclude, these results indicate that, when corrected for question-mark usage, children who are faster had higher fast and slow abilities.

4. Discussion

In this paper, we investigated whether the application of the fast-slow model (DiTrapani et al., 2016; Partchev & De Boeck, 2012) to the domain of multiplication was feasible by testing whether the model would result in the detection of two qualitatively different processes. The fast-slow model uses a split in response times to detect processes and results in parameter estimates for the items (probability of being solved with a fast versus slow process; probability that an item is solved correctly when a fast process is used and the probability that an item is solved correctly when a slow process is used) and in parameter estimates for persons (a person's tendency to use either the slow or the fast process and probability of responding correctly for both the slow and

[6] The presented results were stable under the different RT splits; the within item split to investigate person speed and the within person split to investigate item speed.

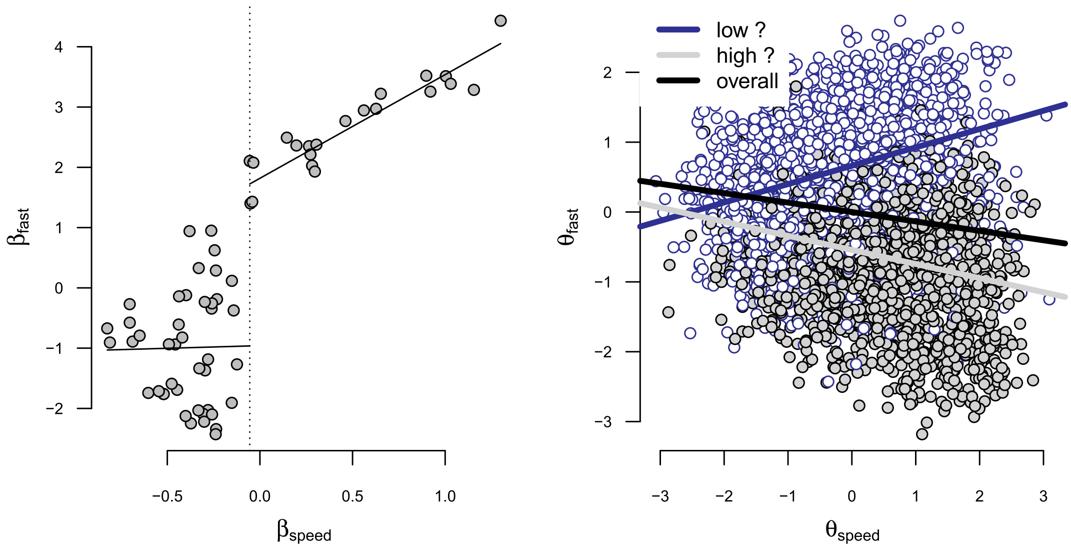

Fig. 3. Relationship between speed and fast accuracy for items and persons. Low ? and high ? denote the frequency of question mark usage.

the fast process). A successful application of the fast-slow model in the domain of multiplication would be indicative for the use of the fast-slow model in other domains of cognition and learning as well.

Indeed, the two detected processes differed in expected ways and matched fast retrieval versus slow computational processes of multiplication (Siegler, 1988). On the item side, the probability to solve items with a fast versus slow process decreased as problem size increased. Also, the probability to solve items correctly decreased with size in both the slow and the fast process. The tie effect was indeed observed for the fast but not the slow process and the tendency to solve items with the fast process was higher for tie than for non-tie problems. Contrary to expectations was the observation that the effect of most special operands (2, 5) was observed in both the fast and slow process. These effects had been expected in the slow process only. On the person side, older children (who are assumed to have more experience) provided more fast responses. Although older children can be faster in multiple ways, the results indicate that this developmental trend is partly due to a higher probability of a retrieval strategy for older compared to younger children. As expected, the probability of correct responses increased with age, in both the fast and slow process. Additionally, although boys and girls did not differ with respect to accuracy in executing the fast and slow processes – in line with the results of Carr and Jessup (1997) – boys provided more fast responses than girls. These results confirm that children's strategies for solving mental multiplication items can be disentangled using a split in observed response times, as is implemented in the fast-slow model.

