Cognitive Analysis of Educational Games

The Number Game

van der Maas, H.L.J.; Nyamsuren, E.

DOI
10.1111/tops.12231

Publication date
2017

Document Version
Final published version

Published in
Topics in Cognitive Science

TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 9 (2017) 395–412

ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12231

This article is part of the topic "Game-XP: Action Games as Experimental Paradigms for Cognitive Science," Wayne D. Gray (Topic Editor). For a full listing of topic papers, see: http://onlinelibrary.wiley.com/doi/10.1111/tops.2017.9.issue-2/issuetoc.

Cognitive Analysis of Educational Games: The Number Game

Han L. J. van der Maas,[a] Enkhbold Nyamsuren[b]

[a]*Department of Psychology, University of Amsterdam*
[b]*Welten Institute, Open University*

Received 3 July 2015; received in revised form 28 January 2016; accepted 19 June 2016

Abstract

We analyze the cognitive strategies underlying performance in the Number task, a Math game that requires both arithmetic fluency and mathematical creativity. In this game all elements in a set of numbers (for instance, 2, 5, 9) have to be used precisely once to create a target number (for instance, 27) with basic arithmetic operations (solution: $[5-2] \times 9$). We argue that some instances of this game are NP complete, by showing its relation to the well-known Partition problem. We propose heuristics based on the distinction in forward and backward reasoning. The Number Game is part of Math Garden, a popular online educational platform for practicing and monitoring math skills using innovations in computerized adaptive testing. These educational games generate enormous amounts of rich data on children's cognitive development. We found converging evidence for the use of forward proximity heuristics in the data of Math Garden, consisting of more than 20 million answers to 1,700 items. Item difficulties and the structure of correct answers were analyzed.

Keywords: Education; Games; Arithmetic; Reasoning; NP Complete; Number game

1. Introduction

Math Garden (Rekentuin.nl) is a popular Dutch educational website for practicing and monitoring math skills in a gamified environment. Math Garden consists of a garden with plants, each representing a Math game. These plants grow as math ability increases.

Correspondence should be sent to Han L. J. van der Maas, Department of Psychology, University of Amsterdam, Weesperplein 4, Room 207, 1018 NX Amsterdam, Netherlands. E-mail: h.l.j.vandermaas@uva.nl

Currently Math Garden contains games for learning basic arithmetic operations, but also for counting, series, fractions, clock reading, working memory, deductive reasoning, and perceptual intelligence.

One especially interesting game is the Number game. It requires arithmetic fluency as well as creativity. The Number game can be defined as follows. Given a set S_N of numbers and a set S_O of arithmetic operators, a player has to make a target number T. Each number in S_N can be used only once, but operators in S_O can be reused. Minimum sizes of S_N and S_O are two. S_O can consist of any combinations of following operators: +,−,×, and /.

Fig. 1 shows a screenshot of an instance of the Number game. The player is required to reach the target number 2 ($T = 2$) using only addition and subtraction ($S_O = +,-$) using three other numbers 1, 5, and 6 ($S_N = 1, 5, 6$). Possible solutions are 6−5 + 1, 6 + 1−5, and 6−(5−1).[1] The input fields of the game are designed such that children can give answers without using brackets. It is easy to build a large item bank containing items of various difficulties (the Number game in Math Garden contains 1,650 items). Difficult items do not necessarily have large sizes of S_N. A notorious difficult case is to create 24 with the number 1, 3, 4, and 6.[2]

When four numbers must be used and the target is fixed to 24, this game is known as the 24-game. It is available as a commercial game (claiming 10 million users) and is also played in many schools over the world as an educational game. Although the game has not been analyzed from a cognitive science perspective before, the Number game clearly

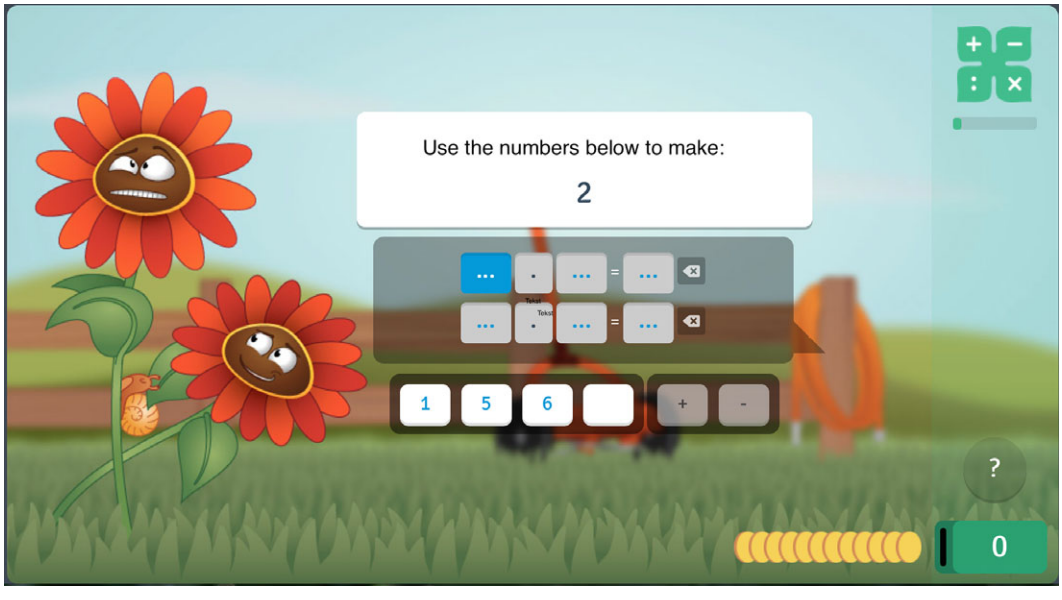

Fig. 1. A screenshot from Math Garden showing an instance of the Number game. The player is required to reach the target number 2 using only addition or subtraction and three other numbers 1, 5, and 6. Possible solutions are 6−5 + 1, 6 + 1−5, and 6−(5−1).

requires fluency in basic arithmetic skills and something that we could call mathematical creativity. Only a few scientific sources describe the game (Eley, 2009; Flaherty, Connoly, & Lee-Bayha, 2005). These unpublished studies report positive learning effects of playing the 24-game on arithmetic development. Several websites discuss the game, some providing theory on solution equivalence and puzzle difficulties (www.4nums.com).

This study is organized as follows. We first describe the setup of Math Garden in more detail. We then provide a first analysis of how people solve the Number task. Based on analyses of similar tasks, we will propose two heuristics, the proximity heuristic and the backward heuristic. We will use the data from Math Garden to test for these heuristics, and we will replicate our results in two additional datasets.

