 doi:10.20944/preprints202010.0314.v1

Article

Creative Mathematical Thinking in a Numbers Game

Joost Kruis[1,*], Claire Stevenson [1] and Han L. J. van der Maas [1]

[1] Psychological Methods Department, University of Amsterdam, Amsterdam, NL
* Correspondence: j.kruis@uva.nl

Abstract: Creative thinking is increasingly recognised as an essential ability that should be part of school curricula.Given the move towards online learning and assessment, we investigate whether mathematical creativity can be assessed at-scale in the *Numbers* game, an arithmetic game in Math Garden, a popular online math practice platform. In the *Numbers* game, a generalisation of the 24 Game, children are asked to figure out how to compute a target number using basic arithmetic operations and a given set of numbers. We argue that creative thinking is required when the search space is complex, and propose that the base-pattern, i.e., the sequence of the operations needed to solve a *Numbers* game item, indicates search space complexity. We then demonstrate that items with disordered base-patterns are more likely to require mathematical creativity to figure out. Specifically, our analysis shows that for items with only one solution sequence, those with disordered base-patterns are more difficult and take longer to solve compared to items with ordered base-patterns. For items where multiple solution sequences are possible, nine times out of ten children choose ordered over disordered base-patterns. We conclude that the *Numbers* game has potential for assessing mathematical creativity at-scale.

Keywords: creative thinking, mathematical creativity, response processes, digital learning environment, online learning platform

1. Introduction

There is a growing consensus that creative thinking is an important skill that should be nurtured and developed from an early age[1]. For example, PISA 2022, a worldwide evaluation of education systems that measures the scholastic performance of 15-year old students, will focus on creative thinking[2,3]. In addition, education is increasingly shifting from a physical to an online setting, whether as a function of technological advancement or forced through disruptive events, such as a global pandemic. As such, assessing creative thinking in large scale online settings, to help students develop their creative thinking skills, has become essential.

In theories about the creative thinking process, a common distinction is made between divergent and convergent thinking[4]. Divergent thinking is the process of generating multiple possible responses to a problem, and convergent thinking is the process of identifying the single correct solution to a problem. As these two concepts are considered paramount to the creative thinking process, it should not come as a surprise that well-known tasks of creative thinking focus on these two processes. Notable tasks used for measuring creative thinking, such as the Torrance Test of Creative Thinking[5] (TTCT), the Wallach–Kogan Creativity Test[6] (WKCT), and Guilford's Alternative Uses Task [4](AUT), focus on divergent thinking, whereas the visual and classic Remote Associations Tasks [7,8](RAT) aim to measure associative skills, convergent thinking and insight[9–12]. The more recently introduced *Evaluation of Potential Creativity*[13–15] (EPoC) battery, includes measures of both divergent and convergent thinking with graphic, numeric and verbal items.

The TTCT, WKCT and EPoC creativity tests all consist of several verbal and visual tasks. In one task, the participant is given a problem and has to come up with as many solutions as possible, for example, "If all the shops are closed, how would you feed yourself?". In another typical verbal divergent thinking task, the participant is provided with a basic shape and has to use this shape to

 doi:10.20944/preprints202010.0314.v1

create an original image. Responses to these tasks are rated on criteria such as fluency, flexibility, originality, uniqueness and elaboration. In the AUT, the participant has to come up with as many uses as possible for a common object, such as a brick. In most cases the creativity measure is established with the so-called Consensual Assessment Technique[16,17], where expert raters judge the responses on creativity as a whole, or as a composite of originality and usefulness.

Scalability is a significant challenge for creative thinking tasks because the creativity of each separate response needs to be judged by multiple raters. In recent years there have been several attempts to take out the human rater. For example, in computerised versions of the WKCT [18,19] and the TTCT [20,21] the creativity of a response is determined by how often it occurs in a database. The benefits of being able to determine the creativity of a response using software, instead of human raters, are easily recognised, as the response can be judged faster and more consistently. A limitation of this approach is that each possible response has to be encoded in the database to be able to judge its creativity. When a new response arises, the creativity of it has to be evaluated by human raters. This is because not only the originality, which would be very high for a word that has a frequency of zero in the database, has to be assessed, but also the usefulness in order to determine if the response is creative. Furthermore, if the judgements in a database are unreliable, so will the judgements of the actual responses.

A partial solution to this has been developed for AUT types of tasks and uses natural language processing techniques (NLP), to automatically score the creativity of responses based on the distance between the response and some reference point created from an initial word corpus [22–25]. While these approaches are very promising, there are still several limitations. For example, NLP-based solutions may be biased with respect to the number of words in a response [26]. Furthermore, the corpus used for NLP strongly influences what is considered creative because it determines the semantic structures from which a creativity score is derived.

In the RAT[7], participants are given three seemingly unrelated words and instructed to find the concept that ties them together. For example, THOUGHT - STEAM - BEARER are connected by the word TRAIN. In contrast to divergent thinking tasks, the RAT requires a participant only to give a single solution, and the creative thinking process is assumed to take place by finding associations between the presented words. For convergent thinking tests such as the RAT, in which only the set of correct responses have to be encoded in the software, scalability is much less of a problem. One drawback, however, is that scores on the RAT are strongly related to verbal ability[11], as such, two people with equal creative abilities might be judged differently as a function of differences in their vocabulary. Recently a visual RAT was introduced that uses graphical representations of the stimulus words[8]. However, as the response still is given in a verbal format, somewhere a translation from a visual to a verbal representation of the task is necessary.It is thus not necessarily the case that proficiency on this task is entirely unrelated to verbal ability.

These challenges are largely absent in the *Numbers* game, which is a generalisation of both the 24 Game created by inventor Robert Sun in 1988[27] and the card game Krypto designed by Daniel Yovich in 1963[28,29]. In a *Numbers* game item participants are presented with a set $x = [x_1, x_2, \ldots, x_n]$ containing n numbers, a set O containing k arithmetic operators from the set $\{+, -, \times, \div, \wedge\}$, and a single target number T. The goal is to find the arithmetic expression that results in T, using operators from O to combine the numbers in x. In Krypto, six numbered cards are drawn from a deck of 56 cards, the numbers on the card ranging from 1 through 25, and the player must use all numbers from the first five cards once, again with any combination of summation, subtraction, multiplication and division, to obtain the target number on the sixth card. Krypto game items are a subset of the *Numbers* game items in which the dimensions of all items are fixed to $n = 4$, $x \in [1, 25]$, $O = \{+, -, \times, \div\}$ and $T \in [1, 25]$. In the 24 Game a player has to make the number 24 from four numbers shown on a card, using each of these numbers once and any combination of summation, subtraction, multiplication and division. The 24 Game type of items are thus a subset of *Numbers* game items in which the dimensions of all items are fixed to $n = 4$, $O = \{+, -, \times, \div\}$ and $T = 24$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

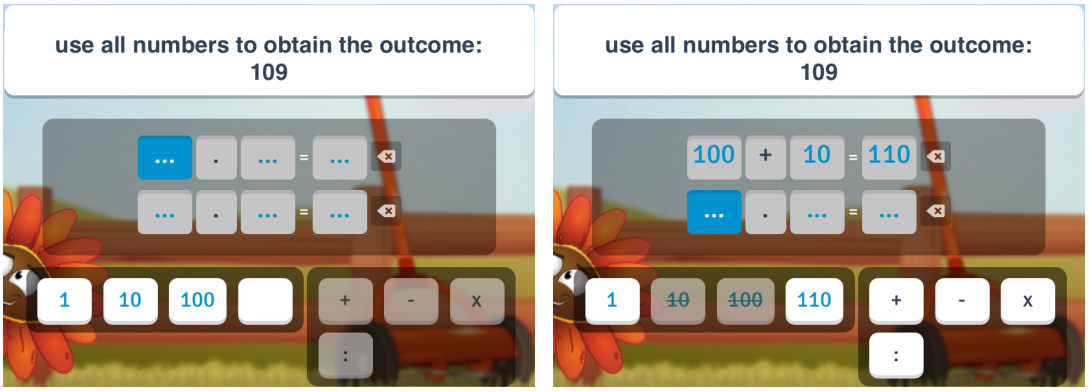
Figure 1. 24 Game card with the numbers 2, 6, 6 and 9. The goal is to use each number once and any combination of summation, subtraction, multiplication and division, to obtain the outcome 24.

For example, in Figure 1 a 24 Game card with the numbers 2, 6, 6 and 9, is shown, and two of the possible correct solutions would be

$$\left.\begin{array}{l} 9-6=3 \\ 6+2=8 \end{array}\right\} 3\times 8=24 \qquad \textbf{OR} \qquad \{9+6=15\ \}\ 15\times 2=30\ \}\ 30-6=24\} \tag{1}$$

The First in Math program, an online version of the 24 Game, is used in schools to help children develop math skills[30], and reports over 25 billion problems solved[31]. The 24 Game also has a large host of players online on other dedicated apps and websites, one of which reports almost 3.5 million 24 Game type puzzles solved thus far. Even though it is a popular game, only a handful of, mostly unpublished, studies exist on the 24 Game [32–34].

The *Numbers* game is one of the games available in the Math Garden[35,36], a large scale online computer-adaptive system for helping primary school children development their math skills in the Prowise Learn environment. Figure 2 shows an item from the Math Garden *Numbers* game, with $x = [1, 10, 100]$, $O = \{+, -, \times, \div\}$, and $T = 109$. The interface makes brackets unnecessary, i.e., the outcome of the first operation on the first line automatically appears as the first number for the second operation (see Figure 2b).



(a) **(b)**

Figure 2. Math Garden *Numbers* game item with $x = [1, 10, 100]$, $O = [+, -, \times, \div]$, and $T = 109$. (**a**) Initial display of the item. (**b**) Updated display of the item after choosing the first operation.