The unexpected finding that the effect of special operands (2, 5 and 9) was comparable for the fast and slow process could be explained by having used rather crude methods to disentangle strategies. These methods may have allocated some retrieval responses as slow and some computational responses as fast, resulting in a lower power to find differential effects. Especially with special operands, the computational process can be fast, which may have resulted in accidentally assigning computational strategies to the fast process and may explain why special operands 2 and 5 also had a positive effect on the accuracy in the fast process. However, using various split methods do show consistent results. Further methodological improvements are possible with developing better ways of splitting response times as the most important one. Ideally, the data itself determines the classification into fast and slow processes, resulting in a more optimal classification of responses to strategies (DiTrapani et al., 2016).

Other limitations of the study are the following. First, we used only a single split, restricting the number of multiplication strategies to only two. LeFevre et al. (1996) showed that children report the use of various computational strategies (e.g., repeated addition, derived facts), rules and number series, next to fact retrieval. Moreover, fact retrieval may be required to execute the computational strategy of derived facts, showing that the fast and slow process are sometimes combined. Second, data are collected under specific settings, that is, children solved math problems to practice their skills in a natural classroom setting. Results may be different in a controlled lab setting, with different instructions and different consequences of performance (e.g., in a test situation).

Exploratory analyses showed that the application of the fast-slow model also provided additional information on children's solution strategies on multiplication items. In the application that was currently used for data collection (Math Garden), children are allowed to “escape” from an item by selecting a question mark. Some children are more inclined than others to provide a question-mark response, and exploratory analyses showed that those who do so have a different speed-accuracy trade-off than those who use the question-mark less often.

Hence, as described by Siegler (2007) and Van der Ven, Boom, Kroesbergen, and Leseman (2012), the application of the fast-slow model also shows that multiplication ability should be seen as a toolbox of different strategies, where both the ability of each child within a certain strategy and individual differences in strategy selection determine the observed performance. This study indicates that these processes, often studied in smaller and controlled experimental settings, also determine multiplication ability in a large-scale online learning platform, supporting the generalizability of the effects and the validity of the Math Garden.

4.1. Implications for education

This line of research may provide applied researchers, teachers and students with valuable information on strategies in multiplication, without using time-intensive methods such as verbal protocols. The application of the fast-slow model indicates for each response of a child whether the child used the fast or the slow strategy and whether the strategy was performed correctly. This enables tailored feedback about proficiency of strategies when learning multiplication, and thereby matches the aims for mathematics education. For instance, in the Netherlands education ultimately aims for both understanding multiplication concepts and memorization of the single-digit tables of multiplication (SLO, 2009). Ideally, a teacher receives the information on each child's tendency to respond fast and on proficiencies in the fast and the slow process from the online learning environment and can next adapt instructions. If a child is proficient in the slow but not the fast process, the teacher could introduce evidence-based interventions for promoting automaticity (see for example Kroesbergen & Van Luit, 2002). If however a child is proficient in the fast but not the slow process, this might indicate that the child has rote-learned various math facts but doesn't possess any successful backup strategies (see for a possible approach Kroesbergen, Van Luit, & Maas, 2004).

4.2. Future directions

It should be noted that the mixture of retrieval and computational processes underlying the responses in multiplication will depend on the testing conditions. In the Math Garden, items were selected to match children's ability, resulting in a mixture of different strategies. Presenting solely easy or hard items will change the mixture of strategies. Additionally, the test conditions were such that children perceived time-pressure. This evokes faster responses, and probably influences the strategies that were used (Hofman, Visser, Jansen, & van der Maas, 2015). Further research should investigate whether children's performances in high-stakes tests also depends on multiple processes. Additionally, next to response latencies, error types also contain information about the used strategy (Siegler, 1988). In a first minimal example, Coomans et al. (2016) already showed that fast errors in response to multiplication items were different from slow errors. Utilizing both response latency and error types could provide additional confidence in estimating the used strategy.

