Universiteit Leiden

Psychologie
Faculteit der Sociale Wetenschappen

Using heuristics in solving arithmetic word problems: An advantage or disadvantage?

Name: Myrthe Torenbeek
Student number: s1055933
Supervisor: B.R. Bocanegra
Second reader: W.L.G. Verschuur
Cognitive Psychology
Thesis Msci Applied Cognitive Psychology

Abstract

In the present study we investigated which training method (structured or non-structured) to solve arithmetic word problems prepares children best for solving them the real world. These arithmetic word problems could be simple and difficult. We hypothesized that a structured training condition would stimulate respondents to use heuristics whereas a non-structured training condition would stimulate respondents to use algorithms. Using heuristics would be beneficial in solving simple arithmetic word problems but disadvantageous in solving difficult arithmetic word problems. On the other hand, using algorithms would be beneficial in solving difficult arithmetic word problems but it would be an unwieldy procedure for simple arithmetic word problems. We expected that respondents would have more difficulties with difficult than simple arithmetic word problems. Furthermore, we expected this difference to be larger in the structured training condition than in the non-structured training condition. We found out that, contradictory to our expectations, there was no difference between difficult and simple arithmetic word problems for the respondents trained in the structured training condition. A possible explanation is that these respondents developed a metaheuristic over the already existing heuristic, giving them the opportunity to solve simple and difficult problems in the same manner.

Introduction

An important part of school is mathematics education. In the Netherlands, 20 percent of the primary school uses the mathematics curriculum 'de wereld in getallen' ('the world in numbers'), making it the second most used mathematics curriculum in the Netherlands (Janssen, Van Der Schoot, & Hemker, 2005). In the fourth grade the goal of this curriculum is to consolidate additions and subtractions of numbers below 100 by practicing and repetition (Van Grootheest et al., 2009). Practicing and repetition is a well-known phenomenon in education, and this practicing and repetition strengthens the connection between stimuli and response. The stimulus is the trial one has to solve and the response is the solving strategy that leads to the correct or incorrect answer. When one gives the incorrect answer, one gets punished and that suppresses the frequency of giving that response. But when one gives the correct answer, one gets reinforced and that increases the frequency of giving that response (Kalat, 2009). So, each time one gives the correct answer, the connection between the stimulus and the correct response will be strengthened. Furthermore, in this mathematics curriculum, trials are presented in structured rows, and get more difficult step-by-step. The aim of structuring the learning material is that subtraction and addition becomes automatic (Van Grootheest et al., 2009). An automatic process "*can be carried out rapidly and without effort or intention"* (VandenBos, 2007, p. 91). Many researchers make similar claims about heuristics, "whereas *heuristics are time saving mental shortcuts that reduce complex judgements to simple rules of thumb"* (Crisp, & Turner, 2014, p. 45). Also, people use heuristic rules to save cognitive costs (Keane, 2013). The structured approach of curricula facilitates the creation of paths through the nervous system and thus the development of heuristics. And in a structured environment, like the classroom, using heuristics is beneficial because children can use these heuristics for the whole row of similar trials they have to solve. Moreover, each correct answer will strengthen the connection between stimulus and response and thus reinforce the idea that they should use the heuristic to solve the trials. However, the real world, everything outside the classroom, is more variable than the classroom since the trials are not presented in a structured manner. There thus are no rows of trials for which one can constantly use the same heuristic. But these heuristics might be activated in some cases, and will not always be beneficial to generate the correct answer. Therefore, one should ignore the heuristic and start from the beginning with solving the trial step-by-step. One is thus actually using an algorithm to solve the trial and generate the correct answer. Using an algorithm, a precisely defined procedure to solve a trial, costs time and effort but, if correctly followed, will result in the correct answer (VandenBos, 2007). So besides striving for automaticity, another teaching method might be to stimulate children to always use algorithms so that they will not get distracted by automatically activated heuristics. A disadvantage is that it will take more time to solve trials in which a heuristic actually would have been beneficial. But the advantage is that one does not have to ignore the automatically activated heuristic. Thus, to solve trials, there are two teaching methods, focused on the development of heuristics or algorithms, and both have advantages and disadvantages. In the present study, we investigated which teaching method prepares children best for solving trials in the real world.