Continuing from Figure 2a, we can easily see that there are two sequences that lead to the target number; the first operation of one of these is shown in Figure 2b.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

$$\left.\begin{matrix}100+10=110\\1\end{matrix}\right\} 110-1=109 \qquad \textbf{OR} \qquad \left.\begin{matrix}10-1=9\\100\end{matrix}\right\} 100+9=109 \tag{2}$$

Van der Maas and Nyamsuren [34] argue that the *Numbers* game requires not only arithmetic skills, but also creative mathematical thinking because of its complex search spaces. If the *Numbers* game is capable of assessing mathematical creativity, adding this game to the current battery of creative thinking tests has several advantages. For one, item responses to this game are simple to score, i.e., they are correct or incorrect, and large numbers of items can be generated automatically and on the fly. Furthermore, in contrast to the RAT, *Numbers* game items only require basic arithmetic skills to solve. Specifically, given infinite time any person who can add, subtract, multiply, and divide would eventually come up with the correct response by going through all possible sequences of operations. In contrast, a correct response to the RAT could never occur if a particular word is not a part of the participants' vocabulary. Furthermore, with the Math Garden interface, it would also be possible to track part of the creative thinking process, which could enrich its assessment, by logging all operations the player performs to check the result of a combination, even if that outcome is not used in the final response.

It is easy to argue that creative thinking is required to solve items such as the infamous, $x = [1,3,4,6]$, $O = \{+,-,\times,\div\}$, and $T = 24$. However, for $x = [3,5,7,9]$, $O = \{+,-\}$ and $T = 24$, finding the correct sequence is pretty straightforward, and requires probably minimal if any creative thinking. An important question then is which *Numbers* game items require creative thinking to solve and can therefore potentially be used in large scale creativity assessment?

The structure of this paper is as follows. We start by reviewing the premise of Van der Maas and Nyamsuren [34] that creative thinking is required in problems with a complex search space. Next, we discuss the distinction between a complex search space as opposed to a large search space. We then propose three problem solving strategies for the *Numbers* game: random, reduction, and creative; and discuss how changes in the size and complexity of a search space (in)validate the use for each of these strategies. Next, we use the similarities between the *Numbers* game and so-called Mate-in-*m* chess problems to determine whether a search space is complex and hence requires creative thinking. We then derive the base-pattern of a *Numbers* game item as one of the simplest indicators for search space complexity and use data from the 24 Game and the *Numbers* game, to study the relation between these base-patterns and the accuracy of item responses and response times.

2. Theory

Creative thinking can be initiated when someone becomes aware of a gap, limitation or inconsistency in their perception[37–39]. When a creative process is subsequently used to close this gap, this requires from the person *'to break away from the usual sequence of thought into an altogether different thought'*[37], that is, the problem should *'... be looked at from a special, and often an unusual point of view, and that it should be recomposed and reinterpreted to achieve the desired state'*[38]. This latter idea is also found in the highly influential work of Boden, who envisions creative thinking as involving *'the exploration - and sometimes the transformation - of conceptual spaces in people's minds'*[40, p. xi].

> *'Conceptual spaces are structured styles of thought. ...Within a given conceptual space many thoughts are possible, only some of which may actually have been thought. Some spaces, of course, have a richer potential than others. Noughts and crosses is such a restricted style of game-playing that every possible move has already been made countless times. But that's not true of chess, in which the number of possible moves, though finite, is astronomically large.'*[40]

In the case of the *Numbers* game, we define the search space as containing all possible response patterns given the set of numbers in x, the operators in O. Specifically, as each item in which x has size n

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

requires $n-1$ operations, for an item with n initial numbers, and k operators, we have $n \times k \times (n-1)$ options for the first operation, $(n-1) \times k \times (n-2)$ options for the second operation, etc., such that the total number of possible operations forming a arithmetic expressions ($|P_r^{nk}|$) is given by:

$$\begin{aligned} |P_r^{nk}| &= \prod_{m=1}^{n-1} m \times (m+1) \times k \\ &= n! \times k^{(n-1)} \times (n-1)! \end{aligned} \tag{3}$$

Even though, as we will discuss later in more detail, not all these possible expressions lead to different outcomes, the amount of different possible outcomes becomes quickly too large to find a solution with a brute force approach in a feasible time. The concept of a space to be explored is also the basis for the claim by Van der Maas and Nyamsuren [34] that the *Numbers* game requires not only arithmetic fluency but also mathematical creativity due to its complex search space. They showed that items in which $O = \{+,-\}$ are equivalent to the NP-complete Partition problem[41–43], which means that solving time increases exponentially with the size of n and hence no efficient algorithms to solve a problem exist. However, while items with $O = \{+,-\}$ are thus NP-complete problems themselves, items with $O = \{+,-,\times,\div\}$ are not, as for the latter case algorithms exist that are increasingly likely to work with the growing size of x[33,34].

While a large search space might thus provide more potential for creative thinking, this does not mean that it is guaranteed. We do think that the premise of Van der Maas and Nyamsuren [34] is correct, i.e., we believe that creative mathematical thinking is required to solve a *Numbers* game item when the search space, or perhaps more appropriate, the route through the search space, is complex. The argument in Van der Maas and Nyamsuren [34] covers, however, only half the story, as distinguishing creative and non-creative items in the *Numbers* game, requires one to determine when the (route through) a search space is complex, and when it is not.

2.1. Three Strategies

To enhance the readability we first clarify some of the mathematical notations used throughout the section. The first distinction is between x with a numerical subscript (x_1), used to denote the number on the corresponding index of x, and x with a character subscript (x_a) to denote index non-specific numbers. For example, if $x = [1,2,4]$, $((x_1 + x_2) - x_3) = ((1+2)-4)$, whereas $((x_a + x_b) - x_c) \in \{((1+2)-4),((1+4)-2),((2+1)-4),((2+4)-1),((4+1)-2),((4+2)-1)\}$. We use the $\sim$ symbol to denote non-specific operations, such that for $O = \{+,-,\times,\div\}, (x_1 \sim x_2) \in \{(x_1 + x_2),(x_1 - x_2),(x_1 \times x_2),(x_1 \div x_2)\}$. Lastly, unless otherwise specified, we will assume that the set of permissible operators is the set $O = \{+,-,\times,\div\}$.

We can view a response to a *Numbers* game item as a sequence of $n-1$ operations of the form $(x_a \sim x_b)$. Given a simple setup, in which x consist of only integers between one and nine, it is often reasonable to assume that there are no individual differences in the ability to perform a single operation, i.e., most people can correctly perform any possible operation of the type $(x_a \sim x_b)$. As such, the difficulty of the *Numbers* game items will not be in performing the single operations themselves, but finding the right sequence of operations with the right combination of number to obtain the target number.

We can translate this response sequence to the concept of a search space by imagining the start of this space as a large hall filled with $n \times k \times n-1$ doors, in which each door represents a possible choice for the first operation. After we have chosen the first operation and walked through the corresponding door, we find ourselves again in a large hall now with $n-1 \times k \times n-2$ doors, in which each door now represents a possible choice for the second operation. This sequence goes on until we find ourselves in the final room with $2 \times k \times 1$ doors, and we make the final decision. The first question is how we decide which doors we should walk through to end up at the correct response; in other words, what is our

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

solution strategy. We propose the random, reduction and creative solution strategy as three different options that can be employed to solve problems such as a *Numbers* game item. As we argue next, the potential success of the first two strategies, which are dependent on the characteristics of the search space, determine if a creative solution strategy is required.

2.1.1. Random

The random solving strategy is a brute-force strategy comparable to unmethodically walking through doors in the hopes of at some point going through the door of the correct response. The probability of this strategy being successful is given by $\frac{R^c}{|P^{nk}|}$, in which R^c is the number of response sequences that lead to the correct solution. In general there will be multiple routes through the search space that lead to the same response, for example, while $|P^{nk}| = 4$ for $x = [2,3]$, $O = \{+,-\}$, and $T = 5$, two of those four sequences, i.e., $2+3=5$, and $3+2=5$ lead to the correct response, hence $R^c = 2$, and the probability of success for the random strategy would in this case be $^1/_2$. There are several forms of algebraic equivalencies that result in multiple routes leading to the same response. However, as the exact value for R^c is always dependent on the particular item characteristics, e.g., when x contains, ones, zeros, or duplicates, the number of unique responses will be lower compared to when x would contain only unique numbers higher than one. We will not go into specifics here and discuss this in more detail in the appendix.

Assuming that a person has unlimited time and can perform all separate operations, eventually, they would walk through the door ending up at the correct response. In practice, this is however not feasible because of time constraints, for example, in the Math Garden application in which there is a time limit of 20 seconds per item. Alternatively, single try restrictions, in which the process is terminated as soon as one walks through a door not leading to the correct response, make such a brute-force approach impossible. The success of the random solving strategy is thus primarily dependent on the search space size and, even with multiple routes leading to the correct response, with increasing n and k this strategy becomes increasingly less likely to work. So for items in which the search space is large and $\frac{R^c}{|P^{nk}|}$ has become too small for a random strategy to be feasible, another approach must be taken. That is, we need ways to reduce the size of the search space.

2.1.2. Reduction

A reduction strategy works by making (informed) assumption about the solution for (part of) the problem, such that the number of possible options for the variables of a problem, and hence the search space, are reduced to a manageable size. By using heuristics that are available for the combination of numbers, operations and the target, the problem, and with that, the search space, can be transformed. The resulting search space becomes smaller in size, and hence the probability of choosing the route leading to the correct response becomes larger. We describe five possible types of reduction heuristics, as several of these are discussed in Van der Maas and Nyamsuren [34] we will mention them here but not discuss them in detail. Naturally, the effectiveness of a each of the reduction strategies is dependent on the extent to which the heuristics and assumptions that underlie the transformation of the search space are applicable to, and hence, appropriate for the problem at hand.