Finally, the fast-slow model may be applied to other domains in cognition and learning where the use of strategies that differ in speed is expected. For example, the fast-slow model can detect strategies in other arithmetic domains, decision making, memory, and spelling. Detection of strategies may not only inform theory-formation but also shed light on the desired instruction in children's education.

Appendix A. Model estimation and comparison

A.1. Estimation of the fast-slow model

We adopt a Bayesian approach to estimate the parameters of our fast-slow model, and wish to quantify our uncertainty about these parameters in a joint posterior distribution: i.e., $f(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_\beta|\text{data})$. To this aim, we need to specify a prior distribution for the population parameters $\{\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\theta, \boldsymbol{\Sigma}_\beta\}$. First, we specify a Jeffreys prior for the between dimension person covariance matrix $\boldsymbol{\Sigma}_\theta$ (Gelman et al., 2014, page 37)

$$f(\boldsymbol{\Sigma}_\theta) \propto |\boldsymbol{\Sigma}_\theta|^{-2},$$

where we assume that $\boldsymbol{\Sigma}_\theta$ is independent of $\{\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta\}$ a priori. Second, we constrain the off-diagonal elements from the between dimension item covariance matrix $\boldsymbol{\Sigma}_\beta$ to be zero and assign independent Jeffreys priors to the mean and variance for each dimension (Gelman et al., 2014, page 64), i.e.,

$$f(\mu_{\beta,\ d}, \quad \sigma^2_{\beta,\ d}) \propto \sigma^{-2}_{\beta,\ d}.$$

Constraining the off-diagonal elements of $\boldsymbol{\Sigma}_\beta$ to zero means that we a priori assume that the item parameter values are independent between dimensions. This is of course highly unlikely, but we have chosen to do this to favor convergence of our estimation procedure; the Gibbs sampler (Geman & Geman, 1984). Recent Bayesian theory shows that such a choice may shrink the posterior estimate of the between-dimension correlation to zero, but also that this shrinkage effect will be minor when there are many observations (Marsman, Maris, Bercher, & Glas, 2016). We expect the shrinkage of these correlations to be small in our analyses, and our suspicion was confirmed by some additional simulations.

A.2. Simulating from the full-conditional distributions

The full-conditional distribution of the between dimension person covariance matrix $\boldsymbol{\Sigma}_\theta$ is easily sampled from:

$$f(\boldsymbol{\Sigma}_\theta) \propto \text{Inverse–Wishart}_{n-1}(\mathbf{S}_\theta),$$

where n refers to the sample size and $\mathbf{S}_\theta$ to the 'sample' covariance matrix $Cov(\boldsymbol{\theta})$. Similarly, we find that the full-conditional distributions for $\{\mu_{\beta,\ d}, \quad \sigma^2_{\beta,\ d}\}$ are easy to sample from:

$$f(\mu_{\beta,\ d}|\sigma_{\beta,\ d}, \quad \boldsymbol{\beta}_d) \propto \mathcal{N}(\bar{\boldsymbol{\beta}}_d, \quad \sigma^2_{\beta,\ d}/k)$$

$$f(\sigma_{\beta,\ d}|\boldsymbol{\beta}_d) \propto \text{Inverse-}\chi^2\left(k-1, \quad \sum_{i=1}^{k}\beta^2_{id}/(k-1)\right),$$

where k refers to the number of items in our analyses.

