Cognitive Analysis of Educational Games: The Number Game

Cognitive Analysis Of Educational Games: The Number Game

Han L.J. van der Maas[1] & Enkhbold Nyamsuren[2]

[1] Department of Psychology, University of Amsterdam, Netherlands

[2] Welten Institute, Open University, Netherlands

Author Note

Correspondence should be addressed to Han van der Maas at h.l.j.vandermaas@uva.nl

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.

Abstract

We developed online educational games for practicing and monitoring math and language skills using innovations in computerized adaptive testing. The educational games generate enormous amounts of rich data on children's development. We first describe this system and discuss advantages and disadvantages of this approach to the study of cognitive development. Second, we present results on one example, the number game. In this creative math 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)*9). We argue that this game is NP complete, by showing its relation to the well- known partition problem. We propose heuristics based on the distinction in forward and backward reasoning. 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 1700 items. Item difficulties and the structure of correct answers are analysed.

Keywords: education, games, arithmetic, reasoning, NP-complete, number game

Cognitive Analysis Of Educational Games: The Number Game

Math Garden and Language Sea (Rekentuin.nl and Taalzee.nl) are popular Dutch educational websites for practicing and monitoring math and language skills in a gamified environment. They originated in basic research on the dynamics of cognitive development. The human cognitive system, and especially its development in the first 10 years, is inconceivably complex. Of all complex systems that are now studied in science, such as ecosystems, the climate, the immune system, and stock markets, the developing human cognitive system is by far the most challenging (van der Maas et. al, 2013). One common element in the study of the dynamic complex systems is the focus on high quality and high frequent measurements. This is clearly a weak spot in the study of cognitive and scholastic development. It has been very difficult to develop reliable and valid measurement instruments for cognitive development that can be used in a large age range. It is even more difficult to construct these instruments in such a way that they can be used in high frequent measurement, say once a week or once a day. Next, even when such tests were available, schools would certainly refuse to allow researchers to test children daily or weekly in the schools.

Our solution to these problems, or at least some of these problems starts with the observation that children do make exercises in math and language daily. Actually, schoolchildren spend about 50% of their time on learning reading, writing, and arithmetic (OECD, 2006). A substantial part of this time is devoted to exercises. If we could obtain the data on these exercises, a new measurement system could be in reach. Our first attempts to use data of existing exercise methods used in schools (written booklets) were however not successful. 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, 2000). Modern test theory consists of a number of item response models that specify the probability of a correct answer given the ability of a person and the difficulty of the item (and in more complex models other person and item characteristics). The person ability and item difficulty are expressed on the same latent ability scale. In the simplest model, the Rasch model (Rasch, 1960), the probability correct equals .5 when the ability of the person difficulty equals item difficulty. If the ability is much higher the probability goes to 1, and vica versa to 0. Person abilities and item difficulties correlate strongly with more traditional test indices such as sum scores for persons and item p-values for items. However, there are important advantages of the Rasch model and its successors. The advantage relevant for our approach concerns adaptive testing. It simply means that persons do not have to make complete tests but are presented with items depending on the successes and failures on earlier items. Based on earlier 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.

Computer adaptive testing is however not directly useful for computer adaptive practicing. To apply these techniques in a practice system that children and schools would be willing to use on a daily basis, we had two problems to solve. The first problem concerns pre-testing. The procedure of computerized adaptive testing only functions when all item difficulties are known. This means that at least hundreds of persons have to be tested on hundreds of items per task, before one can start a CAT. Since our systems Math Garden and Language Sea consists of about 40 games with in total more than 40.000 items, pre-testing is out of the question. The second problem is that in CAT the most informative next item is an item for which the expected probability correct is about .5. In a practice system, a failure rate of 50% is unacceptable. It is not difficult to select more easy 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 the 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 Graden and Language Sea, we use the same system with some modifications. Persons play items, and increase in rating (ability) when they solve the item, and decrease in rating when they fail, and vice versa 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 seconds, a correct answer in 5 seconds gives a score of +15, whereas an error in 15 seconds 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 in our extended Elo system that is used in Math Garden and Language Sea. 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 case of a correct answer, the remaining coins are added to the total number of coins children collected. In 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.