Forward reduction strategy: A forward strategy reduces the search space by combining two values in x through any operation, i.e., $x_{ab} = x_a \sim x_b$, and using the remaining $n-1$ numbers as the reduced set $x_r = [x_{ab}, x_c, \ldots, x_n]$ to solve the item. If one would select the entries for $x_a \sim x_b$ at random, this would, of course, be nothing more than the first step in the random strategy, as such we expect there to be some thought behind the chosen numbers. In their paper Van der Maas and Nyamsuren [34] suggested greediness and convergence with respect to the target number as heuristics to reduce the search space. Greediness entails that one chooses the largest number in x for x_a and the second-largest number for x_b, and convergence denotes the distance from the target number to the outcome of the first operation. For example, when $x = [1,2,10,15]$, and $T = 74$, a greedy approach entails using $15 \sim 10$ as the first operation. Naturally, both these can be used independently of one another; one can shrink

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

the search space with the assumption that the first operation takes place on the two largest numbers, or choose the operation that has the largest convergence with respect to the target number.

Backward reduction strategy: Whereas forward strategies reduce the size of x and hence the search space by combining numbers in x using heuristics, the backwards reduction strategy does this by combining a number in x and T as such making x smaller and changing T. For example, when $x = [2, 10, 15]$ and $T = 35$, a backward strategy would be $T_r = T - 15 = 35 - 15 = 20$, and $x_r = [2, 10]$, which is easy to solve.

Operator reduction strategy: The operator reduction strategy revolves around making assumptions about which operators are most likely necessary or unnecessary to arrive at the target value. In the positive operator reduction strategy, the size of the search space is shrunk by assuming that a particular operator should probably be used to arrive at the target value. For example, if $x = [2, 3, 5]$ and $T = 30$ most people will recognise that a multiplication operation is probably needed, hence the search space shrinks by excluding all those sequences containing no multiplication operation. In the negative operator reduction strategy, on the other hand, the size of the search space is shrunk by assuming that a particular operator is probably not necessary to be used to arrive at the target value. For example, if $x = [2, 3, 5]$ and $T = 30$ most people will recognise that using subtraction here will probably not be needed, hence the search space shrinks by excluding all those sequences containing subtraction operations.

Memory reduction strategy: A more implicit type of strategy is based on reduction through arithmetic memory. This is the idea that the outcome of particular basic arithmetic operations is stored in a person's memory, and hence do not require an actual operation to obtain the outcome, prime examples of these are $1 + 1 = 2$, or $2 \times 2 = 4$. If we think of the cognitive load associated with performing these types of games, we argue that the load required by the types of operations can be ordered as memory operation < heuristic operation < full operation. In a memory operation, the outcome of the specific operation was stored in our memory, and hence, only needed retrieving. In a heuristic operation, not the outcome itself, but a general rule that makes finding the outcome more straightforward was stored, for example, $10 \times x$ or, $x \div 1$.

We suspect that as the cognitive load for these operations is thus minimal, a person can relatively easily swap two numbers in x for any outcome of their operation encoded in their memory which reduces the size of x and hence the search space. Some support for this can be found in the analysis of the 24 Game data by Van der Maas and Nyamsuren [34], who showed that items which could be reduced to 24 Game specific heuristics, namely 8×3 or 6×4, were easier solved.

Algebraic reduction strategy: The final type of reduction strategy is based on the recognition of algebraic equivalencies. For example, in the case of $n = 3$ having considered a response of the form $((x_a \div x_b) \div x_c)$ and concluded that it does not lead to the target value, one with knowledge of algebraic equivalencies would recognise that in this case it no longer makes sense to consider any of the following three response sequences, $((x_a \div x_c) \div x_b), (x_a \div (x_b \times x_c))$, or $(x_a \div (x_c \times x_b))$, as these are all equivalent.

Whereas the success of a random strategy is dependent on the size of the search space, for the reduction strategy, it is dependent on the extent to which reduction opportunities are available, and recognised by the player. Having discussed random and reduction strategies, we arrive at our main point of interest for this paper, the creative strategy.

2.1.3. Creative

At the beginning of this section, we wrote that we agree with the premise of Van der Maas and Nyamsuren [34] that *Numbers* game items require a creative solution if the search space is complex. However, as they focused on NP-completeness, their argument primarily covered the implied failure of random (brute-force) strategies, which is only a function of the size of the search space. Here we provide the second part of the argument and propose that a search space for a *Numbers* game item is complex when the search space for this item is too large for successful application of a random

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

strategy and does not provide sufficient recognisable opportunities for a reduction strategy, and hence, shrinking the search space to a manageable size.

3. Creative mathematical thinking in the *Numbers* game

Even though we established a theoretical definition of what constitutes a complex search space, we must acknowledge that this definition is not very practical. Specifically, because the definition is founded on the absence of sufficient recognisable opportunities for a reduction strategy, this would require us to consider for each individual item which reduction opportunities are available, and whether these are enough to use a reduction strategy. In the context of large scale online assessment, this is of course not desirable. However, what we can do is look for a general characteristic of *Numbers* game items for which variation in the characteristic might say something about the opportunities for reduction and, as such, the complexity of the search space.

One possible item characteristic is found by considering the close relationship between the *Numbers* game and the so-called Mate-in-*m* problems in chess. In this type of chess problems, the player is given a particular setup of a chessboard and instructed to find, for the starting side, the moves such that the other side is under checkmate after exactly m moves. A *Numbers* game item in which x has size n requires $n-1$ operations, in that fashion a Mate-in-m chess problem is similar to a $n = m+1$ *Numbers* game item. An example of such a problem, in which white must checkmate black in three moves, and the solution, is shown in Figure 3.

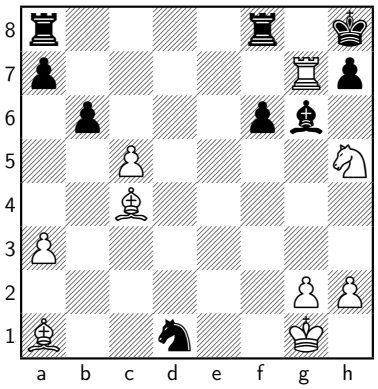


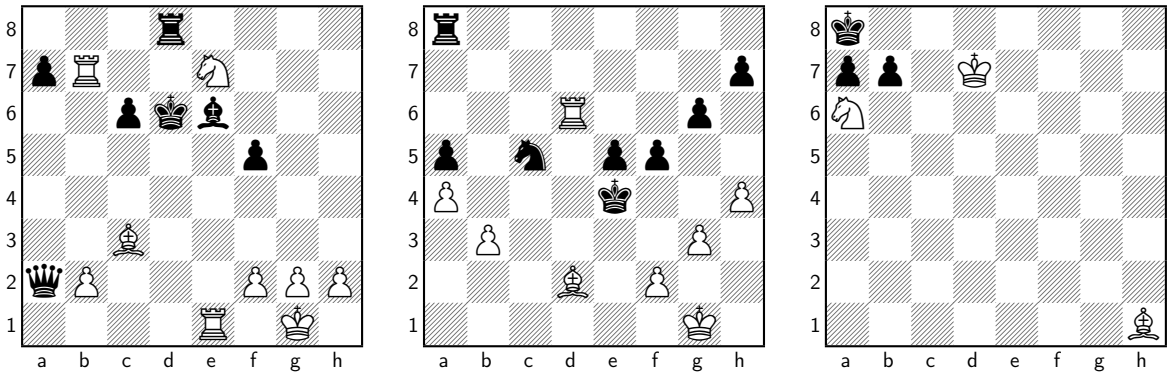
Figure 3. Mate in Three, White to move
Solution: ♖g8+, ♜xg8, ♗xf6+, ♜g7, ♗xg7#

One should read the notation in the solution in the following way. First, the white Rook (♖) moves to square **g8** after which the black king is under check as indicated by the + sign. Then a black Rook (♜) captures a piece, as indicated by the x sign, on the **g8** square, as only one black Rook is capable of this move, i.e., the one on **f8**, we do not specify the square from which this Rook came. This will be the same for all other situations in which multiple pieces of the same class our on the board, but only one piece can make a legal move. Then a white Bishop (♗) captures (x) a piece on the **f6** square, and puts the black king under check (+) once again, and a black Rook (♜) then moves to **g7**. A white Bishop (♗) captures (x) a piece on the **g7** square, resulting in a checkmate (#).

It is clear that chess games like these also have a large search space, as each operation consists of considering each possible move one can make as well as the subsequent counter move by the other colour. However, just as with the *Numbers* game, here we also find particular heuristics that can be applied in a reduction strategy to reduce the search space. For example, a negative reduction could be that as none of the white pawns (♙) could reach the black King (♚) in three moves, these probably should not be moved. A positive reduction strategy would, in this case, be the assumption that one has to use the stronger pieces, e.g., the Rook (♖), Knight (♘), or Bishop (♗). However, a reduction strategy that is most likely to be used would be to consider for moves that put black in check as soon as possible; we call this the early-check strategy. These moves would, in terms of the *Numbers* game

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

strategy, have a high convergence, and also results in making a correct prediction for the counter move more likely, as this move would have to take the black King out of check again, thus severely limiting the possible moves for black. In Figure 4 three puzzles are shown for which this reduction would work from respectively the beginning (**a**), after the first move (**b**), or only on the final move (**c**).