Unfortunately, the full-conditional distributions of the person and the item parameters are not readily sampled from. Standard approaches, such as the Metropolis within Gibbs approach of Patz and Junker (1999a, 1999b), are difficult to apply here due to the need of non-trivial fine-tuning that is required for each of the $n \times 3$ person and $k \times 3$ item parameters. This fine-tuning is particularly problematic as each of the persons responds to a possibly different set of items, and, similarly, each of the items has been responded to by a different set of persons.

To sample from the full-conditional distributions of the person and the item parameters we therefore utilize an independence chain Metropolis algorithm that was proposed by Marsman, Maris, Bechger, and Glas (2015). Their approach is particularly efficient when applied to the Rasch model and is simple to use with incomplete designs.[7]

A.3. Model comparison

Three cross-validation statistics were used, all three based on the deviation between the observed and the predicted response in the test data set. First, the prediction accuracy (ACC):

[7] Details about this algorithm as applied to the Rasch model can be found in Marsman (2014) pages 85–88.

$$ACC = 1 - \frac{1}{n}\sum_{i=1}^{n} o_i - p_i \tag{A.1}$$

where, both o_i is the observed response and p_i is the predicted response and n is the number of responses in the test data. Here, p_i is either correct or incorrect based on the probabilities following from Eq. (1), and the maximum a-posteriori (MAP) estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. The ACC reflects the percentage of correctly predicted responses by the model parameters.

The second cross-validation statistic was the root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2} \tag{A.2}$$

where, p_i is the predicted probability on a continuous scale between 0 and 1.

The third cross-validation statistic was the log-likelihood (LL), defined as follows:

$$LL = \sum_{i=1}^{n} o_i \log(p_i) + (1 - o_i)\log(1 - p_i). \tag{A.3}$$

Both the RMSE and the LL are presented since the LL provides a higher penalty to predictions that were confident and wrong (high deviation between o_i and p_i), whereas the RMSE provides an equal penalty for each deviation between o_i and p_i (Pelánek, 2015).

Appendix B. Robustness analysis

To investigate the stability of the comparison of the full fast-slow model with the more constrained versions of the model, we constructed multiple data sets and replicated the analyses presented in the paper.

B.1. Data selection

For the second data set, we selected responses to the 150 most played items, referred to as the most-played data set. This second data set includes some of the items from the first subset and additionally includes multi-digit multiplication items (such as: 1 × 500, 7 × 100, 9 × 12, 803 × 10 and 80 × 6000). Items with a minimum of 200 encounters were selected, resulting in 145 items. Through analysing the second data set we investigated whether the results from the first data set can be generalised to a data set including responses to a broader set of items. Also, replicating the initial analyses using this second data set provides a check of the robustness of the results.

For the single-digit items, we constructed two data sets based on the selection of children that completed at least thirty items within one day *or* within one week. For the most-played items, we selected data of children that completed at least 30 items within one day *or* one week *or* sixty items within one week. These choices resulted in a total of five different data sets. Within each data set, we selected items with a minimum of 200 responses, and looked at the child's first response to an item (multiple responses can be given to the same item within a set of 30 items). The total number of responses, children, items and percentage of missing responses for each data set are presented in Table B.4.

Table B.4
Data description. The number of responses, children, items, and amount of missing data in the different constructed data sets.

Item selection	Time	N responses	N children	N items	% missing
Single digit	Day	51,284	1164	64	31
	Week	180,651	3551	64	21
Most played	Day	387,882	7403	135	61
	Week	422,634	7860	145	63

For each of the four data sets, the response times were split using the overall median RT, within-person median RT and the within-item median RT. This resulted in a total of twelve model comparisons.

B.2. Model comparison

Table B.5
Fit statistics for each of the four models based on all different data selection procedures based on week data.