2. Math Garden

Math Garden originated in basic research on the dynamics of cognitive development. It has been very difficult to develop reliable and valid measurement instruments for cognitive development that can be used across a large age range. It is even more difficult to construct these instruments in such a way that they can be used in high-frequency measurement, say once a week or once a day. Next, even if such tests were available, schools would likely refuse to allow researchers to test children daily or weekly.

Our solution to these problems, or at least some of these problems, starts with the observation that schoolchildren complete exercises in math daily. If we could obtain the data on these exercises, a new measurement system could be within reach. To acquire measurements that are scientifically useful, we had to start from the point of view of measurement theory.

Measurement models are developed in the field of psychometrics. Modern test theory provides a number of techniques for educational measurement, the most promising being computerized adaptive testing (Wainer, Dorans, Flaugher, Green & Mislevy 2000). It simply means that persons do not have to complete entire tests, but are presented with items depending on the successes and failures on earlier items. Based on prior responses, the most informative item is selected to converge as soon as possible to a reliable final estimate of person ability. In computer adaptive testing (CAT), the item bank consists of at least hundreds of items.

To apply these techniques in a practice system that children and schools would be willing to use on a daily basis, two problems had to be solved. The first problem concerns pre-testing. The procedure of computerized adaptive testing only functions when all item difficulties are known. This means that for each task, hundreds of persons have to be tested on hundreds of items before one can start a CAT. As Math Garden consists of 18 games with more than 15,000 items in total, pre-testing is out of the question. The second problem is that in CAT the most informative item is an item for which the expected probability of being correct is about 0.5. In a practice system, a failure rate of 50% is unacceptable. It is not difficult to select easier items in a CAT, but then the speed of convergence in estimating ability deteriorates quickly (Eggen & Verschoor, 2006).

We solved the first problem using an estimation method originally proposed for chess competitions. In this so-called Elo system, ratings (abilities) of players are updated after each game by simple update formula (Elo, 1978). In this update formula, the outcome of a game is compared with the expected outcome computed from the ratings of the players prior to the game. The advantage of Elo's dynamic estimation method is that it can start with arbitrary initial ratings. We can set all players' ratings to zero, let players play games, and after some time the ratings will converge to values that accurately represent (differences in) playing strength. In Math Garden we use the same system with some modifications. When persons play items, they increase in rating (ability) when they solve the item, and decrease in rating when they fail. The reverse is true for the item ratings. Details of our adaptation of the Elo system can be found in Klinkenberg, Straatemeier, and Van der Maas (2011) and Maris and Van der Maas (2012).

We solved the second problem by using response times in the scoring of answers. On (very) easy items, accuracy is no longer informative on ability but speed of responding is (Van der Maas & Wagenmakers, 2005). We apply an explicit scoring rule to inform players about the weighing of accuracy and speed. This scoring rule, called high speed high stakes, weights accuracy (+1, −1) with the remaining time for an item. Given a time limit of, for instance, 20 s, a correct answer in 5 s gives a score of +15, whereas an error in 15 s yields a score of −5. In Maris and Van der Maas (2012), it is shown that this scoring rule has excellent psychometric properties.

This scoring rule is incorporated into the extended Elo system that is used in Math Garden. In the games, the scoring rule is represented with coins, equal to the time in seconds available for the item. Each second one coin disappears. In the case of a correct answer, the remaining coins are added to the total number of coins collected by children. In the case of an error, the remaining coins are subtracted from the total. In this way, the scoring rule is understandable even for young children and adds gamifying elements to the task. For example, children can go to a prize cabinet and buy virtual prizes, such as flags and trophies, using collected coins. Because the games are adaptive to the level of ability of the players, the coins and prizes won are independent of ability and depend only on how much one plays.

These online games let children practice intensively at their own level with direct feedback, two important requirements of deliberate practice (Ericsson, 2006). Teachers are provided with learning analytics at the class and individual level. Apart from adding children to the system, and providing them with a login name and password, their task is minimal. Math Garden is a self-organizing additional learning tool that does not require work of teachers. Note that these websites do not give any instruction. They take over the practicing and monitoring task, not the instruction.

Math Garden became quickly popular in the Netherlands. Almost 2,000 schools bought subscriptions either for selected groups of students or the whole school. In addition, many families took home subscriptions. In the spring of 2015, more than 150,000 children of preliminary elementary schools in the Netherlands use Math Garden regularly. During weekdays, about 1 million item responses are collected with a speed of 60 per second at peak hours.

A number of studies using Math Garden data have been published (Gierasimczuk, van der Maas, & Raijmakers, 2013; Groeneveld, 2014; Jansen, De Lange, & Van der Molen, 2013a; Jansen et al., 2013b, 2014; Kadengye, Ceulemans, & Van den Noortgate, 2014; Nyamsuren, der Van Maas, & Taatgen, 2015; Van der Ven, Van der Maas, Straatemeier, & Jansen, 2013). One example concerns the counting game. In this game, children have to count fish in an aquarium. The number of fish in an item varies from 1 to more than 50, and displays can be ordered (for instance, dice patterns) or random. Time limit per item is 20 s. An important phenomenon in the study of counting is that counting small numbers takes place by subitizing, a rapid automatic assessment of numbers smaller than 4 or 5. In Jansen et al. (2014), we compared ratings of and response times to counting items with random displays, line displays, and dice displays. Because of the advantage of dice patterns over line and random patterns up to the number six, we argued that subitizing is perhaps based on rather general pattern recognition abilities and not due to some domain-specific ability.

A second example concerns the development of deductive logical reasoning. Gierasimczuk et al. (2013) analyzed data of a variant of the popular Mastermind game. They proposed a logical analytic tableaux model for the deductive reasoning process in this task and verified this model with data from 37,000 children who played the game regularly. More recently, Nyamsuren et al. (2015) found that errors in an SET game depend on various factors such as progression of game play, past experience with the game, strategy, and a structure of a specific game instance. Finally, Braithwaite, Goldstone, van der Maas, and Landy (2016) presented evidence from Math Garden for two non-formal mechanisms, perceptual grouping and opportunistic selection, to determine order of evaluation of arithmetic expressions. Interestingly, the effects associated with these mechanisms increased with age.

3. The Number game

The Number game is one of the 18 games in Math Garden designed to study and improve player's mathematical reasoning skills. Within the Math Garden, games are divided into a basic garden and a bonus garden. The Number game is usually only available in the bonus garden and can only be played when the games in the basic garden are finished. However, it is a popular game and a lot of data have been collected (see Descriptives).