(a) Solution:
♗e5+, ♚c5, ♖c1+, ♝c4, ♙b4#

(b) Solution:
♔g2, ♟f4, ♙f3+, ♚f5, ♙g4#

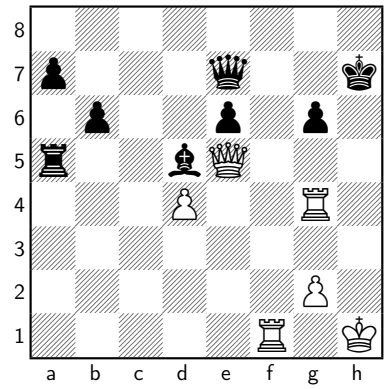
(c) Solution:
♗c6, ♟xc6, ♔c8, ♟c5, ♘c7#

Figure 4. Mate in Three, White to move

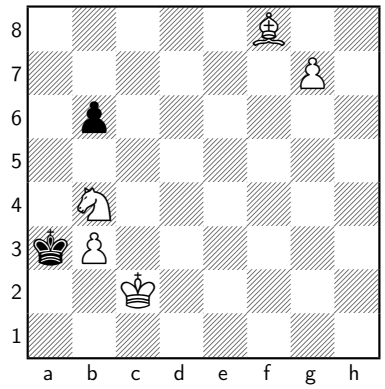
Puzzle **a** has multiple pieces that can put black on check with the first move. ♗e5+ is, however, the clear choice as it is the only move that seriously restricts the possible counter moves for black, i.e., only ♚c5, the same goes for the second move ♖c1, after which checkmate is achieved with the move ♙b4. As such, this puzzle seems pretty straightforward once the available reduction strategies are recognised and applied. For puzzle **b** one will recognise that the first move is a surprising one, and might as such require creative thinking, however after this move, and the subsequent counter move by black, the remainder of the puzzle once again becomes straightforward using reduction strategies. Finally, even though the search space for puzzle **c** is the smallest of the three, and as none of the reduction strategies appear applicable, it is the most difficult and would arguably require the most creative thinking to solve.

Naturally, chess puzzles exist that for which the early-check strategy is possible, however, that still might require creative thinking as other heuristics do not apply. For example in Figure 5a, while the correct response requires the player to put black under check in each move, the sacrifice of the Rook required in the first move (♖f7+, ♛xf7) might not be initially the most attractive one. It is after one recognises that as this forces black to move the Queen, one can move the remaining Rook (♖h4+), leaving the black king with only one option, after which merely moving either the Queen (♕h8#) or the Rook (♖h8#) result in checkmate. So apart from the unavailability of the early-check strategy, we find that puzzles, in which one must progress against suggestions based on common heuristics, might also be indicative of the requirement for creative thinking. Moreover, as we see in Figure 5b combining these two creates even more complex chess puzzles. While a pawn promotion in the first move is a logical step for white, most times one would promote to a Queen as it is the most powerful piece on the board, and not the underpromotion to a Knight, as required in the solution for this puzzle.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1



(a) Solution:
♖f7+, ♕xf7, ♖h4+, ♔g8, ♕h8/♖h8#



(b) Solution:
♙g8→♘, ♟b5, ♘e7, ♔xb4, ♘c6#

Figure 5. Mate in Three, White to move

Based on these examples of chess puzzles we can argue that the extent to which one has to think ahead about the consequences of a move before the move can be validated, might be indicative of creative thinking requirements for the puzzle. When we look to the *Numbers* game, we find that the general item characteristic that most closely resembles this is what we call the base-pattern.

3.1. Base-patterns

A base-pattern is the most fundamental representation of a response sequence in the *Numbers* game as they describe only the order in which the operations take place, and ignore the operations or numbers used in the response sequence itself. For $n = 2$ there is only one operation, and as such only a single base-pattern exists, $(x_a \sim x_b)$. Whereas there only is a single base-pattern for $n = 2$, for $n = 3$ we see that there exist two base-patterns namely, $((x_a \sim x_b) \sim x_c)$ and $(x_c \sim (x_a \sim x_b))$, in which $(x_a \sim x_b)$ denotes an operation on two numbers from the original set, for which the outcome is then used in the operation together with the remaining initial number x_c. As the set of possible base-patterns is a function of the number of operations, which is itself a function of n, namely $n - 1$, we can use P_b^n to denote the set of unique possible base-patterns for an item with n initial numbers, and its cardinality $|P_b^n|$ is given by the Catalan number C_{n-1}:

$$|P_b^n| = C_{n-1} = \frac{1}{n}\binom{2(n-1)}{(n-1)} \text{ for } n \geq 0 \tag{4}$$

Calculating P_b^n for $n = [2, 3, 4]$ we find that there are respectively 1, 2 and 5 base-patterns, which we have written out below. Continuing until $n = 10$ we would need to write out 14, 42, 132, 429, 1430 and finally 4862 different base-patterns respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

$$P_b^2 = \{(x_a \sim x_b)$$

$$P_b^3 = \begin{cases} ((x_a \sim x_b) \sim x_c) \\ (x_c \sim (x_a \sim x_b)) \end{cases}$$

$$P_b^4 = \begin{cases} ((x_a \sim x_b) \sim (x_c \sim x_d)) \\ (((x_a \sim x_b) \sim x_c) \sim x_d) \\ ((x_c \sim (x_a \sim x_b)) \sim x_d) \\ (x_d \sim ((x_a \sim x_b) \sim x_c)) \\ (x_d \sim (x_c \sim (x_a \sim x_b))) \end{cases} \quad (5)$$

Closer inspection of the base-patterns shows that we presented these patterns in a particular order for each level of n. Specifically, we ordered them to the extent the number subscripts appear in alphabetical order. For example, when we look to the second base-patterns for $n = 4$, the subscripts are in perfect alphabetical order, whereas the reverse is true for the fifth base-pattern. In practice, this would mean that for the last pattern, one must solve the item in reverse order. If we compare the *Numbers* game for n with the chess puzzle Mate-in-$n-1$, we can make an analogy between each base-pattern and the moves in which black in under check, i.e., the extent to which one needs to think ahead and the solution requires creative mathematical thinking. For $n = 4$ and Mate-in-3 we find the following correspondence between a base-pattern, and the availability of the early check strategy:

Table 1. Correspondence between $n = 4$ base-patterns and the availability of the early check strategy in Mate-in-3 chess puzzles. x indicates strategy available, o indicates strategy unavailable, # indicates check mate.

Base-pattern	Early Check Available		
	1^{st} move	2^{nd} move	3^{rd} move
$((x_a \sim x_b) \sim (x_c \sim x_d))$	x	x	#
$(((x_a \sim x_b) \sim x_c) \sim x_d)$	x	x	#
$((x_c \sim (x_a \sim x_b)) \sim x_d)$	x	o	#
$(x_d \sim ((x_a \sim x_b) \sim x_c))$	o	x	#
$(x_d \sim (x_c \sim (x_a \sim x_b)))$	o	o	#

Note that the first base-pattern is a special one, as the first two operations, i.e., $\{(x_a \sim x_b), (x_c \sim x_d)\}$ can be performed in any order. As such, if we would consider items for which all correct sequences are of a single base-pattern, while we expect that both base patters would generally not require a creative solution, from a random and reduction strategy we would expect items that require the first base to be easier. That is, from the perspective of a random strategy, one is twice as likely in randomly selecting a correct first operation with the first base-pattern, compared to all other base-patterns, including the second. From a reduction strategy perspective, in most cases, this translates to more opportunities for reduction. However, as we show in the next section when more than one base-pattern is possible for a correct sequence, people will prefer the more sequential second base-pattern to respond.

4. Results

Based on the relation to the Mate-in-m chess puzzles we would conclude that the more reversed the order of the subscripts of a base-pattern is, the more one has to think ahead in order to obtain the correct response, and hence the more likely the item is to require creative thinking. While several other

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

characteristics of items might also be indicative of creative thinking requirements, the applicability of greediness, convergence, and other heuristics are more item-specific. As we are interested in a more general indicator of creative thinking, we will focus on the base-patterns underlying the correct response. Specifically, we investigated the following two hypotheses:

Hypothesis 1. *The difficulty of items increases in the amount of operations of reverse order in the base-pattern.*

Hypothesis 2. *When the correct response to an item can be reached through multiple base-patterns, the frequency of the base-patterns used in the actual responses is decreasing in the amount of operations of reverse order in the base-pattern.*

4.1. 24 Game data

To asses Hypothesis 1, we analysed the percentage correct and mean response time data from 77 of the 1362 different 24 Game items, gathered by the website *https://www.4nums.com*. The 1362 items represent all possible combinations of four integers between one and 13, that for $O = \{+, -, \times, \div\}$ contain a solution for $T = 24$. All 1362 items together had been solved almost 3.5 million times at the time of the analyses. As the data is on the item level and already aggregated over players, we do not know the exact responses players have given to each item. As such, we use only those items in our analysis for which only one base-pattern can be used to obtain a correct response. We will refer to such items as items with a single base-pattern solution. In order to get information on the base-patterns, we calculated for each item all possible response sequences, checked which ones returned 24 as response, and returned the set base-patterns underlying those sequences. Unfortunately, it turned out that for the last three base-pattern categories there are none, or only a few, 24 Game items with a single base-pattern solution:

Table 2. Frequency of base-patterns for 24 Game items with a single base-pattern solution

Base-pattern	N
$((x_a \sim x_b) \sim (x_c \sim x_d))$	63
$(((x_a \sim x_b) \sim x_c) \sim x_d)$	8
$((x_c \sim (x_a \sim x_b)) \sim x_d)$	0
$(x_d \sim ((x_a \sim x_b) \sim x_c))$	1
$(x_d \sim (x_c \sim (x_a \sim x_b)))$	4

We decided to combine the last two base-patterns into one category, and added one more item which required either one of these two patterns to be solved. In Table 3 we show for the three different categories the number of items, the mean log response time until the item solved. In Figure 6, we show both the distribution of the proportion correct and mean log response times for each item across the three categories.