Item selection	RT split	Model	*DIC*	ACC_{CV}	$RMSE_{CV}$	LL_{CV}
Single digit	med	Baseline model	168,836	0.763	0.399	−2498
		$\theta_{fast} \neq \theta_{slow}$	166,932	0.767	0.399	−2507
		$\beta_{fast} \neq \beta_{slow}$	169,443	0.768	0.400	−2535
		Full model	163,923	0.779	0.393	−2427
	us	Baseline model	168,287	0.763	0.399	−2490
		$\theta_{fast} \neq \theta_{slow}$	165,004	0.773	0.396	−2495
		$\beta_{fast} \neq \beta_{slow}$	169,718	0.761	0.405	−2573
		Full model	161,507	0.776	0.391	−2401

(*continued on next page*)

Table B.5 (*continued*)

Item selection	RT split	Model	*DIC*	ACC_{CV}	$RMSE_{CV}$	LL_{CV}
	it	Baseline model	169,285	0.759	0.401	−2515
		$\theta_{fast} \neq \theta_{slow}$	169,907	0.759	0.415	−2760
		$\beta_{fast} \neq \beta_{slow}$	170,249	0.766	0.403	−2566
		Full model	164,251	0.776	0.393	−2428
Most played	med	Baseline model	439,633	0.736	0.424	−5428
		$\theta_{fast} \neq \theta_{slow}$	432,042	0.741	0.422	−5381
		$\beta_{fast} \neq \beta_{slow}$	437,394	0.735	0.424	−5416
		Full model	429,223	0.748	0.418	−5284
	us	Baseline model	439,734	0.731	0.427	−5495
		$\theta_{fast} \neq \theta_{slow}$	426,316	0.748	0.416	−5250
		$\beta_{fast} \neq \beta_{slow}$	437,301	0.737	0.422	−5409
		Full model	420,960	0.752	0.413	−5186
	it	Baseline model	439,659	0.736	0.424	−5437
		$\theta_{fast} \neq \theta_{slow}$	434,761	0.741	0.422	−5375
		$\beta_{fast} \neq \beta_{slow}$	440,873	0.734	0.424	−5448
		Full model	430,136	0.744	0.419	−5301

Note. RT split refers to the split method used for the dichotomization of the response times: med = overall median split, us = within person split, it = within item split. The baseline model refers to the fully constrained fast-slow model. The full model is the model that allows for both different person and item parameters in the fast and slow part. CV indicates that these fit statistics are based on cross validation.

Table B.6
Fit statistics for each of the four models based on all different data selection procedures based on day data.

Item selection	RT split	Model	*DIC*	ACC_{CV}	$RMSE_{CV}$	LL_{CV}
Single digit	med	Baseline model	46,019	0.797	0.377	−675
		$\theta_{fast} \neq \theta_{slow}$	45,392	0.806	0.374	−668
		$\beta_{fast} \neq \beta_{slow}$	45,760	0.789	0.380	−691
		Full model	44,689	0.806	0.371	−658
	us	Baseline model	45,910	0.791	0.377	−670
		$\theta_{fast} \neq \theta_{slow}$	44,244	0.807	0.370	−650
		$\beta_{fast} \neq \beta_{slow}$	45,623	0.784	0.382	−695
		Full model	44,073	0.806	0.367	−636
	it	Baseline model	46,208	0.793	0.380	−683
		$\theta_{fast} \neq \theta_{slow}$	44,858	0.812	0.372	−661
		$\beta_{fast} \neq \beta_{slow}$	47,733	0.803	0.380	−691
		Full model	44,646	0.804	0.371	−656
Most played	med	Baseline model	395,163	0.750	0.416	−5255
		$\theta_{fast} \neq \theta_{slow}$	388,508	0.749	0.415	−5244
		$\beta_{fast} \neq \beta_{slow}$	393,494	0.752	0.414	−5235
		Full model	383,261	0.760	0.409	−5105
	us	Baseline model	397,674	0.747	0.417	−5289
		$\theta_{fast} \neq \theta_{slow}$	380,354	0.764	0.408	−5119
		$\beta_{fast} \neq \beta_{slow}$	391,909	0.753	0.416	−5285
		Full model	376,130	0.768	0.406	−5039
	it	Baseline model	394,142	0.752	0.415	−5241
		$\theta_{fast} \neq \theta_{slow}$	388,683	0.755	0.415	−5338
		$\beta_{fast} \neq \beta_{slow}$	393,792	0.752	0.415	−5284
		Full model	384,519	0.762	0.410	−5116

Note. The baseline model refers to the fully constrained fast-slow model. The full model is the model that allows for both different person and item parameters in the fast and slow part. CV indicates that these fit statistics are based on cross validation.