Math Garden and Language Sea

Based on these innovations we developed Math Garden, and later Language Sea, as a practice and monitoring websites for schools. Math Garden consists of a garden with plants, each representing a Math game, that grow with the increase in math ability. Currently Math Garden contains games for the basic arithmetic operations, but also for counting, series, fractions, clock reading, as well as more cognitive abilities, such as working memory, deductive reasoning, perceptual intelligence and math.

These online games let children practice intensively at their own level with direct feedback, two important requirements of deliberate practice (Ericcson, 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 and Language Sea are self-organizing additional learning tools that don't 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 and Language Sea became quickly popular in the Netherlands. Thousands of 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 200.000 children of preliminary elementary schools in the Netherlands use these tools regularly. During weekdays, about 1.5 million item responses are collected with a speed of 100 per second at peak hours.

Scientific analysis

The data collected in this way have clear advantages and disadvantages. On the one hand, these 'big' data are extremely promising. They contain high frequent 'modern test theory' measurements of the development of wide range of abilities from children of wide range of ages, collected in a natural learning environment. In a sense, they provide a new window on cognitive and scholastic development.

On the other hand, there are several important issues to take into account. First, it is important to guarantee the privacy of users and to de-identify the data carefully. Second, data are collected with a variety of tablets and computers, in different places, and children might occasionally receive help. In general, internet data are less reliable than data collected in the lab. Finally, big data analysis is often explorative. Explorative analyses suffer from the risk of data fishing. In the analysis of Math Garden data, many, often rather arbitrary, choices have to be made about item and person selection, handling of missing data, choice of statistical technique, etc. Our general solution is to check the robustness of results with different data selections and different analytical techniques. Only robust results are reported.

For each game, the data consist of three datasets. One dataset concerns ability estimates (ratings) of players, their ages and sex, the number of items played and the dates of first and last change in rating. The second dataset consist of item ratings, dates of first and last changes, and additional tags that describe the item (for instance, irregular verb). The final dataset is a logfile of each item played. It contains among others, item and user information, ratings, RT, and answer. It is thus possible to follow rating changes over time for person and items, to analyze response times, and to analyze response errors.

A number of papers using Math Garden and Language Sea data have been published (Gierasimczuk, van der Maas & Raijmakers, 2013; Groeneveld, 2014; Jansen, De Lange & Van der Molen, 2013;. Jansen, et al 2014; Jansen, et al, 2013; Kadengye, Ceulemans & Van den Noortgate, 2014; Nyamsuren, van der Maas & Taatgen, 2015; Van der Ven et al, 2103).

One example concerns the counting game. In this game, children have to count fishes in an aquarium. The number of fishes in items varies from 1 to more than 50, and displays can be ordered (for instance dice patterns) or random. Time limit per item is 20 seconds. 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, van der Maas and Raijmakers (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 37 thousand children that played the game regularly. More recently, Nyamsuren, Van der Maas & Taatgen (2015) explored the nature of errors in a SET game depends on various factors such as progression of game play, past experience with the game, strategy and a structure of a specific game instance.

The number game

The number game is one of the games in Math Garden designed to study and improve player's mathematical reasoning skills. It 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 '/'.

Figure 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)".

Figure 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 1650 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.

When always four numbers are 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 played in many schools over the world as an educational game. Although the game has not been analysed from a cognitive science perspective before, the Number game clearly requires fluency in basic arithmetic skills and something that we could call mathematical creativity. Only a few scientific sources describe the game (Flaherty et al., 2002; Eley, 2009). These unpublished papers 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).

The creative part of playing the number game is due to the complex search space of the game. 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 the NP problems particularly difficult even for computers. As in the Number task, checking solutions of NP problems is relatively easy.

To understand the complexity of the number task further, it is useful to compare the task 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 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' 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).[2]

How do humans solve the number problem?

An exhaustive systematic search of the problem space is clearly not an option for human players. Yet, 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 forwards 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, 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 forwards 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:

$$Convergence = \begin{cases} \frac{R_1}{T}, & T > R_1 \\ \frac{T}{R_1}, & T < R_1 \end{cases}$$

The degree of convergence was calculated using two different ratios since convergence can occur either upwards (for multiplication and summation) or downwards (for subtraction

and division). Convergence varies between zero and one with one indicating the fastest convergence on *T*.