Table 3. Mean log-RT's and proportion correct for the 24 Game items with a single base-pattern solution.

			Mean (SD)	
Category	**Base-pattern**	**N**	**mean log-RT**	**prop. cor.**
A	$((x_a \sim x_b) \sim (x_c \sim x_d))$	63	2.49 (0.483)	0.785 (0.135)
B	$(((x_a \sim x_b) \sim x_c) \sim x_d)$	8	4.12 (0.341)	0.375 (0.084)
C	$(x_d \sim ((x_a \sim x_b) \sim x_c))$ $(x_d \sim (x_c \sim (x_a \sim x_b)))$	6	5.00 (0.335)	0.266 (0.038)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

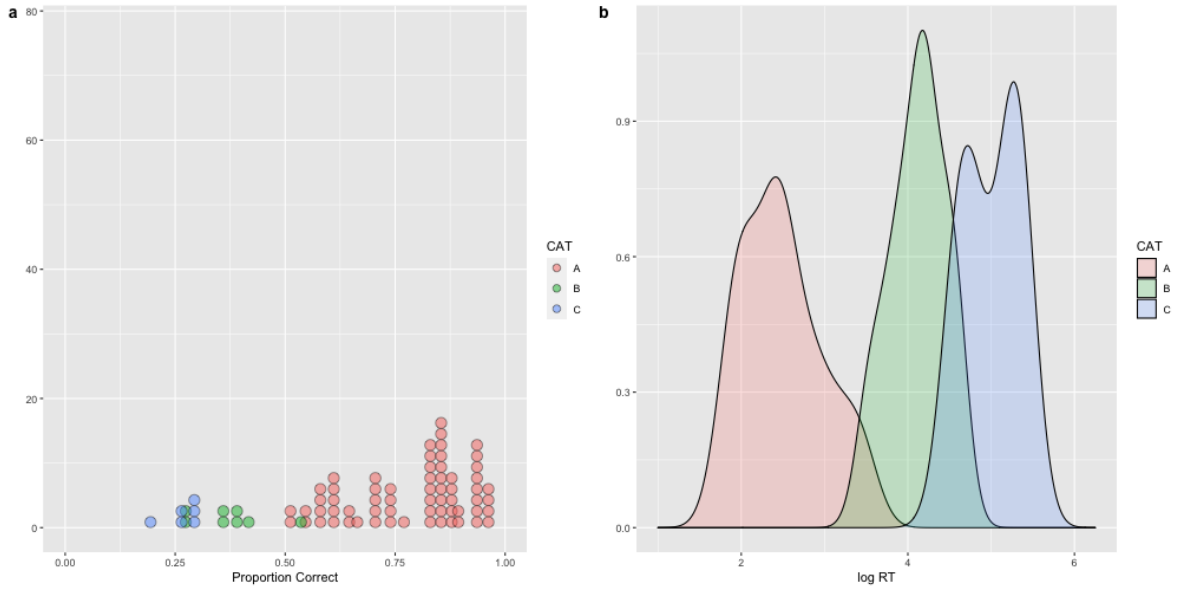


Figure 6. Proportion Correct and Log Response Times for items in the three base-pattern categories

Figure 6 shows that the base-pattern needed for the correct response seems to play a role in the probability for a correct response and how long that response will take. This is confirmed by the results of a linear regression analysis, which showed that 75% of the variance in mean log response times is explained by the base-pattern needed to give the correct response.

Table 4. Fit statistics for the estimated Regression model: log mean RT ~ base-pattern category.

	Dependent variable: log mean response time			
	Coefficient	Std. Error	t-value	p-value
CAT **A** (Intercept)	2.495 [2.38, 2.61]	0.058	42.796	< 1e-10
CAT **B**	1.624 [1.28, 1.97]	0.174	9.349	< 1e-10
CAT **C**	2.505 [2.11, 2.90]	0.198	12.674	< 1e-10
Observations	77			
R^2	0.754			
Adjusted R^2	0.747			
Residual Std. Error	0.463 (df = 74)			
F Statistic	113.394^{***} (df = 2; 74)			

These results of these analyses support the idea the difficulty of items increases in the amount of operations of reverse order in the base-pattern.

4.2. Math Garden data

In comparison to the 24 Game data, there are even fewer items with a single base-pattern solution in the *Numbers* game in the Math Garden. Specifically, of the 273 items with $n = 4$ and $O = \{+, -, \times, \div\}$, only 10 are items with a single base-pattern solution. Of those 10, six have been played only three of four times each, and none of those times a correct response was given. Luckily, in contrast to the data from the previous analysis, we have access to the actual response patterns themselves.

To asses Hypothesis 2, we use the data from 124 items with $n = 4$ and $O = \{+, -, \times, \div\}$, in which the correct response could be given using any of the five base-patterns and for which at least 100 times

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

a correct response was given. For each of these items, we derived the different response sequences that would result in a correct response and counted for each base-pattern the proportion of times it was used in such a correct response. Averaging over all items, we find that if base-patterns for correct responses are selected proportionally to their availability, we would expect to observe each base-pattern approximately the same number of times. Our analyses of the base-pattern used in 220493 responses to these 124 items revealed that the proportion of times the base-patterns are used is far from this expected frequency.

Table 5. Frequency of base-patterns used in correct responses to $n = 4$ and $O = \{+, -, \times, \div\}$ *Numbers* game items with all base-pattern solutions.

Base-pattern	Expected	Observed
$((x_a \sim x_b) \sim (x_c \sim x_d))$	.206	.125
$(((x_a \sim x_b) \sim x_c) \sim x_d)$	.201	.782
$((x_c \sim (x_a \sim x_b)) \sim x_d)$	.198	.031
$(x_d \sim ((x_a \sim x_b) \sim x_c))$	.199	.039
$(x_d \sim (x_c \sim (x_a \sim x_b)))$	.197	.022

The proportions in Table 5 show that, in line with our expectations, the frequency of the base-patterns used in the actual responses is decreasing, in the amount of operations of reverse order in the base-pattern. Interestingly, we find that it is not the $((x_a \sim x_b) \sim (x_c \sim x_d))$ base-pattern that is used most to respond, but instead $(((x_a \sim x_b) \sim x_c) \sim x_d)$. Although at first sight, this finding might seem to be contradictory to our findings with the 24 Game data, more careful consideration of the difference between the two types of items provides a plausible explanation. Specifically, for the 24 Game data, we analysed only items with a single base-pattern solution, in these cases items with base-pattern $((x_a \sim x_b) \sim (x_c \sim x_d))$ have a smaller search space, and the correct response is hence more likely to be found. However, if all base-patterns can be used, it might be simpler to perform the operations using the outcome of the previous step immediately in the next step, as this outcome is already active in the working memory. For example, if $x = \{1, 2, 3, 4\}$ and $T = 10$, we are probably more likely to summarise the correct response as $1 + 2 + 3 + 4 = 10$, and not as $1 + 2 = 3$, $3 + 4 = 7$, $3 + 7 = 10$. In other words, the second base-pattern might be preferred over the first pattern as it allows in many cases a more natural flow of the response sequence.

To make sure this finding is not an artefact of those items in which only addition has to be used, we recalculated the proportions for 93698 responses to the 84 items that could not be answered with addition operations alone and are shown in Table 6.

Table 6. Frequency of base-patterns used in correct responses to $n = 4$ and $O = \{+, -, \times, \div\}$ *Numbers* game items with all base-pattern solutions, solutions requiring only addition excluded.

Base-pattern	Expected	Observed
$((x_a \sim x_b) \sim (x_c \sim x_d))$	.209	.170
$(((x_a \sim x_b) \sim x_c) \sim x_d)$	.201	.730
$((x_c \sim (x_a \sim x_b)) \sim x_d)$	.196	.041
$(x_d \sim ((x_a \sim x_b) \sim x_c))$	.198	.035
$(x_d \sim (x_c \sim (x_a \sim x_b)))$	.195	.025

Even though we see that the proportion of times the first base-pattern is used in these items goes up slightly at the cost of the second pattern, we still see an overwhelming preference for the second base-pattern. As in the previous selection of items, the three non-ordered base-patterns are each used decreasingly in this disorderedness. We further confirmed our findings by analysing the 2241 responses to 5 items that can be answered correctly with all but the first base-pattern (IG1), and the 3502 responses to the five items that can be answered correctly with only one of the first two base-patterns

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

(IG2). The observed proportions for both subsets of items in Table 7 show that base-pattern $(((x_a \sim x_b) \sim x_c) \sim x_d)$ is strongly preferred over $((x_a \sim x_b) \sim (x_c \sim x_d))$ if only these two are possible, and if only $((x_a \sim x_b) \sim (x_c \sim x_d))$ cannot be used for a correct response, we find that the proportion of used base-patterns is decreasing in the disorderedness of the base-pattern.

Table 7. Frequency of base-patterns used in correct responses to $n = 4$ and $O = \{+, -, \times, \div\}$ *Numbers* game items in which only the first two base-patterns can be used in the solution (IG1), and items in which all but the first base-pattern can be used in the solution (IG2).

	IG1		IG2	
Base-pattern	**Expected**	**Observed**	**Expected**	**Observed**
$((x_a \sim x_b) \sim (x_c \sim x_d))$	.500	.067		
$(((x_a \sim x_b) \sim x_c) \sim x_d)$	.500	.933	.245	.755
$((x_c \sim (x_a \sim x_b)) \sim x_d)$			.245	.166
$(x_d \sim ((x_a \sim x_b) \sim x_c))$			.255	.023
$(x_d \sim (x_c \sim (x_a \sim x_b)))$			.255	.056

5. Discussion

Creativity is an important 21st century skill that educators want to instil upon their pupils[1,2]. Given that online education is on the rise, the need arises for including creative thinking in online learning and assessment platforms. Therefore, creative thinking problems need to be developed that can be administered at scale. In this paper we argue that Math Garden's *Numbers* game could be used to assess creative mathematical thinking at scale. In the *Numbers* game children are asked to figure out how to compute a target number using basic arithmetic operations and a given set of numbers. We argue that creative thinking is required in complex search spaces and for the *Numbers* game these can be characterised by disordered base-patterns, i.e., those items in which the sequence of the operations needed to solve a problem is not straightforward. We provide evidence based on data from the 24 Game and from Math Garden's *Numbers* game that disordered base-patterns determine search space complexity and hence the need for creative thinking. We conclude that the *Numbers* game can be a valuable tool in assessing creative mathematical thinking in large-scale online learning and assessment platforms.