References

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117*, 288.

Ackerman, P. L., & Cianciolo, A. T. (2000). Cognitive, perceptual-speed, and psychomotor determinants of individual differences during skill acquisition. *Journal of Experimental Psychology: Applied, 6*, 259.

Ashcraft, M. H. (1995). Cognitive psychology and simple arithmetic: A review and summary of new directions. *Mathematical Cognition, 1*, 3–34.

Ashcraft, M. H., & Guillaume, M. M. (2009). Mathematical cognition and the problem size effect. *Psychology of Learning and Motivation, 51*, 121–151.

Aunola, K., Leskinen, E., Lerkkanen, M.-K., & Nurmi, J.-E. (2004). Developmental dynamics of math performance from preschool to grade 2. *Journal of Educational Psychology, 96*, 699.

Carr, M., & Jessup, D. L. (1997). Gender differences in first-grade mathematics strategy use: Social and metacognitive influences. *Journal of Educational Psychology, 89*, 318.

Coomans, F., Hofman, A., Brinkhuis, M., van der Maas, H. L., & Maris, G. (2016). Distinguishing fast and slow processes in accuracy-response time data. *PloS One, 11*,

e0155149.
De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533–559.
De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software, 48*, 1–28.
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.
de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*, 595–624.
De Smedt, B., Holloway, I. D., & Ansari, D. (2011). Effects of problem size and arithmetic operation on brain activation during calculation in children with varying levels of arithmetical fluency. *Neuroimage, 57*, 771–781.
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.
DiTrapani, J., Jeon, M., De Boeck, P., & Partchev, I. (2016). Attempting to differentiate fast and slow intelligence: Using generalized item response trees to examine the role of speed on intelligence tests. *Intelligence, 56*, 82–92.
Dowker, A. (2005). *Individual differences in arithmetic: Implications for psychology, neuroscience and education*. East Sussex: Psychology Press.
Eggen, T. J., & Verhelst, N. D. (2011). Item calibration in incomplete testing designs. *Psicológica: Revista de metodología y psicología experimental, 32*, 107–132.
Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub.
Freudenthal, H. (1991). *Revisiting mathematics education*. Dordrecht, the Netherlands: Kluwer.
Geary, D. C., Widaman, K. F., & Little, T. D. (1986). Cognitive addition and multiplication: Evidence for a single memory network. *Memory & Cognition, 14*, 478–487.
Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). (3rd ed.). *Bayesian data analysisVol. 2*. Chapman and Hall/CRC Press.
Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721–741.
Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*, 608.
Hofman, A. D., Visser, I., Jansen, B. R., & van der Maas, H. L. (2015). The balance-scale task revisited: A comparison of statistical models for rule-based and information-integration theories of proportional reasoning. *PloS One, 10*, e0136449.
Imbo, I., Vandierendonck, A., & Rosseel, Y. (2007). The influence of problem features and individual differences on strategic performance in simple arithmetic. *Memory & Cognition, 35*, 454–463.
Jansen, B. R., Hofman, A. D., Savi, A., Visser, I., & van der Maas, H. L. (2016). Self-adapting the success rate when practicing math. *Learning and Individual Differences, 51*, 1–10.
Jost, K., Beinhoff, U., Hennighausen, E., & Rösler, F. (2004). Facts, rules, and strategies in single-digit multiplication: Evidence from event-related brain potentials. *Cognitive Brain Research, 20*, 183–193.
Kirk, E. P., & Ashcraft, M. H. (2001). Telling stories: The perils and promise of using verbal reports to study math strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 157.