A typical example of the application of the proximity heuristic is the solution 12 x 5 - 3 - 2 for S_N = (2, 3, 5, 12) and *T* = 55. It is greedy since the biggest numbers are used first and convergence is large (55/60 = .92). Many number puzzles are solvable in this way. Below we will test whether the proximity heuristic is indeed dominant in solving number puzzles by humans by analysing the difficulties of items as they are estimated in Math Garden.

The second general heuristic is based on backwards reasoning. It starts by applying operators of S_O to the target and on one of the numbers 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 = (+, -, x, /), the proximity heuristic fails. But 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 backwards-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 behaviour 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 backwards reasoning are more difficult than items that require forwards 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 in the analysis of item difficulty.

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, Straatemeier, & van der Maas, 2011) and informative on the characteristics that make items difficult (e.g. Gierasimczuk, van der Maas and Raijmakers, 2013).

We will first analyse the types of correct answers given to different types of items. Second, with multiple regression analysis we will investigate item difficulty in the Number game. Third, the results of the regression analyses are illustrated within sets of items. Finally, the results of a small control study are presented. 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 177880 players of which 51405 played more than 100 items.

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 (49681 out of 95942 answers, with 11323 for the second most popular answer).

There are 315 such items that were played at least 30 times in Math Garden. As it is shown in Figure 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 (Figure 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*. Figure 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 ($r(314) = -0.48$, $p < 0.001$) indicating that players' strategy is dependent on a greedy approach. Players find items where first operations converge closer 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 $Size(S_N)$, since higher $Size(S_N)$ inevitably results in increased item difficulty and decreased ratio. To further verify if the significant correlation is an 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 $r(144) = -0.14$, $p = 0.09$ for $Size(S_N) = 3$, $r(76) = -0.25$, $p = 0.02$ for $Size(S_N) = 4$, and $r(67) = -0.54$, $p < 0.01$ 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.