We argued that creative thinking is required when the search space is complex. We then proposed that the base-pattern underlying the response sequence to a *Numbers* game item determines the complexity of the search space, and hence whether creative mathematical thinking is required. In Equation 5 we show the base-patterns, arranged in terms of their disorderedness, for situations with 2, 3 and 4 numbers in the initial set. We show for items from the 24 Game with only one solution sequence, that those with disordered base-patterns are more difficult and take longer to solve compared to items with ordered base-patterns. For the *Numbers* game in Math Garden we found that if multiple solution sequences were possible, then nine times out of ten, children chose ordered over disordered base-patterns. We concluded that the base-pattern can be considered a general indicator for search space complexity.

The base-pattern is arguably a crude indicator of creative thinking, and may not be able to perfectly distinguish creative from non-creative items. We can fine-tune the classification of item complexity by using more distinct response patterns. For example, instead of treating all operators as interchangeable ($\sim$) we can distinguishes operators signifying order-independent operations $\sim \in \{+, \times\}$, from operators signifying order dependent operations $\sim \in \{-, \div\}$. For an item with three numbers in the initial set, this distinction would allow us to compare the complexity of six patterns, instead of two base-patterns. In the appendix, we show how response patterns can be distinguished even further as we move from base-patterns to the full response pattern in several steps.

Through the analysis of 24 Game and the *Numbers* game data, we were able to establish the base-pattern as an indicator for mathematical creativity requirements. However, more data would be

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

needed to confirms this. For example, even though we could establish differences in response times as a function of base-pattern complexity, for items with a single base-pattern solution using the 24 Game data, the overall number of complex base-patterns in the game is not very high. This is because there are fewer items with a single base-pattern solution as a consequence of restricting $x \in [1, 13]$ and $T = 24$. In Table 8 we show that the number of items with single base-pattern solutions goes up when allowing T to be either a positive integer value ($T \in \mathbb{N}$), or any integer value ($T \in \mathbb{Z}$), compared to when $T = 24$.

Table 8. Frequency of base-patterns for $x \in [1, 13]$ items with a single base-pattern solution and different ranges for T

Base-pattern	$T = 24$	$T \in \mathbb{N}$	$T \in \mathbb{Z}$
$((x_a \sim x_b) \sim (x_c \sim x_d))$	63	10872	17152
$(((x_a \sim x_b) \sim x_c) \sim x_d)$	8	391	722
$((x_c \sim (x_a \sim x_b)) \sim x_d)$	0	57	1096
$(x_d \sim ((x_a \sim x_b) \sim x_c))$	1	167	325
$(x_d \sim (x_c \sim (x_a \sim x_b)))$	4	103	168

While platforms such as Math Garden might be a great way to gather such data, for applications that cater to children in early grades of primary school, arithmetic ability might play a role for younger children that have not acquired all of the necessary skills. In the current analysis this was not a problem as we focused on the distribution of base-patterns for correct responses. However, due to the adaptive nature of the system, complex items are more likely to be solved incorrectly, especially given the limited time children have to solve each item. As such, estimated item difficulty will be high, and hence, these items are less often presented to the players. We therefore need to take into account that adaptive settings might not be the optimal way to obtain data for our research question.

Another solution might be to develop a joint measurement model for creative thinking and arithmetic ability. In such a model the probability correct for a *Numbers* game item would depend on item difficulty, which is a function of the difficulty of the required operations, e.g., operators and numbers, and item complexity. In this way, we can examine whether the player solved an item incorrectly because they don't posses the required arithmetic ability or because they did not think creatively.

With creative thinking being increasingly recognised as an essential ability that should be part of school curricula, the need to facilitate learning and assessment of creative thinking skills on a large scale will also increase. Given the complex nature of creativity, developing games that can automatically measure creative thinking, without further taxing educators, is paramount. The *Numbers* game appears to be an excellent candidate for assessing mathematical creativity in online learning and assessment settings. Furthermore, the requirements we derived for creative thinking in the context of the Mate-in-m chess puzzles and the *Numbers* game could be used to develop games that assess creative thinking in other domains.

Author Contributions: conceptualisation, H.M. and J.K.; methodology, H.M. and J.K.; formal analysis, J.K.; writing–original draft preparation, J.K.; writing–review and editing, C.S., H.M, and J.K.; visualisation, J.K.; supervision, H.M.

Funding: This research was funded by the NEDERLANDSE ORGANISATIE VOOR WETENSCHAPPELIJK ONDERZOEK grant number 022.005.0 (J.K.) and 314-99-107 (H.M.), and by the JACOBS FOUNDATION FELLOWSHIP 2019-2021 (C.S.).

Conflicts of Interest: The authors declare no conflict of interest.The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A Patterns

In this section, we discuss the different types of patterns that form a response sequence in the *Numbers* game and the algebraic equivalencies between patterns. Specifically, we will show how we

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

can from $|P_b^n|$ unique base-patterns to all $|P^{nk}|$ unique response sequences, using the example of an item in which $n = 2$, and $O = \{+, -, \times, \div\}$.

Appendix A.1 Base pattern

As we already discussed the base-pattern in the main text, we only reiterate that this is the most fundamental representation of a response sequence as it describes only the order in which the operations take place, and ignores the operations or numbers used in the response sequence itself. We reiterate that as the set of possible base-patterns is a function of the number of operations, which is itself a function of n, namely $n-1$, we use P_b^n to denote the set of unique possible base-patterns for an item with n initial numbers, and its cardinality $|P_b^n|$ is given by the Catalan number C_{n-1}.

Appendix A.2 Operator pattern

A more detailed representation of a response sequence is one that takes the used operators into account. We can distinguish two types of operator patterns, the basic operator pattern, and the full operator pattern. The basic operator pattern only makes a distinction between operators for which the operation outcome is independent of the number order, e.g., $(x_a + x_b) = (x_b + x_a)$, and those operators for which the outcome is dependent on the number order, $(x_a - x_b) \neq (x_b - x_a)$. The full operator pattern makes the distinction between all different operators.

Basic operator pattern

For the basic operator pattern, we distinguish addition and multiplication operations, from subtraction and division operation. To that end, we use $* \in \{+, \times\}$ to denote operations for which the order of the numbers, in an operation, does not matter. We use $\div \in \{-, \div\}$ to denote operations for which the order of the numbers, in an operation, does matter. It is with this distinction that we also see equivalent patterns for the first time, as for $n = 3$ there are eight basic operator patterns (P_{bo}^n). However, only 6 of those will lead to a unique response:

$$P_{bo}^3 = \begin{cases} [((x_a * x_b) * x_c), (x_c * (x_a * x_b))] \\ [((x_a \div x_b) * x_c), (x_c * (x_a \div x_b))] \\ ((x_a * x_b) \div x_c) \\ (x_c \div (x_a * x_b)) \\ ((x_a \div x_b) \div x_c) \\ (x_c \div (x_a \div x_b)) \end{cases} \tag{A1}$$

Note that this is the case under the assumption that the actual operations are the same, e.g., $((x_a + x_b) \times x_c) = (x_c \times (x_a + x_b))$, while $((x_a + x_b) \times x_c) \neq (x_c + (x_a \times x_b))$. As such, the basic operator pattern might be mostly of interest as a indicator for the success of the random strategy, i.e., response sequences containing only operations in $*$ are more likely to be sampled, as the order of the operation does not matter.

Full operator pattern

As for each of the $n-1$ operations, we would have k possible operators the number of possible full operator patterns given n, and k is given by:

$$\begin{aligned} |P_o^{nk}| &= k^{n-1} \times |P_b^n| \\ &= k^{n-1} \times \frac{1}{n} \binom{2(n-1)}{(n-1)} \end{aligned} \tag{A2}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

It is with this representation that we find the first patterns that are always equivalent, while for $n = 3$ and $O = \{+, -, \times, \div\}$ there are 32 different operator-patterns ($P_o^3\{*\}$, in which the symbol $*$ represents that $O = \{+, -, \times, \div\}$), only 24 of those will lead to a unique response:

$$P_o^3\{*\} = \begin{cases} ((x_a + x_b) + x_c) = (x_c + (x_a + x_b)) \\ ((x_a + x_b) - x_c) \neq (x_c - (x_a + x_b)) \\ ((x_a + x_b) \times x_c) = (x_c \times (x_a + x_b)) \\ ((x_a + x_b) \div x_c) \neq (x_c \div (x_a + x_b)) \\ \\ ((x_a - x_b) + x_c) = (x_c + (x_a - x_b)) \\ ((x_a - x_b) - x_c) \neq (x_c - (x_a - x_b)) \\ ((x_a - x_b) \times x_c) = (x_c \times (x_a - x_b)) \\ ((x_a - x_b) \div x_c) \neq (x_c \div (x_a - x_b)) \\ \\ ((x_a \times x_b) + x_c) = (x_c + (x_a \times x_b)) \\ ((x_a \times x_b) - x_c) \neq (x_c - (x_a \times x_b)) \\ ((x_a \times x_b) \times x_c) = (x_c \times (x_a \times x_b)) \\ ((x_a \times x_b) \div x_c) \neq (x_c \div (x_a \times x_b)) \\ \\ ((x_a \div x_b) + x_c) = (x_c + (x_a \div x_b)) \\ ((x_a \div x_b) - x_c) \neq (x_c - (x_a \div x_b)) \\ ((x_a \div x_b) \times x_c) = (x_c \times (x_a \div x_b)) \\ ((x_a \div x_b) \div x_c) \neq (x_c \div (x_a \div x_b)) \end{cases} \tag{A3}$$

Appendix A.3 Sequence pattern

The most detailed representation of a response pattern for the *Numbers* game that is still a general representation takes not only the operators but also the order of the numbers into account. Specifically, whereas with the operator patterns the operations take place in alphabetical order, e.g., the first operation is always represented by a sequence $(x_a \sim x_b)$, in as sequence pattern the first sequence can also be of the form $(x_b \sim x_a)$ or if $n > 2$ $(x_c \sim x_b)$.