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. https://doi.org/10.1016/j.compedu.2011.02.003.
Kroesbergen, E. H., & Van Luit, J. E. (2002). Teaching multiplication to low math performers: Guided versus structured instruction. *Instructional Science, 30*, 361–378.
Kroesbergen, E. H., Van Luit, J. E., & Maas, C. J. (2004). Effectiveness of explicit and constructivist mathematics instruction for low-achieving students in the Netherlands. *The Elementary School Journal, 104*, 233–251.
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.
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.
Liu, O. L., Wilson, M., & Paek, I. (2008). A multidimensional Rasch analysis of gender differences in PISA mathematics. *Journal of Applied Measurement, 9*, 18.
Maris, G., & Van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika, 77*, 615–633.
Marsman, M. (2014). *Plausible values in statistical inference*. Enschede, the Netherlands: University of Twentehttps://doi.org/10.3990/1.9789036537445 Ph.D. thesis.
Marsman, M., Maris, G., Bechger, T., & Glas, C. (2015). Bayesian inference for low-rank ising networks. *Scientific Reports, 5*.
Marsman, M., Maris, G., Bercher, T., & Glas, C. (2016). What can we learn from plausible values? *Psychometrika, 81*, 247–289.
Miller, K., Perlmutter, M., & Keating, D. (1984). Cognitive arithmetic: Comparison of operations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 46.
Mislevy, R. J., & Wu, P.-K. (1996). Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing. *ETS Research Report Series*.
Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence, 40*, 23–32.
Patz, R. J., & Junker, B. W. (1999a). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342–366.
Patz, R. J., & Junker, B. W. (1999b). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146–178.
Pelánek, R. (2015). Metrics for evaluation of student models. *JEDM-Journal of Educational Data Mining, 7*, 1–19.
Price, G. R., Mazzocco, M. M., & Ansari, D. (2013). Why mental arithmetic counts: Brain activation during single digit arithmetic predicts high school math scores. *The Journal of Neuroscience, 33*, 156–163.
Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research.
Reed, H. C., Stevenson, C., Broens-Paffen, M., Kirschner, P. A., & Jolles, J. (2015). Third graders' verbal reports of multiplication strategy use: How valid are they? *Learning and Individual Differences, 37*, 107–117.
Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581–592.
Schwarz, G., et al. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.
Siegler, R. S. (1988). Strategy choice procedures and the development of multiplication skill. *Journal of Experimental Psychology: General, 117*, 258.
Siegler, R. S. (2007). Cognitive variability. *Developmental Science, 10*, 104–109.
SLO. (2009). *Leerlijnen rekenen/wiskunde [learning program arithmetic/mathematics]*http://www.tule.slo.nl.
Straatemeier, M. (2014). *Math Garden: A new educational and scientific instrument*http://hdl.handle.net/11245/1.417091 Ph.D. thesis.
Tenison, C., Fincham, J. M., & Anderson, J. R. (2014). Detecting math problem solving strategies: An investigation into the use of retrospective self-reports, latency and fMRI data. *Neuropsychologia, 54*, 41–52.
Van der Ven, S. H., Boom, J., Kroesbergen, E. H., & Leseman, P. P. (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.
van der Ven, S. H., Straatemeier, M., Jansen, B. R., Klinkenberg, S., & van der Maas, H. L. (2015). Learning multiplication: An integrated analysis of the multiplication ability of primary school children and the difficulty of single digit and multidigit multiplication problems. *Learning and Individual Differences, 43*, 48–62.
Verguts, T., & Fias, W. (2005). Interacting neighbors: A connectionist model of retrieval in single-digit multiplication. *Memory & Cognition, 33*, 1–16.