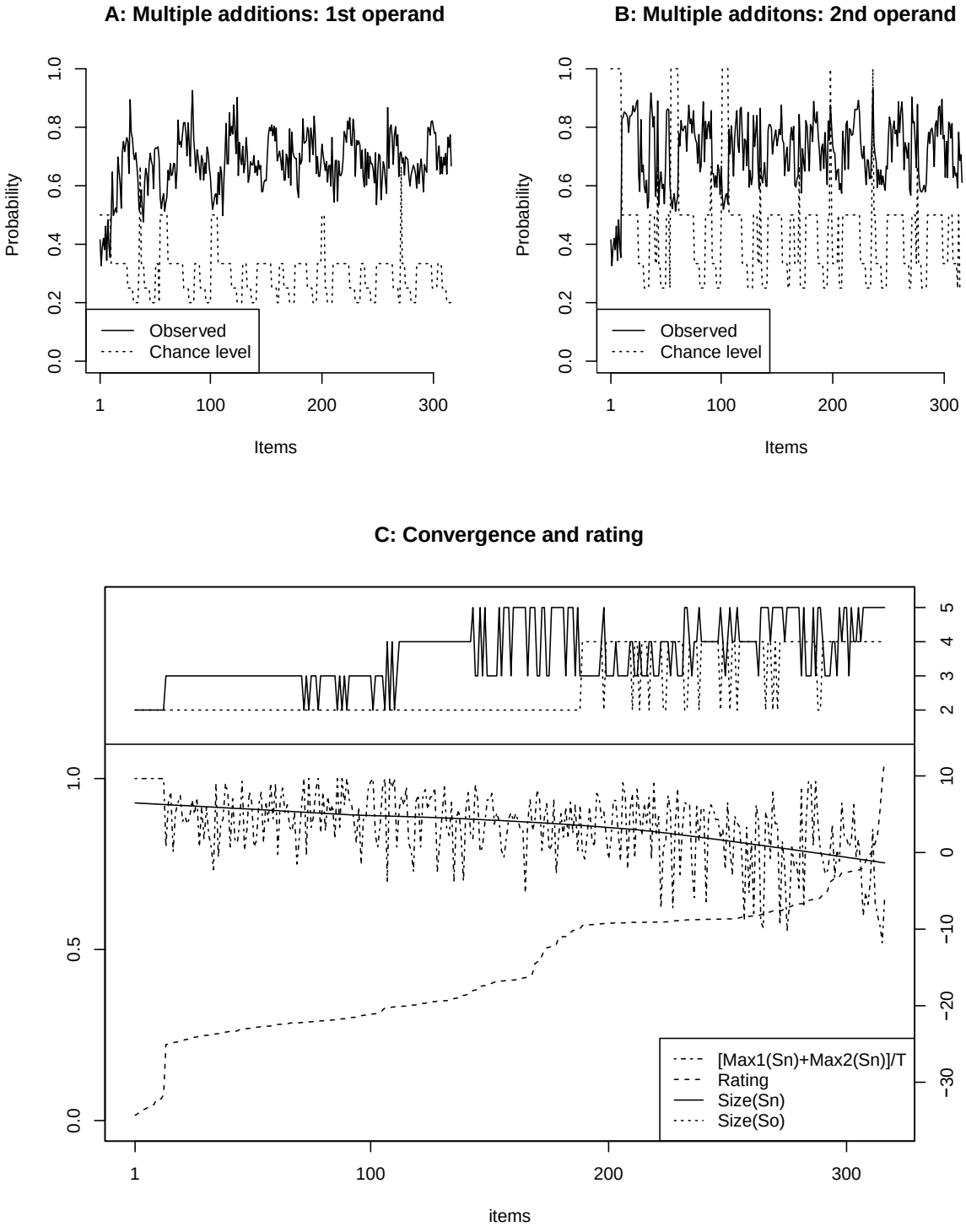

Figure 2: The data is shown for 315 items that were played at least 30 times, have positive target number and require only addition. For each item, the graph shows observed proportions of trials in which (a) the biggest number from S_N, $Max_1(S_N)$, was used as the first operand of the first operation, 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 ratings and ratio of the sum of two largest numbers to the target number, $(Max_1(S_N)+Max_2(S_N))/T$. Smoothing curves are also shown for both data types. Items in panel c were ordered by increasing ratings of difficulty. The upper part of the graph shows $Size(S_N)$ and $Size(S_O)$ for each item.

It is possible that the proximity heuristic is used in addition-only items since 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 may also play important role in those items. A similar bias toward $Max_1(S_N)$ as in addition-only items is 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 were random. 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).

Regression analysis

Variables for greediness and convergence together with other items' properties were included in a linear regression analysis. Results are reported in Table 1. The intercept indicates the difficulty rating of an item with $Size(S_N) = 3$ and $Size(S_O) = 2$. *NumbersIncrease* and *OperatorsIncrease* indicate increases in $Size(S_N)$ and $Size(S_O)$, respectively. *Fractional* is one if T is a fractional number and otherwise zero. *UseAdd*, *UseSubstract*, *UseMultiply* and *UseDivide* are one if corresponding operators are used at least once in the solution and otherwise zero. *Greedy* is one if the solution is greedy and zero otherwise. Finally, *Convergence* is the degree of convergence between zero and one.

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, while,

interestingly, subtraction seem 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.

Table 1:

Linear regression model on items' difficulty ratings: $R^2 = .81$, $F(10, 1085) = 451$, $p < 0.001$.

Predictors	Coefficients	SE	*t* values	*p* values
Intercept	-10.27	0.76	-13.53	< 0.001
NumbersIncrease	0.27	0.23	1.17	0.243
OperatorsIncrease	4.38	0.19	23.00	< 0.001
Fractional	5.47	0.76	7.22	< 0.001
UseAdd	2.57	0.44	5.80	< 0.001
UseSubstract	8.46	0.35	24.43	< 0.001
UseMultiply	3.55	0.41	8.68	< 0.001
UseDivide	3.50	0.46	7.70	< 0.001
Greedy	1.83	0.71	2.57	0.010
Convergence	-1.84	0.83	-2.22	0.026
Greedy:Convergence	-9.33	1.20	-7.77	< 0.001

To explore the interaction effect we analyse 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 convergences 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 prediction is supported by empirical evidence shown in Figure 3. There are 8 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. However, multiplications in these items are not greedy and do not involve both largest numbers from S_N. Similarly, there are 5 items that require a greedy approach but on addition as the first operation. All these items have a varying degree of difficulty not forming any cluster. These results indicate that the combination of greediness and fast convergence makes items considerably easier.