As there are n number placeholders $\{x_a, x_b, \ldots, x_n\}$ in an operator pattern, for a sequence pattern, there have n possible assignments for the first placeholder, $n - 1$ possible assignments for the second placeholder, $n - 2$ possible assignments for the third placeholder, etc., until we only have one left for the last placeholder. In other words, the number of ways in which we can assign our n characters to one operator pattern is given by $n!$. Hence the number of possible sequence patterns is given by:

$$\begin{aligned} |P_s^{nk}| &= n! \times |P_o^{nk}| \\ &= n! \times k^{n-1} \times \frac{1}{n} \binom{2(n-1)}{(n-1)} \end{aligned} \tag{A4}$$

The attentive reader will notice that the sequence patterns describe all possible response sequences possible for a given *Numbers* game item with n and k. Hence, we would expect the number of possible sequence patterns ($|P_s^{nk}|$) to be equal to the total number of possible sequences of operations forming an arithmetic expression ($|P_r^{nk}|$) as given by Equation 3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

$$\begin{aligned} n! \times k^{(n-1)} \times (n-1)! &= n! \times k^{n-1} \times \frac{1}{n}\binom{2(n-1)}{(n-1)} \\ \cancel{n! \times k^{(n-1)}} \times (n-1)! &= \cancel{n! \times k^{(n-1)}} \times \frac{1}{n}\binom{2(n-1)}{(n-1)} \\ (n-1)! &= \frac{1}{n}\binom{2(n-1)}{(n-1)} \end{aligned} \tag{A5}$$

Solving for the last line of the equation we find that $|P_s^{nk}| = |P_r^{nk}|$ only if $n \in \{2,3\}$. The difference for $n > 3$ is explained when we look to the first base-pattern for $n = 4$ from Equation 5, $((x_a \sim x_b) \sim (x_c \sim x_d))$. This pattern exist of three different operations, $x_{ab} = (x_a \sim x_b)$, $x_{cd} = (x_c \sim x_d)$, and $(x_{ab} \sim x_{cd})$, however while for all other base-patterns in P_b^4 the order of these operations is fixed, i.e., you have to start with $(x_a \sim x_b)$, then a operation with x_c, and finally an operation with x_d, in the case of the first pattern you can also start with the operation $(x_c \sim x_d)$. Another way to look at this is that there technically exists a sixth base-pattern for $n = 4$, $((x_c \sim x_d) \sim (x_a \sim x_b))$, if we would count that one we would find that $|P_r^{nk}| = |P_s^{nk}|$ as $(n-1)! = 6$. However, as every possible outcome of $((x_c \sim x_d) \sim (x_a \sim x_b))$ is already captured in $((x_a \sim x_b) \sim (x_c \sim x_d))$, we will use $|P_s^{nk}|$ as our baseline as it gives us the number of syntactically different response sequences.

As in the case of operator patterns we already found that syntactically different patterns can be algebraically equivalent, it should come as no surprise that we find even more of these algebraic equivalencies in sequence patterns. We will use $P_s^n\{O\}$ to denote all sequence patterns that are algebraically unique give n and O, and $P_{s^m}^n\{O\}$ to denote the subset of sequence patterns that share the same outcome among m sequence patterns, e.g., $P_{s^1}^3\{*\}$ is the subset of patterns for $n = 3$ and $O = \{+,-,\times,\div\}$, for which no algebraically equivalent pattern exists, whereas $P_{s^{12}}^3\{*\}$ is the subset of patterns that each are algebraically equivalent to 11 other sequence patterns. Furthermore, if we let $|P_{s^m}^n\{O\}|$ denote the number of subsets of size m in which all m sequences in that subset have the same outcome, e.g., if $|P_{s^{12}}^3\{*\}| = 2$ this means that there are two subsets each containing 12 sequences that are algebraically equivalent, and $|P_{s^m}^n\{O\}| \times m$ is the total number of sequence patterns in that collection. Hence the number of algebraically unique sequence patterns is given by:

$$|P_s^n\{O\}| = \sum_{i=1}^{|P_s^{nk}|} |P_{s^i}^n\{O\}|. \tag{A6}$$

Note that as $\sum_m |P_{s^m}^n\{O\}| \times m$ must sum to $|P_s^{nk}|$, the maximum number of non-empty sets is given by $\frac{1}{2}\left\lfloor \sqrt{1+8|P_s^{nk}|} - 1 \right\rfloor$, and hence, most subsets will be empty. Which, order of equivalencies actually exist is of course dependent on n and O, for example if $O = \{+\}$ all sequence patterns are algebraically equivalent and hence up until $\sum_{i=1}^{|P_s^{nk}|-1} |P_{s^i}^n\{O\}|$, $|P_s^n\{O\}| = 0$. For example, with $n = 3$ and $O = \{+,-,\times,\div\}$ we find that only for $m \in \{1,2,4,8,12\}$, $|P_{s^m}^n\{O\}| > 0$.

Looking to the actual number of algebraically unique sequence patterns, we will find that $|P_s^{34}| = 192$, we will not write all 192 sequences and their equivalent patterns out, but instead give examples of the different types of m^{th} order equivalencies. The 1^{st} order sequence patterns contain those in which either changing the order of the numbers, or the order of the operations will lead to a different outcome, i.e., the sequence patterns that have no algebraically equivalents. For $n = 3$ these are the patterns in which the two operations consist of one subtraction and one division operation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

$$
\begin{aligned}
|P^3_{s1}\{\ast\}| &= 24 \\
1 \times |P^3_{s1}\{\ast\}| &= 24 \\
P^3_{s1}\{\ast\} &= \begin{cases} ((x_a - x_b) \div x_c) \neq (x_c \div (x_a - x_b)) \neq ((x_b - x_a) \div x_c) \\ ((x_a \div x_b) - x_c) \neq (x_c - (x_a \div x_b)) \neq ((x_b \div x_a) - x_c) \\ \dots \end{cases}
\end{aligned} \tag{A7}
$$

The 2^{nd} order sequence patterns contain those patterns in which the two operations consist of either one subtraction and one multiplication operation, or one addition and one division operation, such as:

$$
\begin{aligned}
|P^3_{s2}\{\ast\}| &= 24 \\
2 \times |P^3_{s2}\{\ast\}| &= 48 \\
P^3_{s2}\{\ast\} &= \begin{cases} ((x_a + x_b) \div x_c) = ((x_b + x_a) \div x_c) \\ ((x_a \times x_b) - x_c) = ((x_b \times x_a) - x_c) \\ ((x_a \div x_b) + x_c) = (x_c + (x_a \div x_b)) \\ ((x_a - x_b) \times x_c) = (x_c \times (x_a - x_b)) \\ \dots \end{cases}
\end{aligned} \tag{A8}
$$

The 4^{th} order sequence patterns contains 12 sets of 4 patterns with different types of sequences in which not only the combination of operations, but also the order in which these operations takes place, play a role. For example, even though both of the following patterns consist of two subtraction operations, whereas $((x_a - x_b) - x_c)$ is a 4^{th} order sequence pattern, $(x_c - (x_a - x_b))$ turns out to be an 8^{th} order sequence pattern. Generally, the 4^{th} order sequence pattern contains those sequences with one addition and one multiplication operation, or those sequence patterns in which the operations are either both subtraction or both division and the base-pattern is of the form $((x_a \sim x_b) \sim x_c)$, such that we obtain equivalent representation either by letting x_b and x_c switch places, or using the base-pattern $(x_a \sim (x_b \sim x_c)$ and changing the operation $x_b \sim x_c$ from subtraction to addition, or division to multiplication.

$$
\begin{aligned}
|P^3_{s4}\{\ast\}| &= 12 \\
4 \times |P^3_{s4}\{\ast\}| &= 48 \\
P^3_{s4}\{\ast\} &= \begin{cases} ((x_a + x_b) \times x_c) = ((x_b + x_a) \times x_c) = (x_c \times (x_a + x_b)) = (x_c \times (x_b + x_a)) \\ ((x_a - x_b) - x_c) = ((x_a - x_c) - x_b) = (x_a - (x_b + x_c)) = (x_a - (x_c + x_b)) \\ ((x_a \times x_b) + x_c) = ((x_b \times x_a) + x_c) = (x_c + (x_a \times x_b)) = (x_c + (x_b \times x_a)) \\ ((x_a \div x_b) \div x_c) = ((x_a \div x_c) \div x_b) = (x_a \div (x_b \times x_c)) = (x_a \div (x_c \times x_b)) \\ \dots \end{cases}
\end{aligned} \tag{A9}
$$