An ability children need to acquire during their education is to solve simple arithmetic word problems (Lubin et al., 2013). An example of a simple arithmetic word problem is: "John has 6 marbles, Harry has 4 more than John, how many marbles does Harry have?". An important element of a simple arithmetic word problem is the relational term, in this case 'more than'. Furthermore, to solve this problem one has to do an arithmetic operation, in this case 'addition'. In this example, the relational term is consistent with the arithmetic operation. But in the next example, the relational term is inconsistent with the arithmetic operation: "John has 6 marbles, John has 4 more than Harry, how many marbles does Harry have?". The relational term here is 'more than' whilst the arithmetic operation one should do to generate the correct answer is 'subtraction'. Whether the relational term and arithmetic operation are consistent with each other is important because it influences how easy one can solve the problem. Following the "consistency theory" (Lewis & Mayer, 1978) the inconsistent simple arithmetic word problems should be more cognitive demanding to solve than the consistent ones because one needs to rearrange the relational term to solve the trial. Therefore, we expect that solving trials in which the relational term is inconsistent with the arithmetic operation (incompatible trials) costs more time and has a lower accuracy than solving trials in which the relational term is consistent with the arithmetic operation (compatible trials).

To solve simple arithmetic word problems, people use the heuristic “add if more than and subtract if less than” (Lubin et al., 2013). That is, perform addition if the words ‘more than’ are present in the trial, and perform subtraction if the words ‘less than’ are present in the trial. Using this heuristic is beneficial in consistent simple arithmetic word problems because one will quickly generate the correct answer (Lubin et al., 2013). This heuristic is not beneficial in inconsistent simple arithmetic word problems because when one uses the heuristic in inconsistent simple arithmetic word problems one will generate the incorrect answer (Lubin et al., 2013). Whether one is able to solve incompatible trials depends on one’s ability to inhibit the heuristic “add if more than and subtract if less than” (Lubin et al., 2013). Thus, instead of following the automatically activated heuristic one should ignore it and use an algorithm to step-by-step find the correct answer. Hence, if people use the heuristic “add if more than and subtract if less than” they can easily solve compatible trials but they have to inhibit that heuristic and use an algorithm, to solve incompatible trials.

The real world is, in terms of trials, comparable to the classroom, but an important difference is the lack of structure in the real world. And structure, as offered in the classroom, facilitates developing heuristics which is beneficial for solving compatible trials but disadvantageous for solving incompatible trials. Another way of teaching children might be to present all trials randomly so that they cannot develop heuristics and have to use algorithms every time they solve a trial. That is beneficial for solving incompatible trials but time consuming for solving compatible trials. In the present study we investigated which method works best for preparing children to solve trials in an unstructured environment. We expect that learning in a structured environment will lead to fast reaction times and high accuracy on compatible trials and slower reaction times and lower accuracy on incompatible trials. We expect the same difference but smaller for learning in an unstructured environment because one can use the same strategy for solving compatible and incompatible tasks.

Method

Respondents

63 Respondents (mean age (±SD): 22.3 ± 5.4 years; 12 males and 35 females) participated in this study. They received the link to participate in the internet survey via Leiden University or social media. Leiden University students earned credits for completing the survey. Non-students earned nothing.

