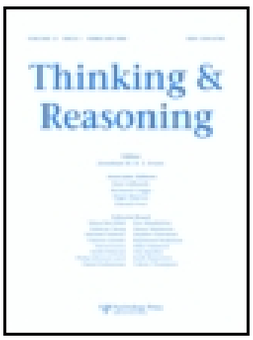

Thinking & Reasoning

Routledge
Taylor & Francis Group

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ptar20

Schoolchildren's transitive reasoning with the spatial relation 'is left/right of'

Kevin Demiddele , Tom Heyman & Walter Schaeken

To cite this article: Kevin Demiddele , Tom Heyman & Walter Schaeken (2020): Schoolchildren's transitive reasoning with the spatial relation 'is left/right of', Thinking & Reasoning, DOI: 10.1080/13546783.2020.1843536

To link to this article: https://doi.org/10.1080/13546783.2020.1843536

THINKING & REASONING
https://doi.org/10.1080/13546783.2020.1843536

Schoolchildren's transitive reasoning with the spatial relation 'is left/right of'

Kevin Demiddele[a], Tom Heyman[a,b] and Walter Schaeken[a]

[a]Laboratory for Experimental Psychology, KU Leuven, Leuven, Belgium; [b]Laboratory for Experimental Psychology, Leiden University, KU Leuven, Leuven, Belgium

ABSTRACT

We examine schoolchildren's reasoning with spatial relations, such as 'is to the left of'. Our aims are to obtain a more precise account of the effect of working memory on reasoning, a more detailed understanding of the internal representation of mental models and a developmental perspective. We discuss two experiments in which 348 children, between eight and twelve years old, needed to verify conclusions for 24 reasoning problems describing the spatial relations between pieces of clothing. In both experiments, children in the experimental condition were allowed to take notes by means of paper and pencil. We find that the participants spontaneously draw iconic representations of the items' spatial ordering, have a strong preference for only considering one possible state of affairs even when more are relevant, and that an explanation in terms of working memory capacity alone cannot fully explain the data.

ARTICLE HISTORY Received 25 February 2020; Accepted 25 October 2020

KEYWORDS Spatial reasoning; mental models; working memory; developmental psychology

Developmental accounts of reasoning come in different flavours. One strand of theories focusses on the syntactic aspect of reasoning. This includes computationalist accounts, ultimately going back to the work of Fodor (1975, 1983), explaining relational reasoning as symbolic reasoning. Alternatively, but still in the syntactic camp, explanations in terms of proportional reasoning are grounded in Inhelder and Piaget (1958)'s concept of formal operational reasoning. A third group of syntactically focussed theories is that of the mental logic theories, rooted in work of Braine (1978) and Rips (1983). In contrast to these syntactically focussed theories, mental model theory (cf. Johnson-Laird, 1983, 2006) is more semantically focussed, and particularly well suited to deal with reasoning with spatial relations. Hence, we will use its framework as our theoretical point of departure.

CONTACT Kevin Demiddele kevin.demiddele@kuleuven.be Laboratory for Experimental Psychology, KU Leuven, Box 3711, Tiensestraat 102, Leuven, 3000, Belgium

Mental model theory says that reasoners mentally construct representations that are iconic to the information they have processed. Consider, for example, the following premise.

The hat is to the left of the shirt

This can be represented by the mental model below.

H S

Keeping such models in mind taxes working memory and the more complex they are, the more will be demanded in terms of working memory. For example, consider the following premise being added to the information.

The jacket is to the right of the shirt

Reasoners will not keep the premises in mind, but, according to the principle of economicity (Cf. Manktelow, 1999), incrementally add the information in the premises to their mental model. In this case, this is still relatively simple, as only one model representation is consistent with this information:

H S J

On the other hand, consider the following premise to be the third piece of information.

The cap is to the right of the shirt

Now at least two different representations can be constructed.

H S J C and H S C J

Both are consistent with the premises. Anyone realising this and representing these possibilities mentally, needs to invest more working memory effort than someone only representing one of both. In fact, it has convincingly been shown (Goodwin & Johnson-Laird, 2005; Jahn et al., 2007; Ragni et al., 2007) that reasoners by default only do construct one of both, the explanation being that this is done because of parsimonious use of working memory. In general, working memory capacity is known to limit reasoning capability (Cf. Bara et al., 1995; Gilhooly et al., 1992; Klauer, 1997).