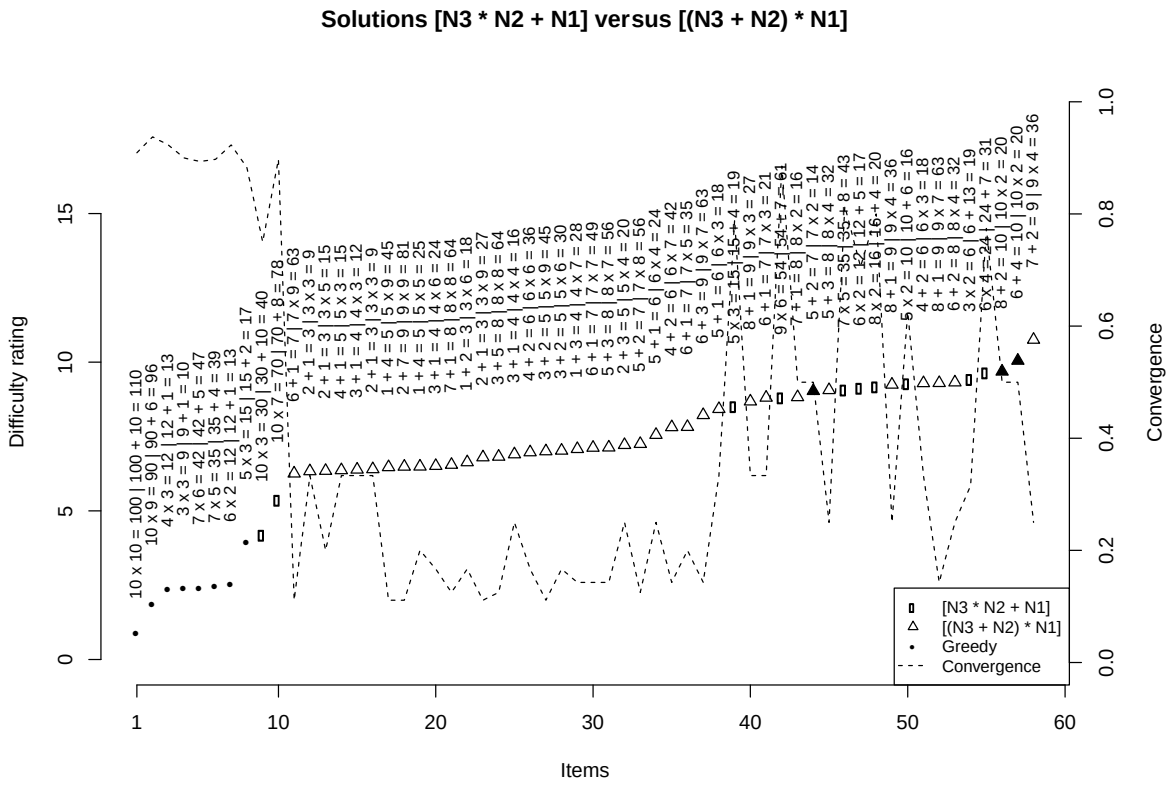

Figure 3: Items that have three numbers and require exactly one addition and one multiplication. Items are shown in increasing order of difficulty ratings. Oval plot points indicate items with a multiplication as a first operation: $N_3 * N_2 + N_1$. Triangles indicate items

that require addition first: $(N_3 + N_2) * N_1$. The striped line depicts convergence to the target number calculated as a ratio of first operations results to the target number: $[N_3 * N_2 / T]$ or $[(N_3 + N_2) / T]$. Black circles and triangles (in contrast to open) denote greediness.

Our second illustration focuses on an even smaller set of items. Figure 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 with only one item requires only addition and 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.