The 8^{th} order sequence patterns contain 6 sets of 8 patterns, in which all patterns in the set contain either one addition and one subtraction operation, or one multiplication and one division operation. As in the previous subset, not only the combination of operations but also the order in which these operations takes place plays a role to determine if there are respectively four or eight algebraically equivalent sequences.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 October 2020 doi:10.20944/preprints202010.0314.v1

$$
\begin{aligned}
|P^3_{s8}\{ * \}| &= 6 \\
8 \times |P^3_{s8}\{ * \}| &= 48 \\
P^3_{s8}\{ * \} &= \begin{cases} ((x_a - x_b) + x_c) = ((x_a + x_c) - x_b) = ((x_c - x_b) + x_a) = (x_a - (x_b + x_c)) = \dots \\ ((x_a \times x_b) \div x_c) = ((x_a \div x_c) \times x_b) = (x_a \times (x_b \div x_c)) = (x_a \div (x_c \div x_b)) = \dots \\ \dots \end{cases}
\end{aligned} \tag{A10}
$$

The final 12^{th} order sequence patterns contain 2 sets of 12 patterns in which all patterns in the set contain either two addition or two multiplication operations.

$$
\begin{aligned}
|P^3_{s12}\{ * \}| &= 2 \\
12 \times |P^3_{s12}\{ * \}| &= 24 \\
P^3_{s12}\{ * \} &= \begin{cases} ((x_a + x_b) + x_c) = ((x_a + x_c) + x_b) = ((x_c + x_b) + x_a) = (x_a + (x_b + x_c)) = \dots \\ ((x_a \times x_b) \times x_c) = ((x_a \times x_c) \times x_b) = ((x_c \times x_b) \times x_a) = (x_a \times (x_b \times x_c)) = \dots \end{cases}
\end{aligned} \tag{A11}
$$

If we would calculate $\sum_{i=[1,2,4,8,12]} i \times |P^n_{si}\{O\}|$ for $n = 3$ and $O = \{+, -, \times, \div\}$ we find that $|P^3_s\{ * \}| = 68$, i.e., while there are 192 syntactically unique sequences, there are only 68 algebraically unique sequences, and hence potentially unique responses.

The reason we talk about potentially unique responses is that even though the 68 sequences themselves are algebraically distinct, this only means that there exist an $x = [x_1, x_2, x_3]$ that contains 3 numbers such that there are 68 different outcomes possible from all 192 possible operational sequences, but this is not necessarily the case for all combinations of 3 values in x. Mainly when x contains duplicate numbers, ones, and/or zeros, the number of algebraically unique patterns is reduced. With duplicates, two values can be interchanged without changing the outcome, dividing and multiplying by one leads to the same outcomes, and adding and subtracting zero from any number as well. As such, $|P^n_s\{O\}|$ is an upper bound on the number of unique outcomes, which can be generated from a particular item with x and O. As there is no easy analytical way to derive the exact number of unique outcomes for a particular item, one has to use statistical software such as R[44] to calculate all possible outcomes for a particular problem.

References

1. Lucas, B.; Spencer, E. *Teaching Creative Thinking: Developing learners who generate ideas and can think critically*; Pedagogy for a Changing World, Crown House Publishing: Camarthen, UK, 2017.
2. OECD. Framework for the assessment of creative thinking in PISA 2021. Draft report 3, Organisation for Economic Co-operation and Development, 2019.
3. OECD. (Organisation for Economic Co-operation and Development). PISA Homepage. http://www.oecd.org/pisa, accessed on 01-09-2020.
4. Guilford, J.P. *The Nature of Human Intelligence*; McGraw-Hill: New York, 1967.
5. Torrance, E.P. *The Torrance Tests of Creative Thinking (Norma-technical Manual)*. Scholastic Testing Service, Inc., Bensenville, 1974.
6. Wallach, M.A.; Kogan, N. *Modes of Thinking in Young Children: A Study of the Creativity-intelligence Distinction*; Holt, Rinehart and Winston: New York, 1965.
7. Mednick, M.T.; Andrews, F.M. Creative thinking and level of intelligence. *The Journal of Creative Behavior* **1967**, *1*, 428–431.
8. Olteţeanu, A.; Zunjani, F.H. A visual remote associates test and its validation. *Frontiers in Psychology* **2020**, *11:26*, 1–9.
9. Baas, M.; Van der Maas, H.L.J. De (on) mogelijkheid van een valide meting van creatief potentieel voor selectiedoeleinden. *Gedrag en Organisatie* **2015**, *28*, 78–97.
10. Said-Metwaly, S.; van den Noortgate, W.; Kyndt, E. Approaches to measuring creativity: a systematic literature review. *Creativity. Theories – Research – Applications* **2017**, *4*, 238–275.
11. Runco, M.A.; Pritzker, S.R. *Encyclopedia of Creativity*; Academic Press, 2011.
12. Lee, C.S.; Huggins, A.C.; Therriault, D.J. A measure of creativity or intelligence? Examining internal and external structure validity evidence of the Remote Associates Test. *Psychology of Aesthetics, Creativity, and the Arts* **2014**, *8*, 446 – 460.
13. Lubart, T.I.; Besançon, M.; Barbot, B. *EPOC: évaluation du potentiel créatif*; Hogrefe: Paris, France, 2011.
14. Barbot, B.; Besançon, M.; Lubart, T.I. Assessing creativity in the classroom. *The Open Education Journal* **2011**, *4*, 58 – 66.
15. Barbot, B.; Besançon, M.; Lubart, T.I. The generality-specificity of creativity: Exploring the structure of creative potential with EPoC. *Learning and Individual Differences* **2016**, *52*, 178–187.
16. Amabile, T.M. Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology* **1982**, *43*, 997 – 1013.
17. Baer, J.; McKool, S.S. Assessing creativity using the consensual assessment technique. In *Handbook of research on assessment technologies, methods, and applications in higher education*; Schreiner, C., Ed.; Information Science Reference: New York, 2009; pp. 65 – 77.
18. Cheung, P.C.; Lau, S. Gender differences in the creativity of Hong Kong school children: Comparison by using the new electronic Wallach–Kogan creativity tests. *Creativity Research Journal* **2010**, *22*, 194–199.
19. Lau, S.; Cheung, P.C. Creativity assessment: Comparability of the electronic and paper-and-pencil versions of the Wallach–Kogan Creativity Tests. *Thinking Skills and Creativity* **2010**, *5*, 101–107.
20. Palaniappan, A.K. Web-based creativity assessment system. *International Journal of Information and Education Technology* **2012**, *2*, 255–258.
21. Pásztor, A.; Molnár, G.; Csapó, B. Technology-based assessment of creativity in educational context: the case of divergent thinking and its relation to mathematical achievement. *Thinking Skills and Creativity* **2015**, *18*, 32–42.
22. Forster, E.A.; Dunbar, K.N. Creativity evaluation through latent semantic analysis. Proceedings of the Annual Meeting of the Cognitive Science Society, 2009, pp. 602–607.
23. Harbinson, J.I.; Haarman, H. Automated scoring of originality using semantic representations. Proceedings of the Annual Meeting of the Cognitive Science Society, 2014, pp. 2327–2332.
24. Beketayev, K.; Runco, M.A. Scoring divergent thinking tests by computer with a semantics-based algorithm. *Europe's Journal of Psychology* **2016**, *12*, 210–220.
25. Stevenson, C.; Smal, I.; Baas, M.; Dahrendorf, M.; Grasman, R.P.P.P.; Tanis, C.; Scheurs, E.; Sleiffer, D.; Van der Maas, H.L.J. Automated AUT scoring using a Big Data variant of the Consensual Assessment Technique. Technical report, University of Amsterdam, 2020.

26. Forthmann, B.; Oyebade, O.; Ojo, A.; Günther, F.; Holling, H. Application of latent semantic analysis to divergent thinking is biased by elaboration. *The Journal of Creative Behavior* **2018**, *53*, 559 – 575.
27. Suntex International. 24® Game History. https://www.24game.com/t-about.aspx, accessed on 2020-06-05.
28. Irvine, E.B. Review of Krypto. *The Mathematics Teacher* **1974**, *67*, 438–438.
29. Greenes, C.E. A Review of Mathematical Games. *Simulation & Games* **1975**, *6*, 408–422.
30. Flaherty, J.; Connolly, B.; Lee-Bayha, J. Evaluation of the First In Math® Online Mathematics Program. Evaluation report, National School District, San Diego, CA, 2005.
31. Suntex International. First in Math®. https://www.firstinmath.com, accessed on 2020-10-09.
32. Eley, J. How Much Does the 24 Game Increase the Recall of Arithmetic Facts? Unpublished manuscript.
33. Kurzen, L. Some Ideas for the Cijferstaak. Internal report, University of Amsterdam, 2011.
34. Van der Maas, H.L.J.; Nyamsuren, E. Cognitive analysis of educational games: The number game. *Topics in cognitive science* **2017**, *9*, 395–412.
35. Klinkenberg, S.; Straatemeier, M.; Van der Maas, H.L.J. Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education* **2011**, *57*, 1813–1824.
36. Maris, G.; Van der Maas, H. Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika* **2012**, *77*, 615–633.
37. Simpson, R.M. Creative Imagination. *The American Journal of Psychology* **1922**, *33*, 234–243.
38. Bartlett, F.C. *Thinking: An Experimental and Social Study*; Allen & Unwin: London, UK, 1958.
39. Torrance, E.P. *Guiding Creative Talent*; Prentice-Hall: Englewood Cliffs, 1962.
40. Boden, M.A. *The Creative Mind: Myths and Mechanisms*; Routledge: London, UK, 2004.
41. Hayes, B. Computing science: The easiest hard problem. *American Scientist* **2002**, *90*, 113–117.
42. Mertens, S. A physicist's approach to number partitioning. *Theoretical Computer Science* **2001**, *265*, 79–108.
43. Mertens, S. The Easiest Hard Problem: Number Partitioning. *Computational Complexity and Statistical Physics* **2006**, *125*, 125–139.
44. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.