What is currently lacking, is a precise account of the effect of working memory, more detailed understanding of the internal representation of mental models and a developmental perspective. The argument by Wright (2001) that 'the issue of how and when children are able to handle transitive inference remains unsettled' (p.385) still holds today. To obtain the developmental perspective, we have chosen children in different age groups as participants for our experiments. For more insight in their internal representation and a more specific understanding of the effect of working memory capacity, we allow them to use paper and pencil in the experimental condition. The assumption that reasoners who are given paper and pencil will write down something that matches their mental representations to some extent, has been made earlier by, for example, Van der Henst et al. (2002) and Bucciarelli and Johnson-Laird (1999). Moreover,

taking notes by means of paper and pencil can be an auxiliary to remember things and thus, alleviate the weight on working memory, as explained by Bauer and Johnson-Laird (1993). In order to maximise this effect of the working memory aid, it also makes sense to try this kind of experiment with participants that are known to have limited working memory, i.e. children (see Gathercole et al., 2004). At the same time, we should be ascertained that participants have the ability to conduct the type of reasoning exercises we want to present them. Our participants ranged from eight- to twelve-year olds. This is certainly old enough to deal with transitive inferences (see Andrews & Halford, 1998; Pears & Bryant, 1990), even taking into account that 'young children (e.g., of 4 or 5 years) can reason transitively but not as logically underpinned' (Wright, 2001, p.414). Moreover, the link between a spatial mental model and reality is tight. Depicting a relation such as 'is to the left of' is simple and intuitive by putting it to the left of the other item in the relationship. At the same time, our chosen age group is young enough to have more limited working memory capacity than adults. As 'to the left/right of' is the only relation we question, combined with 'above/below' in the premises, this is an investigation into these specific spatial relations more than it is an investigation into transitive reasoning in general. The ability to judge whether a relation is transitive or not hardly comes into play, while this should be part of general transitive reasoning research, as described by Wright (2001). Compared to basic three-term transitive reasoning (where on the basis of two premises, like A $<$ B and B $<$ C the relation between two mentioned items must be inferred, like A ? C), the task in our experiment was more complex, as it involved understanding multiple possibilities. Also at least three premises, i.e. four terms, were required, for the MMv problems. Even when working with blocks instead of text, this proved to be more difficult for younger children (cf. Markovits et al., 1995) than basic two-premise problems. Finally, our experimental setup was one that essentially involved reading of the premises and note taking. Hence, fluent reading skills were required. This explains why we considered third-graders to be the youngest possible age group for this experiment. We erred on the side of caution by choosing fourth-graders for the first experiment, to ascertain that the task was not too difficult for the younger group. With sixth-graders as older group, we were still focussing on children rather than adolescents and remained within the Piagetian concrete operational stage. As such, our age group complements lots of research that has been done on slightly younger children or children matching our youngest group (cf. Ameel et al., 2007; Andrews & Halford, 1998; Markovits et al., 1995; Pears & Bryant, 1990; Wright & Smailes, 2015).

If children's reasoning performance is impaired because of limited working memory, and we provide them with a mechanism, i.e. taking notes, to

overcome this limitation, we would expect the result to be a significant improvement of reasoning performance. Moreover, we would expect a clear effect for the cases with multiple possibilities, as these put the heaviest load on working memory. That is, unless the bias to construct only one single model is based on more than working memory limitation alone.

Experiment 1

Method

Participant info

We tested 216 children. There were 106 boys and 110 girls; 120 sixth-graders (M = 11.48 years; SD = 0.28) and 96 fourth-graders (M = 9.47 years; SD = 0.28). All data was collected at schools in Flanders. All children's data was anonymized before processing. The experiment was approved by the social and societal ethics committee of KU Leuven and all participating children had an informed consent signed by their parents.[1] Among the fourth-graders, 55 children were assigned to the control condition and 41 to the experimental condition. Among the sixth-graders, 65 and 55 children were in the control and experimental condition, respectively.