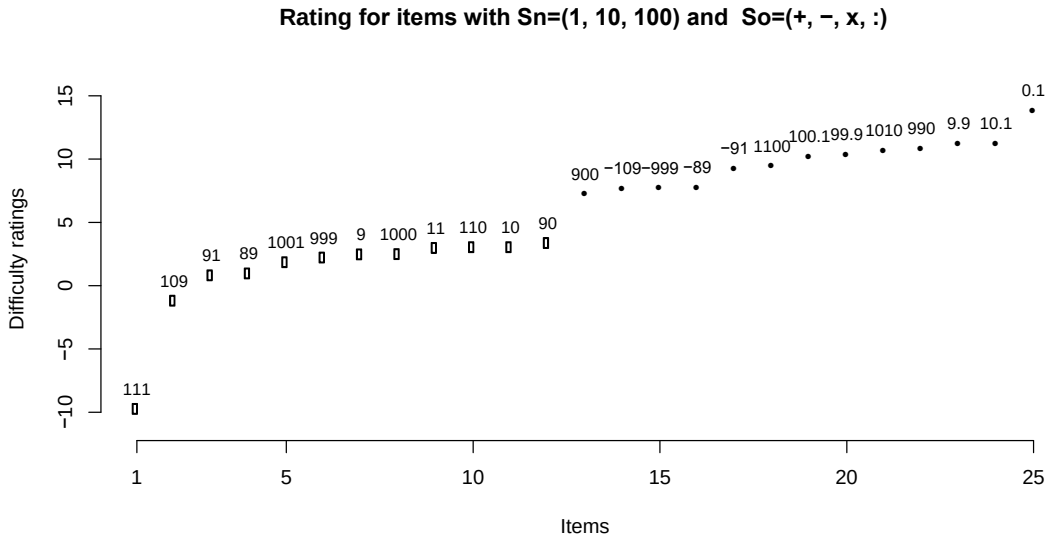

Figure 4: List of items ordered by difficulty ratings. All items have S_N = (1, 10, 100) and S_O = (+', '−', '×', '/'), but require different combinations of operations to reach target numbers shown above plot points. The first 12 items are compatible with the proximity heuristic. The difficult items also often have fractional targets, but note that targets 990 and 1010 belong the most difficult items. With backwards reasoning (dividing the target by 10) the solution is easily found, but the first forward step (100 ± 1) is not greedy and has low convergence.

Alternative explanations

The preference for forwards reasoning in Math Garden's number task might be due to the user interface of the game as shown in Figure 2. First, the numbers in S_N are always presented in increasing order with the biggest numbers closest to the operators S_O (see figure 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 (mean age 21.9, *SD* = 3.5) participated in this study. None of them had a 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 backwards heuristic applies well (the backward items). These forward items were: S_N = (7,10,100), T = 1007; 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 = 3002. The backward items were 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 = 1600. The tests were presented in a paper and pencil format and subjects were allowed to notate their answers in any format. The order of items was randomized. The time limit per item was 60 seconds.

The average percentage correct on the forward items .95 (SD = .12) was much higher than on the backward items 0.53 (SD = .32), t(55) = 9.2, p<.001. Subjects were also much faster on forward items, 15.28 (4.94) than the backwards items 37.13 (12.26), t(55) = -14.5, p<.001. Both differences were highly significant. Hence, the preference for forwards reasoning with the proximity heuristic was replicated with older subjects, not trained in Math Garden, and with an answer format that does not provoked 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 one of 1362 solvable quadruples. By June 2015, 604985 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 percentage correct after a logarithmic transformation to normalize its distribution. Similar to earlier regression analysis, we also included as predictors item's greediness, convergence ratio and types of operation involved. However, all operations except division had insignificant coefficients, and therefore were removed from the final regression analysis. We used centered values of greediness (0 or 1) and convergence (between 0 and 1). The results are shown in Table 2 and indicate that division considerably and significantly increased difficulty of quadruples requiring it. Greediness, convergence and their interaction contributed significantly. These results replicate effects from Math Garden and indicate that the combination of greediness and fast convergence makes quadruples easier. Therefore, we can reasonably assume that proximity heuristic is also frequently used by players in the 24 game outside of Math Garden.

Table 2:

Linear regression model on quadruples' difficulty ratings: $R^2 = .11$, $F(4, 510) = 15.01$, $p < 0.001$.

Predictors	Coefficients	SE	*t* values	*p* values
Intercept	2.41	0.02	107.8	< 0.001
UseDivide	0.12	0.04	2.97	< 0.01
Greedy	-0.17	0.04	-4.66	< 0.001
Convergence	-0.13	0.06	-2.26	<0.05
Greedy:Convergence	-0.22	0.11	-1.98	<0.05

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 paper, 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, attest 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 (Flaherty et al., 2002; Eley, 2009). It was also not the focus of the analyses in this paper.

We are primarily interested in the cognitive processes involved in solving number task problems. As there is no theory or data available, we only made the first steps here. By relating the Number task to the well-known partition problem, we discovered that the search problem of the Number task is extremely hard. Many instances of the Number task are NP complete (but see footnote 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 is 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 backwards reasoning heuristics in the number task. Next step should include investigation of underlying cognitive processes that implement these heuristics and allow humans to solve complex problems. The ultimate goal is to create biologically inspired algorithms that can tackle NP-complete problems with high accuracy and efficiency.