The Number game requires creativity due to its complex search space. The game resembles many elements of so-called NP-complete problems. In such problem, the time necessary for finding an optimal solution increases exponentially with increasing size of an initial set. This property makes NP-complete problems particularly difficult, even for computers. As in the Number game, checking solutions of NP-complete problems is relatively easy.

To understand the complexity of the Number game further, it is useful to compare the game with the famous and thoroughly studied Partition problem. The goal in the Partition

problem (Hayes, 2002) is simple: Given a set of N positive numbers, one should create two non-overlapping subsets, such that the sums of the two subsets are equal. The following is an example originally given by Hayes. Given a set of numbers 2 10 3 8 5 7 9 5 3 2, two subsets can be created so that numbers in both of them add up to 27: 10 7 5 3 2 and 9 8 5 3 2.

Interestingly, Partition problems can be viewed as instances of the Number game (Kurzen, 2011). We can prove this by reformulating Hayes's example Partition problem into the format of the Number game: given S_N = (2, 10, 3, 8, 5, 7, 9, 5, 3, 2) and S_O = (+, −), reach the target number $T = 0$. Then, the solution for the problem is (10 + 7 + 5 + 3 + 2)−(9 + 8 + 5 + 3 + 2). Hence, any Partition problem can be reformulated into an instance of the Number game, where S_O = (+, −) and $T = 0$. Consequently, the Number game is also NP complete with the set of operators S_O = (+, −) (Kurzen, 2011).[3]

4. How do humans solve the number problem?

An exhaustive systematic search of the problem space is clearly not an option for human players. However, they do play the game and often find solutions. The problem-solving literature (Willingham, 2007) suggests two general heuristics for this type of search task.

The first heuristic is based on forward reasoning. It resembles a well-known heuristic for the Partition problem. The Partition problem arises in real life when children need two teams of equal strength to play a game of soccer, for instance. In the so-called soccer heuristic, the two strongest players pick their team members in turns. The key is to pick players in a decreasing order of their skills. This strategy usually results in teams closely matched in skills. The soccer heuristic can be directly applied to the Partition problem. Most of the time, the soccer heuristic will result in a solution that is either optimal or close to optimal within polynomial time.

We hypothesize that human players will use a heuristic closely related to the soccer heuristic. This proximity heuristic, as we will call it, is characterized by forward reasoning, greediness, and convergence. Forward reasoning does not involve the target number in operations. Greediness entails that subjects will start with the biggest numbers in S_N. We define a solution as greedy if the largest and the second largest numbers in S_N (further denoted as $Max_1(S_N)$ and $Max_2(S_N)$, respectively) are used as the first and second operands of the first operation. Convergence implies that subjects attempt to get as close as possible to the target in the first step. The degree of convergence is measured as a ratio of the result of the first operation, R_1, and the target number T:

The degree of convergence is calculated using two different ratios as convergence can occur either upward (for multiplication and summation) or downward (for subtraction and division). Convergence varies between 0 and 1, with 1 indicating the fastest convergence on T. The proximity heuristic can be understood as a form of greedy hill climbing.

A typical example of the application of the proximity heuristic is the solution $12 \times 5-3-2$ for $S_N = (2, 3, 5, 12)$ and $T = 55$. It is greedy as the biggest numbers are used first and convergence is large (55/60 = 0.92). Many Number game puzzles are solvable in this way. Below we will test whether the proximity heuristic is indeed dominant in solving Number game puzzles by humans by analyzing the difficulties of items as they are estimated in Math Garden.

The second general heuristic is based on backward reasoning. It starts by applying operators of S_O to the target, and on one value of S_N. This would lead to a new T, T', and a reduced set S_N'. The process is then repeated for T' and S_N'. For $T = 120$, $S_N = (6, 15, 35)$ and $S_O = (+, -, \times, /)$, the proximity heuristic fails. However, after realizing that 120/6 = 20, the reduced problem $T' = 20$, $S_N' = (15, 35)$ is easily solved. We have no precise hypotheses on the backward reasoning heuristic of human players for this game, except that we expect that they are rarely used. This hypothesis is mainly based on unsystematic observations of human playing behavior in this game in Math Garden but also in other versions, such as the 24-game. Nevertheless, it will be tested below. We hypothesize that items that require backward reasoning are more difficult than items that require forward reasoning.

Clearly, the type of reasoning required is not the only defining characteristic of item difficulty. Items that only require addition and subtraction are expected to be easier. These operations are learned before multiplication and division. Evidently, the size of S_N matters as well as the actual numbers in S_N and T. These aspects will be incorporated into the analysis of item difficulty. Note that in this study we focus on the cognitive analysis. We think this is a prerequisite for further analysis of developmental and other individual differences.

5. Results

In Math Garden, item difficulties are continually updated according to the modified Elo algorithm that uses both the accuracy and response time of answers to items. In earlier publications, we have shown that these estimates are highly reliable (Klinkenberg et al., 2011) and provide information on the characteristics that make items difficult (e.g., Gierasimczuk et al., 2013).

We will first analyze the types of correct answers given to different types of items. Second, we will investigate item difficulty in the Number game using multiple regression. Third, the results of the regression analyses are illustrated within sets of items. Finally, the results of two control studies are presented.

6. Descriptives

All data were extracted from Math Garden in June 28, 2015, and replicated the results of initial analyses on the data extracted in January 2015. The data consisted of

20,949,410 answers from 177,880 players. The frequency of number of played items shows a typical exponential distribution, where 128,271 subjects played more than 10 items, 51,405 subjects played more than 100 items, and 2,228 played more than 1,000 items.

Ability ratings differed by age and frequency of play. Within the primary school range, ratings correlated 0.49 with age and 0.53 with frequency of play. Age and frequency of play were uncorrelated.

7. Type of answer analysis

The preference for the proximity heuristics can be demonstrated by using subsets of items that require only addition. For an item that requires only addition, the order of numbers should not matter if players do not apply the proximity heuristic. One example is the item $T = 125$, $S_N = (1, 2, 2, 20, 100)$, and $S_o = (+, -)$. The correct answer only requires addition and the order of numbers is clearly irrelevant. Still, the typical proximity answer $(100 + 20 + 2 + 2 + 1)$ was by far the most popular answer (49,681 of 95,942 answers, with 11,323 for the second most popular answer).

There are 316 addition-only items that were played at least 50 times in Math Garden. As it is shown in Fig. 2A, players are highly likely to pick $Max_1(S_N)$ as the first operand of the first operation. Similarly, players are likely to pick $Max_2(S_N)$ as the second operand of the first operation (Fig. 2B). Players' biases toward $Max_1(S_N)$ and $Max_2(S_N)$ are much higher than the probabilities of choosing these numbers randomly. Therefore, even when it is unnecessary, players prefer to start calculations with largest numbers.