Procedure

The experiment leader collected the data per class. Each class as a whole was assigned to either the experimental or control condition. During the data collection, she followed a script to ensure that the same instructions were given in each class. Before handing the exercises to the children, she introduced the topic by collectively solving a real life example of the type of problem they were to encounter in the exercises. The example consisted of describing the relative positions of real pieces of clothing, for which no prior expectations could be assumed, and inferring information from that. She showed that a sweater was to the right of a pair of trousers and a cap was to the right of the sweater. Then she agreed with the children that the cap was also to the right of the trousers. The children were instructed that if they thought there was no correct answer, there were multiple possibilities or they did not know, they should choose the answer option 'none of the above'. We will discuss this option in more detail later. Furthermore, they were instructed that they would not receive any grade on the test but also told that we needed them to do their best in order to be able to compare the data with those of other children. Once the introduction was over, they each received their exercise sheets and could start working individually.

[1]The reference number of the ethics committee approval is G-2017 11 970.

Material

All children received 24 reasoning problems. These consisted of premises describing relative positions of pieces of clothing, on the basis of which they had to draw a conclusion that could be chosen from multiple choice options.

There were three problem types. The first problem type, single model problems (M1), consisted of premises that describe an unambiguous arrangement of items. Hence, for all of these problems it was possible to infer what the relation between the question items was and the answer was either option a or option b. Here is an example of a single model problem, translated to English from the Dutch original. In Dutch all nouns were singular.

The trousers are to the left of the cap.
The skirt is to the left of the trousers.
Where are the skirt and the cap relative to each other?
a. The skirt is to the right of the cap
b. The cap is to the right of the skirt
c. None of the above

The correct, unambiguous mental model that can be constructed for these premises is

S T C

The second problem type, multiple-model problems with a valid conclusion (MMv), consisted of premises that describe a situation consistent with two different arrangements of items, but posed a question on items that had the same relation in both of these representations. Thus, for these problems, too, the correct left-right relation between the question items could be inferred. It is important to realise that these problems could be answered correctly by constructing only one of the two possible representations and judging the conclusion based on that one model, possibly without even realising that multiple possible representations are involved. Add 'the jacket is to the left of the trousers' as third premise to the example above. The result is a description of a situation where the jacket can be either left or right of the skirt, but this does not matter for the question at hand, as in both possible representations, the skirt is to the left of the cap.

J S T C and S J T C

Finally, the third problem type were multiple-model problems with no valid conclusion (MMnv). In these problems, premises also described a situation consistent with two different arrangements of items, but now a question was posed on items that had a different relation in both of these representations. No valid conclusion could be inferred from these premises so the correct answer to these problems was always option c, 'none of the above'. To obtain such a problem, again add 'the jacket is to the left of the

trousers' as third premise to the problem above, obtaining the same two possible mental models, but now change the question to 'where are the skirt and the jacket relative to each other?'. As this relation is different in each of the two models, a valid conclusion is not possible.

Next to these one-dimensional problems, there were also two-dimensional ones, that also included the relations 'is above' and 'is below'. The question, however, was always on a left-right relation. Half of the problems were two-dimensional, counterbalanced with problem type. Two-dimensional problems always had four premises. One dimensional problems had two premises for the M1 and MMnv problems and three premises for the MMv problems, which are not possible to construct with two premises only. As there was no significant main effect of dimension, nor any significant interaction with dimension, in both experiments, it is safe to conclude that two-dimensional problems were of comparable complexity and difficulty to the one-dimensional problems. To avoid overcomplicating the results, dimension has been left out of the reported analyses.

There were five different items (the skirt, the jacket, the trousers, the sweater, the cap), which were re-used for all problems. Each child solved eight M1 problems, eight MMv problems and eight MMnv problems. Problems were presented in different randomised orders. In the experimental condition, there was blank space (7 cm x 16 cm) below the question and answer options, where the children could write or draw what they wanted. We did not provide instructions on how to use this space. The experiment leader only mentioned 'you can take notes if you think this can help you'. In the control condition, there was no blank space and children were not allowed to take notes, which was checked by the experiment leader.