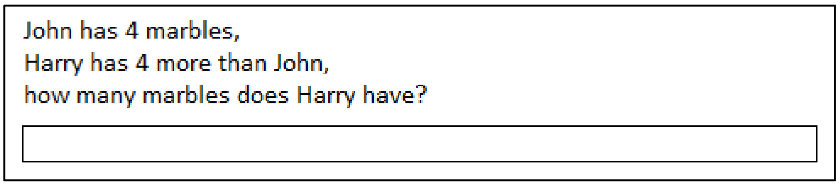
We could only use data from respondents that completed the whole survey so we had to exclude 13 respondents (21% of the sample). Furthermore, we excluded one respondent (2% of the sample) because the reaction times were not saved and two respondents (3% of the sample) because they had given an incorrect answer to more than 10% of the trials.

Procedure

Each respondent participated in a training set and a test set. These two sets were separated by a break of at least 120 seconds.

The training set consisted of two conditions: homogeneous and heterogeneous. Each respondent participated in one of these conditions. The trials in the homogeneous condition were presented structured, namely, the trials were subdivided in eight blocks of ten trials per block. In the heterogeneous condition and the test set, all trials were randomly presented.

In each condition, respondents solved 80 simple arithmetic word problems (trials) of which 40 were compatible and 40 were incompatible. Each trial was built up from three sentences that each were presented on a new line, followed by a text entry box in which respondents could type their answer (Figure 1).



John has 4 marbles,
Harry has 4 more than John,
how many marbles does Harry have?

Figure 1 An example of a compatible trial. The trials were given in Dutch but the structure is the same.

In the lower right corner of the screen, respondents could click the box ‘next’ after they had given their answer. They could only continue to the next trial if they had typed in an answer and they did not get any feedback about the correctness of their answer.

We created trials with four arithmetic operations: addition, subtraction, multiplication and division. The addition and subtraction operations always were plus or minus 4. The multiplication and division operations always were multiplication or division by 2. These arithmetic operations were linked to four relational terms, namely: ‘more than’, ‘less than’, ‘double’ and ‘half’. In compatible trials the relational term was consistent with the arithmetic operation. In incompatible trials the relational term was inconsistent with the arithmetic operation. Note that one could use the heuristic “add if more than and subtract if less than” in the compatible trials but not in the incompatible trials. For each relational term, twenty trials of which ten compatible and ten incompatible were created, resulting in eight blocks of ten trials per block. In the homogeneous condition, the eight blocks of trials were randomly presented, as well as the trials within the blocks. In the heterogeneous condition and the test set, all trials were randomly presented.

To create the trials, we used 40 names and 20 objects. We randomly made Name1-Name2-Object-combinations. We also made counterbalanced versions of each condition, to make sure that any potential interference effect of a certain combination would be present in both versions and thus not influence the difference between conditions. The Name1-Name2-Object-combinations in compatible trials in version A were used in the incompatible trials in version B. The Name1-Name2-Object-combinations in incompatible trials in version A were used in the compatible trials in version B.

Furthermore, to filter out any unwanted learning effects, we used various numbers for arithmetical operations in the training and test set. In the training set, we used the following numbers: 4,6,8,12,20,24,32,40,44, 52,60. And in the test set, we used: 6,10,14,22,30,34,42,50,54,62. We left out the numbers where, when multiplied, the answer would be more than ten (e.g. 14 × 2 = 28 but 16 × 2 = 32).

Respondents were randomly assigned to a training condition (homogeneous or heterogeneous) and to the test set.

Measurements

To measure whether there were any differences between the conditions, we looked at the average reaction times and the accuracy (percentage of trials answered correct). We measured the reaction times as the time that passed before respondents clicked the 'next' button after they had typed in an answer. And we measured the accuracy by dividing the number of correct answered trials through the total number of trials.

Furthermore, we did exploratory analysis because we expected a relationship between impulsivity and the ability to inhibit heuristics. We measured impulsivity with the Barratt Impulsiveness Scale-11 (translated by Lijffijt & Barratt, 2005), and we used the score as defined by Chen (2013). The Barratt Impulsivity Scale-11 has three subscales: the BIS-motor, BIS-cognitive and BIS-nonplanning scales. The BIS-motor scale measures control of motor action, the BIS-cognitive scale measures the quality of attention span and the BIS-nonplanning scale measures the self-control in planning for the future (Chen, 2013).