Moreover, players seem to perform better when the result of the first operation is closer to the target number T. Fig. 2C shows difficulty ratings for previously mentioned 316 items. For each item, the figure also shows $(Max_1(S_N) + Max_2(S_N))/T$ ratio. There is a negative correlation between rating and ratio ($r(314) = -0.48$, $p < .001$), indicating that players' strategy is dependent on a greedy approach. Players find items where first operations converge faster to target numbers to be easier than items where convergence is slower. Alternatively, the significant correlation can be explained by increasing size of S_N, denoted as $Size(S_N)$, as higher $Size(S_N)$ inevitably results in increased item difficulty and decreased ratio. To further verify if the significant correlation is indicative of a greedy strategy or caused by increasing $Size(S_N)$, separate correlation tests were performed on groups of items with the same $Size(S_N)$. The results are $r(144) = -0.14$, $p = .09$ for $Size(S_N) = 3$, $r(76) = -0.25$, $p = .02$ for $Size(S_N) = 4$, and $r(67) = -0.54$, $p < .01$ for $Size(S_N) = 5$. The negative correlation between ratio and difficulty rating is consistent for items of the same $Size(S_N)$. What is even more interesting is that the correlation becomes stronger with larger $Size(S_N)$ despite the decreasing number of observations. A possible explanation is that players rely more on the proximity heuristics with the increasing number of choices to consider. In easier items with few combinations of numbers to consider, players may go through these combinations without a need to rely on heuristics.

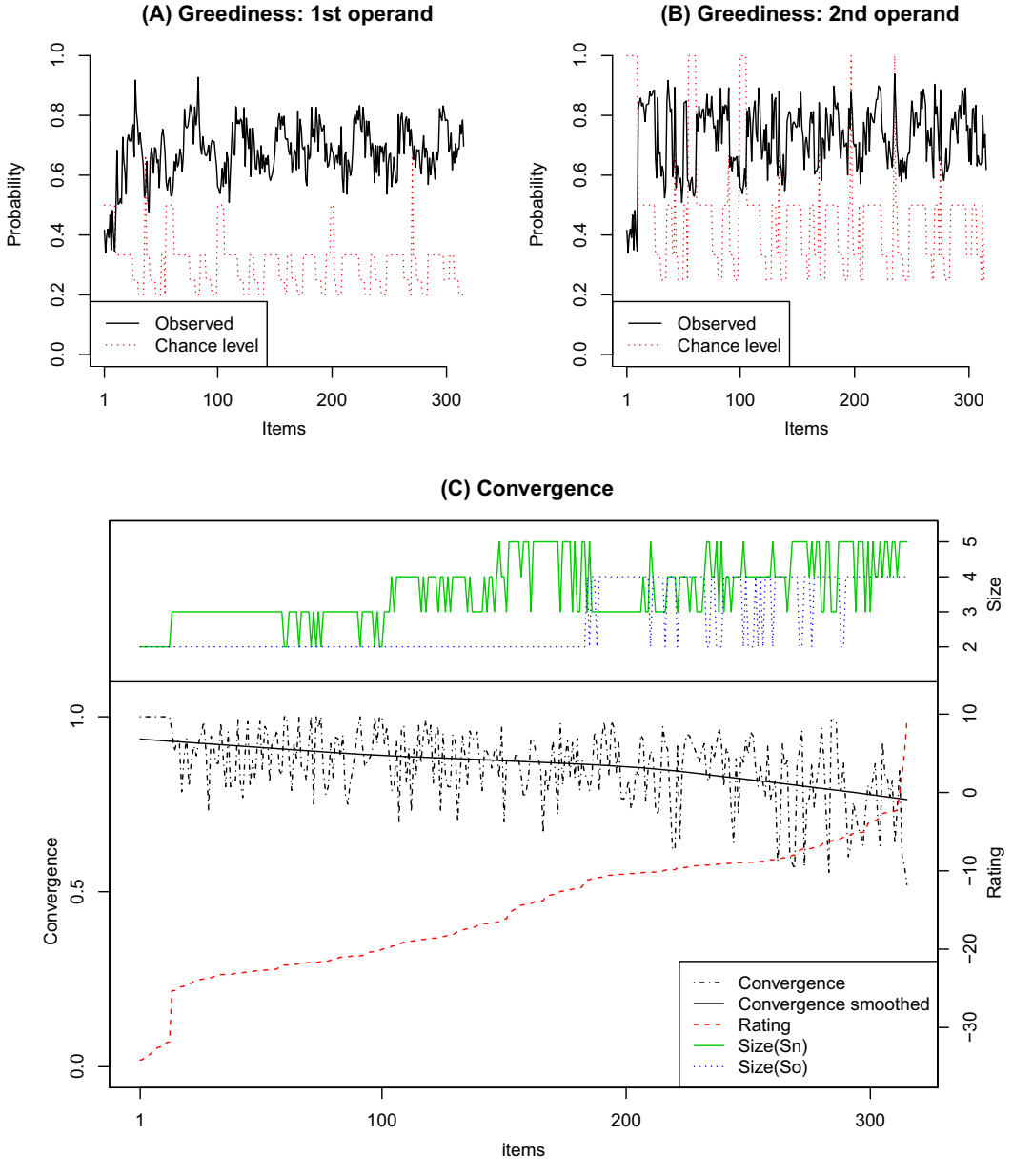

Fig. 2. The data are shown for 316 items that were played at least 50 times, have positive target numbers, and require only addition. For each item, the graph shows observed proportions of trials in which (A) the biggest number from S_N, $Max_1(S_N)$, was used as the first operand of the first operation, and (B) the second biggest number from S_N, $Max_2(S_N)$, was used as the second operand of the first operation. Observed proportions are contrasted against base probabilities of random choices. (C) For each item, the graph shows difficulty rating and ratio of the sum of two largest numbers to the target number, $(Max_1(S_N) + Max_2(S_N))/T$. For this ratio a smooth curve is added. The upper part of the graph shows $Size(S_N)$ and $Size(S_O)$, the number of operators allowed, for each item. Items in panel C were ordered by increasing ratings of difficulty.