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_0, 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 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 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 that did not make them easy. We replicated the preference for the proximity heuristic with a paper and pencil task in college students. The preference for forward reasoning is not due to the setup of Math Garden, the layout of the Number task, or the age group.

These findings are only the first discoveries in the study of the Number task. 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 is 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 backwards reasoning or heuristics that we not have yet detected. Combinations of backwards and forwards reasoning are possible too. Take for instance SN={1,3,4,9}, T=111. After the backwards step 111/3 = 37, the remaining problem SN={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 heuristic more advanced than the proximity heuristic, study of expert players might reveal such forms of reasoning.

New hypotheses on Number task problem solving can be investigated with the Math Garden data set. 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 task in Math Garden to evaluate training methods in number task problem solving. It might be the case that players do not use backwards reasoning spontaneously, but do use it after some training. Our follow-up training study with college students from which we only reported the pre-test data, concerned such training. Since it was only a small scale pilot study with very short training sessions (8 minutes) we decided not to report the results. But the main result was that there were no positive effects of training compared to the control group. The only effect was negative effect of backward training on

the forward items. This should not be taken too seriously but it indicates that training of more advanced heuristic should not be taken lighthearted.

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

References

Borgs, C., Chayes, J., & Pittel, B. (2001). Phase transition and finite-size scaling for the integer partitioning problem. *Random Structures and Algorithms, 19,* 247-288.

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? Retrieved from http://eric.ed.gov/PDFS/ED508367.pdf

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. Retrieved from http://explore.firstinmath.com/media/280/FIM_WestEDstudy.pdf

Gent, I. P., & Walsh, T. (1996). Phase transitions and annealed theories: Number partitioning as a case study. In W. Wahlster (Ed.), *Proceedings of the 12th European Conference on Artificial Intelligence (ECAI-96)* (pp. 170-174). New York : John Wiley & Sons.

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 *Computer Assisted Assessment. Research into E-Assessment* (pp. 53-58). 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. (2013). 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. J., Louwerse, J., Straatemeier, M., Van der Ven, S. H., Klinkenberg, S., & Van der Maas, H. L. (2013). The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and Individual Differences, 24*, 190-197.

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.

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

Larkin, J., McDermott, J., Simon, P. D., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science, 208*, 1335-1342.

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.

Mertens, S. (2001). A physicist's approach to number partitioning. *Theoretical Computer Science, 265*, 79-108.

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

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

OECD (2006). Education at a glance. Parijs: OECD (PAC).

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research

van der Maas, H. L. J. & Wagenmakers, E.J. (2005). The Amsterdam Chess Test: a psychometric analysis of chess expertise. *American Journal of Psychology, 118*, 29-60.

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

van der Maas, H. L. J., Kan, K. J., Hofman, A., & Raijmakers, M. E. J. (2013). Dynamics of development: a complex systems approach. In P.C. Molenaar, Lerner, R. M., & Newell, K. M. (Eds.). *Handbook of developmental systems theory and* methodology (pp. 270-286.). Guilford Publications, New York.

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.

Footnotes

[1] There are good reasons to regard these three solution as similar (see http://www.4nums.com/theory/)

[2] However, this does not prove that the number game with S_O = ('+', '−', 'x', '/') is NP-complete. It is easy to see that number problems with S_O = ('x', '/') are equivalent to the S_O = ('+', '−') case by taking logarithms of all elements in *Sn*. But the case S_O = ('+', '−', 'x')) is different. Suppose n of *Sn* is extremely large, then *T* will be an element of *Sn.* We call this element *Sa. Sn* will also hold two equal numbers *Sb* and *Sc* having difference 0. After reordering *Sn to (Sa, Sb, Sc, S4…Sn),* the solution is given by *Sa + (Sb-Sc)*(S4+S5+…+Sn) = T*. The probability that this algorithm works increases with n, which clearly violates the main property of NP-complete problems. Note this algorithm can be altered to make in applicable to cases with smaller n, by searching for subsets within *Sn* for which *T = Sa + Sb* and *Sc = Sd + Se*, implying a solution (*Sa + Sb)* + (*Sc – Sd - Se*) x (*S6 + S7 +…+Sn*) = *T*.