Data-analysis

We only analyzed data gathered from the test set. We had two measurements: reaction times (RTs) and accuracy. RTs for incorrect trials were excluded. Furthermore, outliers were defined as RTs that were greater than 20 seconds and, after deleting these, RTs that were greater than 2 SDs from the mean for that respondent. We chose the limit of 20 seconds because the trials were such that respondents should be able to answer them within 20 seconds and we wanted to exclude RTs of respondents that were stuck. After deleting the RTs for incorrect trials and the outliers, we calculated the mean RT for compatible trials and the mean RT for incompatible trials, for each respondent. Before calculating the accuracy, we deleted answers to trials (independently of correctness) for outlier RTs. After that, we calculated the accuracy for compatible and incompatible trials by dividing the number of correct answers through the total number of answers (excluding the ones that were linked to outlier RTs).

We performed two 2x2 mixed ANOVA's; one for reaction time and one for accuracy, with difficulty (compatible or incompatible) as within variable and training condition (homogeneous or heterogeneous) as between variable.

For the exploratory analysis concerning impulsivity, we calculated the mean impulsivity for each scale as defined by Chen (2013). We performed Pearson correlations between the scores on these three scales and the RTs and accuracy.

Results

Analyses were conducted on 47 respondents (12 male, 35 female). There was no significant difference in male – female ratio between the homogeneous and heterogeneous condition: $\chi^2(1) = .86$, $p = .35$.

We conducted two 2x2 mixed ANOVA's, with difficulty (compatible and incompatible) as within variable and training condition (homogeneous or heterogeneous) as between variable for both RT and accuracy.

We did not find a significant effect of difficulty on reaction time. Furthermore, we did not find a significant interaction effect between trial difficulty and training condition on RT. Respondents thus were not slower on incompatible trials than compatible trials.

We found a significant effect of difficulty, $F\ (1, 45) = 7.34$, $p < .01$ on accuracy. Consistent with our expectations, pairwise comparisons revealed that compatible trials were more often answered correct than incompatible trials ($M_{compatible} = .97$ (SD = .01) and $M_{incompatible} = .96$ (SD = .01)). We found an interaction effect that approached significance between difficulty and training condition $F\ (1, 45) = 1.85$, $p = .18$ on accuracy. Considering that interactions including a between subjects variable often require a large sample size to reach significance, we split the file on the training condition and did a paired samples T-test. We have to interpret these results with caution, but we found a significant difference $t(24) = 3.07$, $p < .01$ of difficulty for the heterogeneous training condition (*M* = .02, *SD* = .04), and not for the homogeneous training condition *t*(21) = 0.90, *p* = .38. Thus, the heterogeneous trained respondents made less mistakes on compatible trials than incompatible trials and there is no significant difference for the homogeneous trained respondents.

Furthermore, we did exploratory analysis because we expected a relationship between impulsivity and the ability to inhibit heuristics. We questioned the Barratt Impulsivity Scale-11 (Lijffijt & Barratt, 2005), split the file on training condition and calculated Pearson's correlations. We will discuss, per training condition, all correlations that have a significant effect with p <.10. We set a liberal threshold of p<.10 because we wanted to do exploratory analysis that included the effects that not quite reach significance. Furthermore, this is a first step in investigating whether there is a correlation in the first place.

In the heterogeneous training condition, we found a moderate positive correlation between the BIS-motor scale and average accuracy r = .36, p <.10. Thus, counter intuitively, respondents who do not hold off impulsive actions score high on accuracy. Furthermore, we found a strong positive correlation between the BIS-cognitive scale and average accuracy *r* = .40, *p* <.05. Thus, respondents with a short attention span score high on accuracy.