Results and discussion

Effect of type, notes and grade on accuracy

The data that support the findings of this study are openly available on OSF.[2] No data was excluded. We performed a multilevel analysis, predicting scores from type, notes and grade including all their interactions, with individual participants as random intercept.[3] The R^2 of the resulting model was .38 (based on the theoretical variance) and .33 (based on the observation-level variance via the delta method of Nakagawa et al., 2017). If we only look at the fixed effects, the R^2 estimates were .29 and .26, respectively. Critically, there was a significant three-way interaction between notes, grade, and problem type ($\chi^2(2) = 12.81$, $p = .002$), but the pattern of results was not one we had expected a priori (see Figure 1).

[2] https://osf.io/ukep3/

[3] We used the *afex* package (v 0.20.2) in R (v 3.4.3 (2017-11-30) "Kite-Eating Tree")

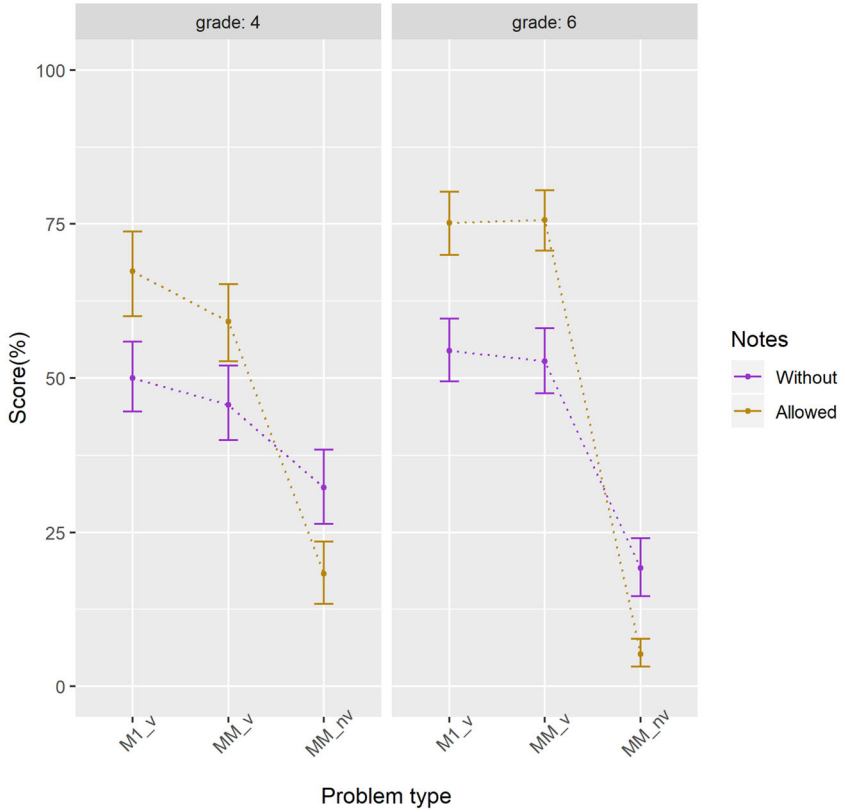

Figure 1. Experiment 1 mean scores in percentages by notes group, grade and problem type.

Taking notes proved to be beneficial for only two of the three problem types. Both fourth- and sixth-graders performed significantly better on single-model problems and multiple-model problems with valid conclusion when given the opportunity to take notes. Surprisingly, they performed worse on multiple-model problems without valid conclusion when allowed to take notes. Moreover, compared to fourth-graders, sixth-graders showed a stronger beneficial effect of note taking on multiple-model problems with valid conclusion, but also a stronger adverse effect of note taking on multiple-model problems without valid conclusion (see Tables 1 and 2).

More detailed analysis of the note taking

The children in the notes condition had the opportunity to take notes, but not all of them took advantage of this. We coded the notes that were taken, with the following results. Overall, notes were taken for 80% of all problems, with 74% for the fourth-graders and 85% for the sixth-graders. Most of the time, participants took notes for all problems, as can be seen in

Table 1. Effect of note taking for each problem type by grade combination.

Problem type	Grade	Estimate	Standard error	Z	p
M1_v	6	1.05	0.19	5.39	< .001
MM_v	6	1.15	0.20	5.91	< .001
MM_nv	6	−1.58	0.28	−5.62	< .001
M1_v	4	0.79	0.21	3.72	< .001
MM_v	4	0.59	0.21	2.82	.005
MM_nv	4	−0.82	0.23	−3.57	< .001

Table 2. Differential effect of note taking across grade for each problem type.