It is possible that the proximity heuristic is used in addition-only items as the operation is highly compatible with the heuristic. However, analyses on data from 72 items that require one or more multiplications indicate that the proximity heuristic also plays important role in those items. A similar bias toward $Max_1(S_N)$ as observed in addition-only items is also observed in multiplication-only items. On average, players choose $Max_1(S_N)$ as the first operand of the first operation in 69% of trials compared to 51%, if choices of first operand were random. This difference is significant (c^2 (1, N = 72) = 4.1, $p < .05$). Among the 72 items, there are only 18 items where $Size(S_N) > 2$. For these items, in 81% of the cases $Max_2(S_N)$ was chosen as the second operand (54% chance level). This difference is non-significant, likely due to the low number of items.

8. Regression analysis

Greediness and convergence were included alongside other item properties as variables in a linear regression analysis. We included an interaction effect of greediness and convergence because we expect that the combination of the two will make items particularly easy. The dependent variable is item difficulty as determined by the adapted Elo system used in Math Garden. Results are reported in Table 1. The intercept indicates the difficulty rating of an item with $Size(S_N) = 3$ and $Size(S_O) = 2$. $Size(S_O)$ represents the number of unique operators available for the item. *NumbersIncrease* and *OperatorsIncrease* indicate increases in $Size(S_N)$ and $Size(S_O)$, respectively. *Fractional* is 1 if T is a fractional number and otherwise 0. *UseAdd*, *UseSubstract*, *UseMultiply,* and *UseDivide* are 1 if corresponding operators are used at least once in the solution and otherwise 0. *Greedy* is 1 if the solution is greedy and 0 otherwise. Finally, *Convergence* is the degree of convergence between 0 and 1.

As expected, item's difficulty increased with increases in $Size(S_N)$ and $Size(S_O)$. Also, fractional numbers significantly increased difficulty. Addition is the easiest operation,

Table 1
Linear regression model on items' difficulty ratings: $R^2 = 0.84$, $F(10, 985) = 513$, $p < .001$

Predictors	Coefficients	*SE*	β	*t* values	*p* values
Intercept	−10.80	0.72		−14.94	< .001
NumbersIncrease	0.43	0.22	0.23	1.99	.047
OperatorsIncrease	4.30	0.18	0.43	24.14	< .001
Fractional	6.40	0.94	0.09	6.82	< .001
UseAdd	3.14	0.43	0.12	7.35	< .001
UseSubstract	8.85	0.33	0.38	26.78	< .001
UseMultiply	3.71	0.40	0.16	9.38	< .001
UseDivide	3.85	0.44	0.13	8.77	< .001
Greedy	0.18	0.70	0.01	0.26	.796
Convergence	−2.02	0.81	−0.06	−2.49	.013
Greedy:Convergence	−8.22	1.16	−0.29	−7.10	< .001

while, interestingly, subtraction seems to contribute the most to the difficulty of an item. Most interestingly, the interaction effect, *Greedy:Convergence*, indicates that items where solutions are both greedy and convergent are both significantly and considerably easier. Therefore, the regression model supports our hypothesis that the proximity heuristics is a major strategy in the Number game.

To explore the interaction effect, we analyze two special cases. We first consider items that have three numbers (Size(S_N) = 3) and require exactly one addition and one multiplication to reach the target number. In these items, multiplications should result in faster convergence on target numbers than summations. Therefore, the proximity heuristic predicts that easiest items should have first operations involving multiplications of two largest numbers in S_N. The evidence shown in Fig. 3 supports the prediction. There are eight items that are compatible with the proximity heuristic. These items also have the lowest difficulty ratings. Finally, there are 10 other items where multiplication is the first operation (red circles in Fig. 3). However, multiplications in these items are not greedy and do not involve both largest numbers from S_N. Similarly, there are five items that require a greedy approach but on addition as the first operation. All these items have a

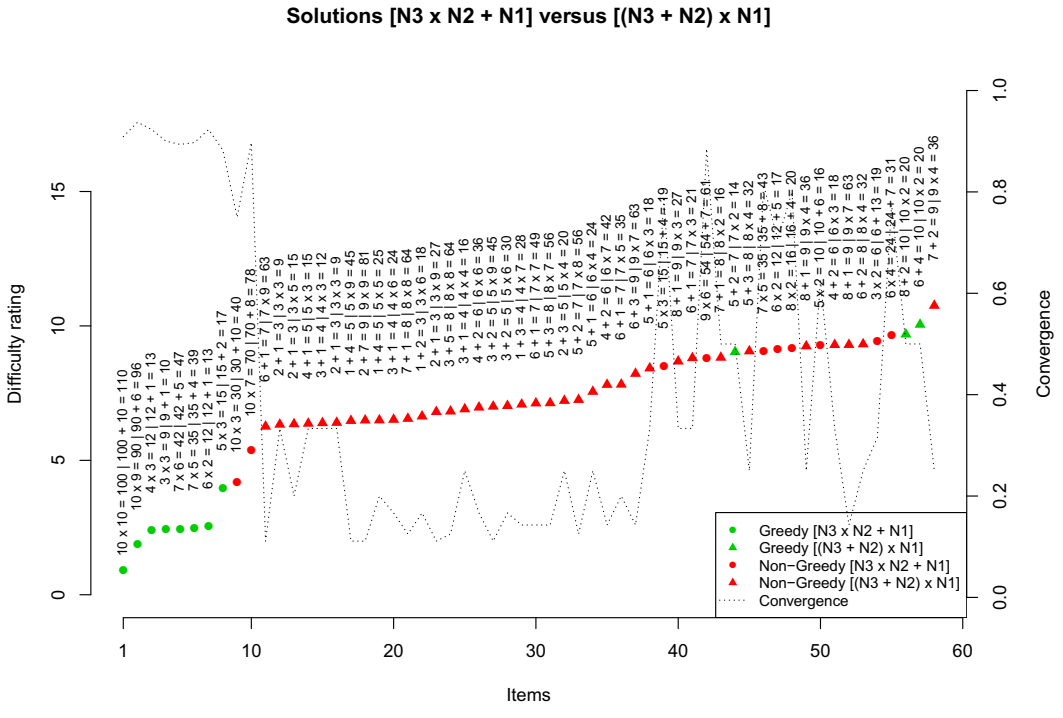

Fig. 3. Items that have three numbers and require exactly one addition and one multiplication. Items are shown in increasing order of difficulty ratings. Colors denote greediness. Circles indicate items with a multiplication as a first operation: $N_3 \times N_2 + N_1$. Triangles indicate items that require addition first: $(N_3 + N_2) \times N_1$. The striped line depicts convergence to the target number calculated as a ratio of first operations results to the target number: $[N_3 \times N_2/T]$ or $[(N_3 + N_2)/T]$.

varying degree of difficulty not forming any cluster. These results indicate that especially the combination of greediness and fast convergence makes items considerably easier.