In the homogeneous training condition, we found a strong positive correlation between the BIS-motor scale and average accuracy *r* = .45, *p* <.05. Thus, respondents who do not hold off impulsive actions score high on accuracy. Furthermore, we found a strong negative correlation between the BIS-cognitive scale and the average reaction times *r* = -.57, *p* <.01. Thus, respondents with a short attention span have fast reaction times.

Discussion

In the present study we investigated which training method (homogeneous versus heterogeneous) to solve arithmetic word problems prepares children best for solving them in an unstructured environment.

Regarding accuracy, we expected that the difference between compatible and incompatible trials would be largest in the homogeneous condition. But we observed exactly the opposite where the difference between compatible and incompatible trials was largest in the heterogeneous condition. In fact, there was no difference in the homogeneous condition which means that the accuracy on compatible and incompatible trials was equal. A possible explanation might be that the homogeneous trained respondents developed a new heuristic, a metaheuristic. This metaheuristic should be applicable to both compatible and incompatible trials and might explain why there is no difference in accuracy between these two types of trials. This metaheuristic, based on the structure of the trials, should then be "if the first two sentences start with a similar name, do the opposite of the "add if more than and subtract if less than" heuristic". In incompatible trials, the first two sentences start with the same name, while in the compatible trials the second sentence starts with another name. So, for solving compatible trials, the "add if more than and subtract if less than" heuristic still is applicable and for solving incompatible trials, one can use the metaheuristic. This metaheuristic was probably developed in the homogeneous condition only because respondents had to solve the same kind of trial ten times in a row. This repetition might have given them the opportunity to strengthen the connection between stimulus and response, where the response for incompatible trials developed from using an algorithm to using the metaheuristic. In the heterogeneous condition, where all trials were presented randomly and thus were not repeated in a row, respondents did not have the possibility to develop this metaheuristic. Future research should be done to investigate whether this metaheuristic is actually developed the way we think it is. A possible manner of doing that might be to develop two training conditions, one which is focused on developing a metaheuristic and one that is focused on not developing a metaheuristic. In the test set, participants get two types of trials where the metaheuristic can be beneficial or not. An important difference with the present study is that in the suggested study the two trials are equally easy for the participants that have not developed a metaheuristic. A possible outcome might be that the group that developed the metaheuristic performed better on the trials for which the metaheuristic is beneficial and worse on the trials for which the metaheuristic is not beneficial. If the group that had not developed the metaheuristic performed similar on both trials, the difference in performance might be explained by the development of that metaheuristic. To conclude, regarding accuracy, our findings seem to suggest that, in contrast to our hypothesis, the homogeneous training method to teach respondents how to solve arithmetic word problems worked best for solving trials in an unstructured environment. The respondents did not have more difficulties in generating an answer for incompatible trials than compatible trials.

In contrast to our expectations, we found no significant differences between compatible and incompatible trials on RTs for both training conditions. A possible explanation is that the respondents made the survey at home, and although we advised them to minimize the amount of distraction, they will probably be distracted sometimes. Furthermore, we measured the reaction time as the point in time where the respondent clicked 'next' to go to the next trial. In future research, the RT could be measured more precise by measuring it as the moment in time where one types in the answer. A slightly different approach is to give multiple choice instead of open questions. Then the RT could be measured as the moment in time when one ticks a box. Or, if two choices, when one presses a button. An additional advantage of that approach is that typo's are not possible.

We cannot generalize these results to children because our respondents were all adults, and as discussed in Lubin et al. (2013), inhibitory control is less developed in children than in adults. So, more research might be done with children of at least 10 years old as participants. They should be at least 10 years old because then they are able to solve the simple arithmetic word problems (Verschaffel, 1994, as described in Lubin et al., 2013). They could participate in the same conditions as the adults in the present study did. They will probably make more mistakes on incompatible trials because their inhibitory control is less developed (Lubin et al., 2013) and they will thus have difficulties in inhibiting the heuristic. Therefore, they might benefit more from training in an unstructured environment where they thus cannot develop the heuristic and do not have to inhibit it for incompatible trials. If children indeed perform better when trained in an unstructured environment, that might have consequences for the use and structure of mathematics curricula as used nowadays.