Problem type	Estimate	Standard error	Z	p
M1_v	0.26	0.29	0.92	.356
MM_v	0.57	0.29	1.99	.046
MM_nv	−0.76	0.36	−2.09	.037

Figure 2. Whether notes were taken or not did not depend on problem type ($\chi^2(2) = 3.35$, $p = .187$), but did depend on problem order ($\chi^2(1) = 24.65$, $p < .001$), as shown by multilevel analysis predicting amount of notes taken by problem type and problem order, with individual participants as random intercept. Less notes were taken on later problems, possibly because some participants grew tired of taking notes throughout the experiment. The type of notes taken were always spatial models of the items, with the items represented mostly by words describing the pieces of clothing, followed by drawings of the clothing, followed by a combination of both and exceptionally initial letters or symbols, as can be seen in Figure 3.

Of the notes taken, 73% were correct, i.e. the notes reflected the information in the premises without mistakes, but possibly without being complete. Fourth-graders took correct notes in 61% of the cases, while sixth-graders did so in 81% of the cases. Of these correct notes, 93% were complete (fourth-graders 85% and sixth-graders 97%), i.e, the notes fully described a *single* mental model as described in the premises. Note that this did not take into consideration information on multiple models. We kept track of this separately. Of the 1224 occasions in which notes were actually taken for multiple model problems, only six (i.e. only 0.49%) showed information that we interpreted as representing multiple models, with all 6 problems answered correctly. The comparison of the accuracy scores between the cases in which notes were taken and those in which notes were not taken showed a similar pattern as when comparing the cases in which notes were allowed (the 'notes' condition) with those in which notes were not allowed (the 'no notes' condition). The latter comparison was shown in Figure 1, while the former, exclusively with data within the 'notes' condition, can be observed in Figure 4.

Whether the notes taken were correct or not significantly predicted the score for the problems, in a mixed model predicting score from notes

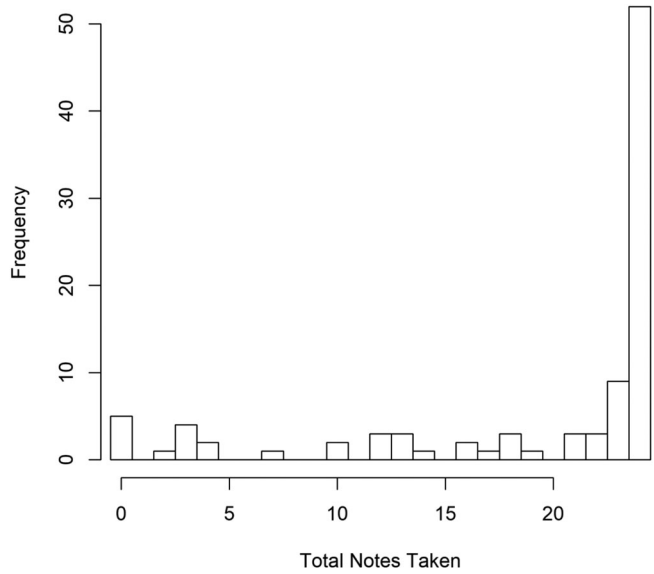

Figure 2. Frequency distribution of total amount of notes taken in experiment 1.

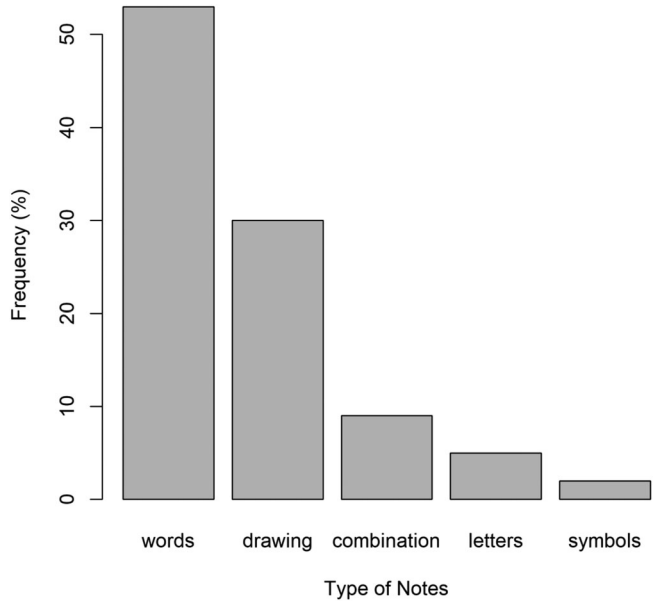

Figure 3. Type of notes taken in experiment 1.