Our second illustration focuses on an even smaller set of items. Fig. 4 shows a set of items ordered by increasing difficulty. All items have S_N = (1, 10, 100) and S_O = (+, −, ×, /), but require different combinations of operations to reach target numbers shown above plot points.

The figure shows three distinct clusters of items. The left-most cluster for only one item that requires only addition is compatible with proximity heuristic, as was discussed previously. The second cluster includes more difficult items that require various combinations of operations but are still compatible with the proximity heuristic. For example, solutions for the easiest and the hardest items in the cluster are 100 + 10−1 = 109 and (100−10) × 1 = 90. In both items, first operations involve largest numbers and result in numbers that are close or equal to target numbers.

All items in the third cluster violate the second criteria of the proximity heuristics, namely, that there should be fast convergence to the target number. Instead, first operations in these items result in numbers that are far from target numbers. For example, the solution for the easiest item in the third cluster is (10−1) × 100 = 900 The first operation results in 9 that which is not close to 900. Similarly, the solution for the hardest item in the third cluster is 10/1/100 = 0.1 where the result of the first operation, 10, is also far from the target number 0.1.

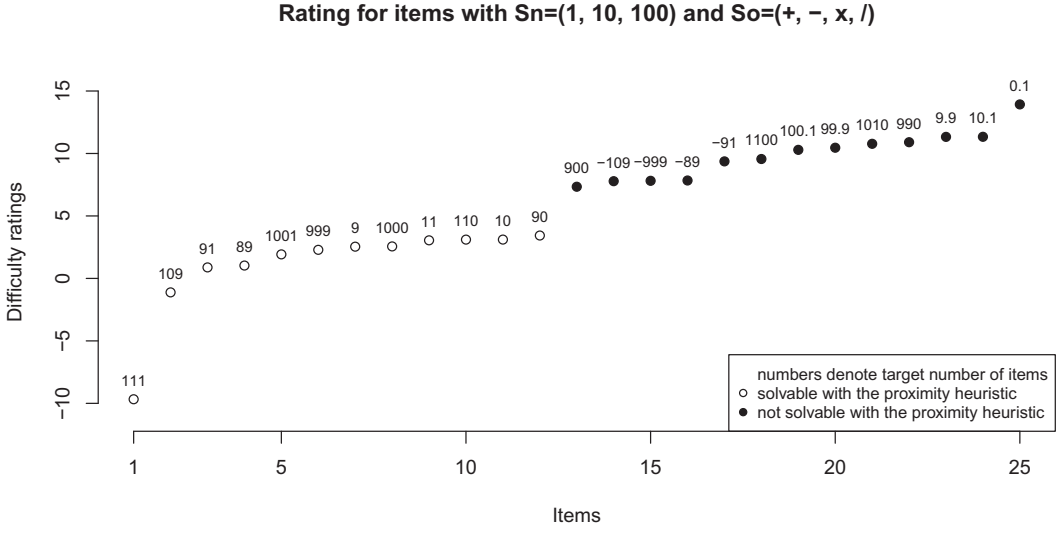

Fig. 4. List of items ordered by difficulty ratings. All items have S_N = (1, 10, 100) and S_O = (+, −, ×, /), but require different combinations of operations to reach target numbers shown above plot points. The first 12 items are compatible with the proximity heuristic. The difficult items also often have fractional targets, but note that targets 990 and 1,010 belong to the most difficult items. With backward reasoning (dividing the target by 10) the solution is easily found for these items, but the first forward step (100 ± 1) is not greedy and has low convergence.

9. Alternative explanations

The preference for forward reasoning in Math Garden's Number game might be due to the user interface of the game as shown in Fig. 2. First, the numbers in S_N are always presented in increasing order with the biggest numbers closest to the operators S_O (see Fig. 1). Second, the input fields require the input of the solution in a forward manner. Third, due to Math Garden's adaptive algorithm for administering items, subjects first receive many easy items that can be solved with forward reasoning. Only more difficult items require backward reasoning. By the time they get these items they might have developed a bias for forward reasoning.

To test whether forward reasoning bias is the product of the user interface, we performed a separate study in a controlled laboratory environment. Fifty-six college students between the age of 18 and 40 (M_{age} = 21.9, SD = 3.5) participated in this study. None of them had prior experience with the Number game in Math Garden. The pre-test consisted of 10 instances of the Number Game: 5 items on which the proximity heuristic is easily applicable (the forward items), and 5 items on which the backward heuristic applies well (the backward items). These forward items were as follows: S_N = (7, 10, 100), T = 1,007; S_N = (3, 20, 100), T = 83; S_N = (5, 20, 100), T = 2005; S_N = (1, 10, 100), T = 10; S_N = (2, 30, 100), T = 3,002. The backward items were as follows: S_N = (1, 10, 100), T = 900; S_N = (6, 10, 100), T = 940; S_N = (2, 10, 100), T = 1020; S_N = (2, 10, 100), T = 9.8; S_N = (4, 20, 100), T = 1,600. The tests were presented in a paper-and-pencil format, and subjects were allowed to report their answers in any format. The order of items was randomized but equal for all subjects. The time limit per item was 60 s.

The average percentage correct on the forward items 0.95 (SD = 0.12) was much higher than on the backward items 0.53 (SD = 0.32), $t(55) = 9.2$, $p < .001$. Subjects were also much faster on forward items, 15.28 (4.94) than the backward item 37.13 (12.26), $t(55) = -14.5$, $p < .001$. Both differences were highly significant. Hence, the preference for forward reasoning with the proximity heuristic was replicated in older subjects whom were not trained in Math Garden, using an answer format that did not provoke forward reasoning.

To further verify whether the proximity heuristics is used outside of Math Garden, we analyzed the data gathered by the www.4nums.com on the 24-game (with N = 4 and T = 24). On this website, a player can randomly play 1 of 1,362 solvable quadruples. By June 2015, 604,985 puzzles were solved by players. The website reports a number of statistics among which the percentage of correct answers. We have selected 515 quadruples with single unique solutions (a unique solution may still have different possible orders to perform the same operations).

We performed a linear regression analysis on the data of 515 quadruples. The predicted variable was the average solution time, which underwent a logarithmic transformation to normalize its distribution. Similar to the regression analysis reported in Table 1, greediness, convergence ratio, and types of operation involved were included as predictors. We used a stepwise selection procedure with backward elimination using BIC. This

resulted in a model with a rather low explained variance (0.09) in which Greediness but not Convergence was included.