We did exploratory analysis to investigate the relationship between being impulsive and using heuristics because we expected that impulsive people would have more difficulties in inhibiting the heuristic. That reasoning is supported by Logan, Schachar, and Tannock (1997) who revealed that impulsive people appear to have difficulty inhibiting automatic responses because their inhibitory responses are very slow. It thus takes more time and effort for impulsive people to inhibit a heuristic. However, that does not accord to our results. In the homogeneous condition we found a strong positive correlation between impulsivity and accuracy and a strong negative correlation between impulsivity and reaction times. Impulsive people thus were more accurate and had faster reaction times. A possible explanation is that impulsive people developed and used that metaheuristic so that they could easily and quickly generate the correct answer for both compatible and incompatible trials. Participants thus did not have to inhibit a heuristic that was automatically activated but not beneficial for incompatible trials. Instead, they could always use automatically activated heuristic for compatible and metaheuristic for compatible trials. In the heterogeneous condition we found a strong positive correlation between impulsivity and accuracy. Future research should be done to explain what is in these trials that impulsive respondents also are more accurate. A possible starting point might be to give two answer opportunities instead of letting participants type in the answer themselves because then participants have no chance to correct their answer and their reaction times will be more accurate. It would be interesting to see if the same pattern can be found using this measuring method.

In conclusion, our findings do not support our idea that people perform better in situations where they cannot use heuristics if they're trained not to do so. In fact, people seem to be so flexible that they create a metaheuristics that helps them generating the correct answer in a quick manner. Therefore, the question should not be whether we should or should not avoid that children develop heuristics but rather the question should be what heuristic we should teach them that is most beneficial in the real world.

References

Chen, W.Y. (2013). *Neuroinvesting: Build a New Investing Brain.* Singapore: John Wiley & Sons Pte. Ltd.

Crisp, R. J., & Turner, R. N. (2014). *Essential Social Psychology (3rd ed.).* London, UK: SAGE Publications Ltd.

Janssen, J., Van Der Schoot F., & Hemker, B. (2005). *Balans [32] van het reken-wiskundeonderwijs aan het einde van de basisschool 4.* Arnhem, The Netherlands: Stichting Cito Instituut voor Toetsontwikkeling.

Kalat, J. W., (2009). *Biological Psychology (10th ed.)*. Belmont, USA: Wadsworth

Keane, C. (2014). *Modeling behavior in complex public health systems: Simulation and games for action and evaluation.* New York, NY: Springer Publishing Company, LLC.

Lewis, A.B., & Mayer, R.E. (1987). Students' miscomprehension of relational statements in arithmetic word problems. *Journal of Educational Psychology, 79,* 363-371.

Logan, G. D., Schachar, R. J., Tannock, R. (1997). Impulsivity and inhibitory control. *Psychological Science, 8,* 60-64.

Lubin, A., Vidal, J., Lanoë, C., Houdé, O., & Borst, G. (2013). Inhibitory control is needed for the resolution of arithmetic word problems: A developmental negative priming study. *Journal of Educational Psychology, 105,* 701-708.

Lijffijt, M., & Barratt, E.S., (2005). Persoonlijke evaluatie: BIS-11. Retrieved from http://www.impulsivity.org/measurement/BIS-11Dutch.pdf

VandenBos, G.R. (2007). *APA Dictionary of Psychology (1st ed.)*. Washington, DC: American Psychological Association.

Van Grootheest, L., Huitema, S., De Jong, M., Munsterman, B., & Osinga, H. (2009). *De wereld in getallen, leerkrachtenmap groep 4 (1e druk, 4e oplage).* 's-Hertogenbosch, The Netherlands: Malmberg.