This bad fit suggested that possibly other predictors are required. The restriction to just one target, 24, and a fixed size of S_N, may lead participants to apply dedicated "24" heuristics. It seems that there are at least two such heuristics.

The first is pairing of numbers into a following format: (N1 [AO] N2) [AO] (N3 [AO] N4) where [AO] is any operator. An example of pairing will be: $(7 + 1) \times (2 + 1)$ and $(3-1) \times (12 \times 1)$. The second additional heuristic is using multiplication of two factors in the last step: 1×24 (or 24×1), 2×12 (or 12×2), 3×8 (or 8×3), and 4×6 (or 6×4). This heuristic seems to require pattern recognition skills. Examples of the applications of this heuristic are as follows: $(10-1-1) \times 3 = 8 \times 3$, and $(5 + 1) \times (3 + 1) = 6 \times 4$.

To check for these heuristics, we added five Boolean predictors to the regression analysis: *Paired*, *F124*, *F212*, *F38,* and *F46*. *Paired* is 1 if solution can be written in a paired format, and 0 otherwise. *F124* is 1 if solution can be reduced to 1×24 (or 24×1), and 0 otherwise. *F212* is 1 if the solution can be reduced to 2×12 (or 12×2), and 0 otherwise. *F38* and *F46* are 1 if solutions can be reduced to 3×8 and 4×6, respectively, and 0 otherwise. All five predictors were centered at the mean. Again, we used a stepwise procedure. The results are summarized in Table 2.

Table 2 shows that adding the dedicated "24" heuristics led to much better results. The explained variance increased from 0.09 to 0.24. All predictors related to these heuristics are significant. Furthermore, Greediness and the interaction between Convergence and Greediness make items significantly easier. Thus, these results replicate the effects from the Math Garden analysis and indicate that the combination of greediness and fast convergence makes quadruples easier.

Note that the explained variance is still low compared to Table 1. This might be due to the fact that the items of the 24-task are much more difficult than the items of the Number task in Math Garden. Also the range of difficulties is smaller. Math Garden contains many very simple items, using only addition, for instance, but also extremely hard items. Finally, the dependent variable is the mean solution time, which is less

Table 2
Linear regression model on quadruples' difficulty ratings: $R^2 = 0.24$, $F(8, 505) = 18.9$, $p < .001$

Predictors	Coefficients	*SE*	β	*t* values	*p* values
Intercept	2.41	0.02		116.69	< .001
Convergence	−0.05	0.07	−0.04	−0.78	.437
Greedy	−0.23	0.04	−0.27	−6.59	< .001
Paired	−0.19	0.04	−0.22	−4.64	< .001
F212	−0.35	0.05	−0.30	−7.06	< .001
F38	−0.31	0.05	−0.25	−5.94	< .001
F46	−0.33	0.06	−0.23	−5.41	< .001
UseDivide	0.15	0.04	0.16	3.84	< .001
Convergence:Greedy	−0.34	0.11	−0.13	−3.18	.002
Paired:F46	0.30	0.12	0.11	2.54	.011

sophisticated than the difficulty ratings obtained in Math Garden. It is likely that the latter are more reliable.

10. Discussion

Math Garden serves both as an educational and a scientific instrument. Arithmetic, like many other scholastic abilities, requires extensive practice. Training basic arithmetic skills by educational adaptive games is attractive to children. Educational games that adapt to the ability level of the child and that provide direct feedback fulfill two important requirements of deliberate practice, which is essential in expertise development (Ericsson, 2006). Furthermore, teachers are released of the task of checking student work and are provided with sophisticated learning analytics. At the same time, the data of Math Garden, because of its size and the measurement frequency, open a new window on cognitive development for scientists.

In this study, we focused on one Math Garden game, the so-called Number game. This game in itself is of educational and scientific interest. The popularity of the 24-game, a restricted case of the Number game, attests to its educational relevance. As it requires both fluency in basic arithmetic skills and creative thinking, it meets important requirements of educational programs in learning math. The evidence on its effectiveness in advancing arithmetic thinking is still limited (Eley, 2009; Hayes, 2002). It was also not the focus of the analyses in this study.

We are primarily interested in the cognitive processes involved in solving Number game problems. As there was no prior theory or data available, we made the first steps here. By relating the Number game to the well-known Partition problem, we discovered that the search problem of the Number game is extremely hard. Many instances of the Number game are NP complete (but see note 2). For NP-complete problems, no optimal fast algorithms are known. That is, the time required to solve these problems increases exponentially as the size of the problem grows. Determining whether or not it is possible to solve these problems quickly is one of the principal unsolved problems in computer science today.

Wikipedia's list of NP-complete problems includes many popular games. How humans solve these special puzzles remains largely unknown. However, there is a vast literature on novel problem solving, a research tradition going back to the seminal work of Newell and Simon (1972). As a starting point, we proposed to investigate the use of forward and backward reasoning heuristics in the Number game. The next step should include investigation of underlying cognitive processes that implement these heuristics and allow humans to solve complex problems.

The specific forward heuristic we have proposed is the proximity heuristic. It is based on greediness (taking the largest remaining numbers from S_N) and convergence (select an operator from S_O, such that the target is closely approximated). We presented a number of empirical analyses of the Math Garden data that all converged to the same conclusion. The proximity heuristic is indeed dominant in children's problem-solving behavior. First, players prefer correct answers that fit the proximity heuristic above correct answers that do

not, especially for sets with larger N. Second, in the regression analysis the interaction effect of greediness and convergence added significantly to the prediction of item difficulty. Third, for the subset of problems with $N = 3$ that require one addition and one multiplication, we showed that the combination of greediness and high convergence in the first step made items systematically easier. Finally, we zoomed in on a subset of items all based on the same set (1, 10, 100). Items with targets for which the proximity heuristic leads to the correct answer are easier than items with non-compatible items. Interestingly, the latter items are solvable with backward heuristics, but this did not make them easy. We replicated the preference for the proximity heuristic with a paper-and-pencil task in college students and with online data collected with the 24-game. The preference for forward reasoning is not due to the setup of Math Garden, the layout of the Number game, or the age group.

These findings are only the first discoveries in the study of the Number game. To start with, the origins of the proximity heuristic still remain an open question. Answering this question may help us understand heuristics people choose to use in other problem-solving domains. Furthermore, it could well be the case that further specification of the proximity heuristic is possible. It remains unclear how players continue when the first attempt to apply the proximity heuristic fails. They may continue with the forward search with different choices of numbers, but at some point they might switch to backward reasoning or heuristics that we have not yet detected. Combinations of backward and forward reasoning are possible, too. Take, for instance, $S_N = (1, 3, 4, 9)$, $T = 111$. After the backward step $111/3 = 37$, the remaining problem $S_N = (1, 4, 9)$, $T = 37$, is easily solved with the proximity heuristic. Whether such combinations occur requires further study. Although we did not find much evidence for heuristics more advanced than the proximity heuristic, study of expert players might reveal such forms of reasoning.

New hypotheses on Number game problem solving can be investigated with the Math Garden dataset. It is, for instance, possible to analyze errors and response times. It is also possible to add items to Math Garden games, to test specific hypotheses on item difficulty, error types, or preferences for correct answers. It is also possible to investigate the relations between Number game performance and performance on other arithmetic tasks. Finally, it would be interesting to use the Number game in Math Garden to evaluate training methods in Number game problem solving. It might be the case that players do not use backward reasoning spontaneously, but do use it after some training.

To summarize, we see a bright future of educational games for both educational and scientifically purposes as a method for the study of cognition and cognitive development. The Number game is a typical example of this.

Acknowledgments

This research was supported by a grant from the Netherlands Organization for Scientific Research (NWO). We thank Wessel de Jong, Alexandra Roos, Rogier Hetem, and Nik Goedemans for their contribution to the data collection of the training study. We thank Cheng Chang for using the data of www.4nums.com.

Notes

1. There are good reasons to regard these three solutions as similar (see http://www.4nums.com/theory/).
2. The solution is 6/(1−3/4).
3. However, this does not prove that the Number game with S_O = (+, −, ×, /) is NP complete. It is easy to see that Number problems with S_O = (×, /) are equivalent to the S_O = (+, −) case by taking logarithms of all elements in S_N. So number problems with S_O = (×, /) are NP complete, too. However, the case S_O = (+, −, ×) is different. Suppose N of S_N is extremely large, then T will be an element of S_N. We call this element S_a. S_N will also hold two equal numbers S_b and S_c having difference 0. After reordering S_N to $(S_a, S_b, S_c, S_4 \ldots S_n)$, where S_n is the last element of S_N, the solution is given by $S_a + (S_b - S_c) \times (S_4 + S_5 + \cdots + S_n) = T$. The probability that this algorithm works increases with N, which clearly violates the main property of NP-complete problems. Note that this algorithm can be altered to make it applicable to cases with smaller N, by searching for subsets within S_N for which $T = S_a + S_b$ and $S_c = S_d + S_e$, implying a solution $(S_a + S_b) + (S_c - S_d - S_e) \times (S_6 + S_7 + \cdots + S_n) = T$. A similar line of reasoning applies to S_O = (+, −, /) and S_O = (+, −, ×, /).

References

Braithwaite, D. W., Goldstone, R. L., van der Maas, H. L. J., & Landy, D. H. (2016). Non-formal mechanisms in mathematical cognitive development: The case of arithmetic. *Cognition*, *149*, 40–55.

Eggen, T. J. H. M., & Verschoor, A. J. (2006). Optimal testing with easy or difficult Items in computerized adaptive testing. *Applied Psychological Measurement*, *30*, 379–393.

Eley, J. (2009). How much does the 24-game Increase the Recall of Arithmetic Facts? Available at: http://eric.ed.gov/PDFS/ED508367.pdf. Accessed June 10, 2016

Elo, A. (1978). The Rating of Chessplayers, Past and Present, Arco.

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, P. Feltovich, & R. R. Hoffman (Eds), *Cambridge handbook of expertise and expert performance* (pp. 685–706). Cambridge, Cambridge University Press.

Flaherty, J., Connoly, B., & Lee-Bayha, J. (2005). Evaluation of the first in Math Online Mathematics Program. Available at: http://explore.firstinmath.com/media/280/FIM_WestEDstudy.pdf.

Gierasimczuk, N., van der Maas, H. L. J., & Raijmakers, M. E. J. (2013). An analytic tableaux model for deductive mastermind empirically tested with a massively used online learning system. *Journal of Logic, Language and Information*, *22*, 297–314.

Groeneveld, C. M. (2014). Implementation of an adaptive training and tracking game in statistics teaching. In M. Kalz & E. Ras (Eds.), *Computer assisted assessment. research into E-assessment* (pp. 53–58). Switzerland: Springer International Publishing.

Hayes, B. (2002). The easiest hard problem. *American Scientist*, *90*, 113–117.

Jansen, B. R. J., De Lange, E., & Van der Molen, M. J. (2013a). Math practice and its influence on math skills and executive functions in adolescents with mild to borderline intellectual disability. *Research in Developmental Disabilities*, *34*, 1815–1824.

Jansen, B. R., Hofman, A. D., Straatemeier, M., Bers, B. M., Raijmakers, M. E., & Maas, H. L. (2014). The role of pattern recognition in children's exact enumeration of small numbers. *British Journal of Developmental Psychology*, *32*(2), 178–194.

Jansen, B. R. J., Louwerse, J., Straatemeier, M., Van der Ven, S. H., Klinkenberg, S., & Van der Maas, H. L. (2013b). The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and Individual Differences*, *24*, 190–197.

Kadengye, D. T., Ceulemans, E., & Van den Noortgate, W. (2014). A generalized longitudinal mixture IRT model for measuring differential growth in learning environments. *Behavior Research Methods*, *46*(3), 823–840.

Klinkenberg, S., Straatemeier, M., & Van der Maas, H. L. J. (2011). On the fly item calibration using a new CAT procedure for computerized student practice and monitoring of maths ability. *Computers & Education*, *57*, 1813–1824.

Kurzen, L. (2011). Some Ideas for the Cijferstaak. Internal report, University of Amsterdam

Maris, G., & Van der Maas, H. L. J. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*, 615–633.

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104, No. 9). Englewood Cliffs, NJ: Prentice-Hall.

Nyamsuren, E., der Van Maas, H. L., & Taatgen, N. A. (2015). How does prevalence shape errors in complex tasks? *International Conference on Cognitive Modeling*, *13*, 160–165.

Van der Maas, H. L. J., & Wagenmakers, E. (2005). A psychometric analysis of chess expertise. *The American Journal of Psychology*, *118*, 29–60.

Van der Ven, S., Van der Maas, H. L. J., Straatemeier, M., & Jansen, B. R. J. (2013). Visuospatial working memory and mathematical ability at different ages throughout primary school. *Learning and Individual Differences*, *27*, 182–192.

Wainer, H., Dorans N. J., Flaugher R., Green B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.

Willingham, D. T. (2007). *Cognition: The thinking animal* (3rd ed). Englewood Cliffs, NJ: Pearson/Prentice Hall.