correctness and problem type with individual participants as random intercept ($\chi^2(1) = 22.14$, $p < .001$). There was also a significant interaction with problem type ($\chi^2(2) = 61.90$, $p < .001$), which can be explained by the fact that a correct, and even complete representation of an MMnv model does

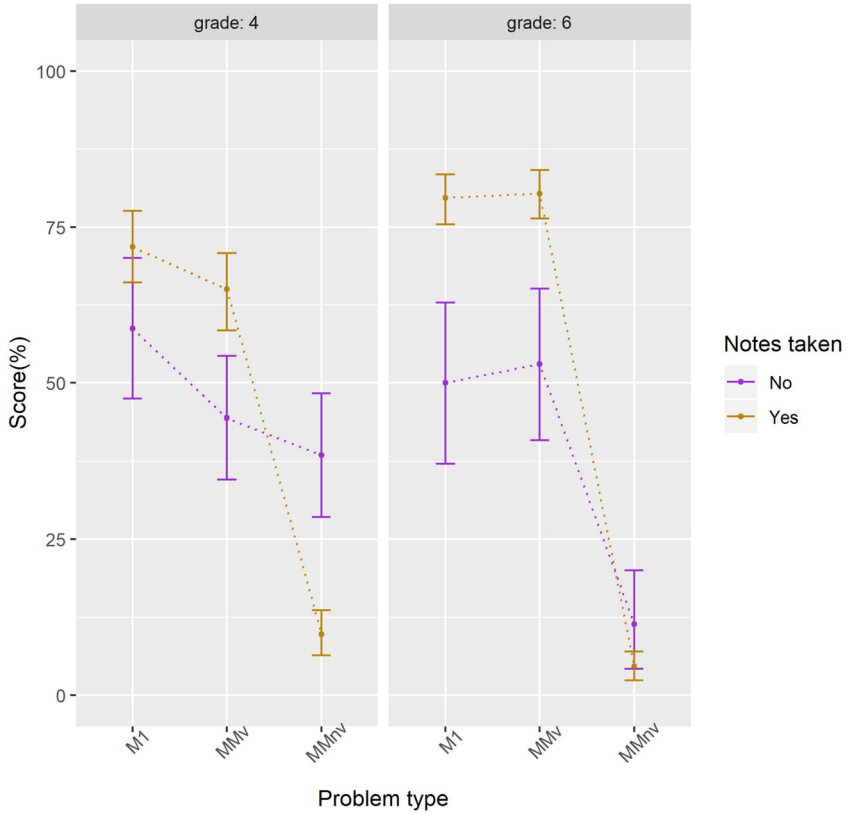

Figure 4. Experiment 1 mean scores in percentages for the notes group, by notes taken, grade and problem type.

not lead to a correct answer as long as there is no understanding of the multiple possibilities. Multilevel analysis predicting note correctness from grade and problem type, with individual participants as random intercept, showed that sixth-graders had significantly more correct notes than fourth-graders ($\chi^2(1) = 9.19$, $p = .002$), as pictured in Figure 5.

Our interpretation of the results is that most children always constructed just one model. This is in accordance with the claim (in Goodwin & Johnson-Laird, 2005; Jahn et al., 2007; and Ragni et al., 2007) that, not only children, but all reasoners in general by default build a single, simple, and typical mental model but neglect other possible models. The children who could take notes were better at this than the ones that could not and consequently scored better on the M1 and MMv problems. However, for the MMnv problems this was not the right strategy. The one model that was constructed in the note taking condition was only one of the possible models, which resulted in a worse score for that problem type when compared to the situation where note taking was not allowed. In the latter condition,

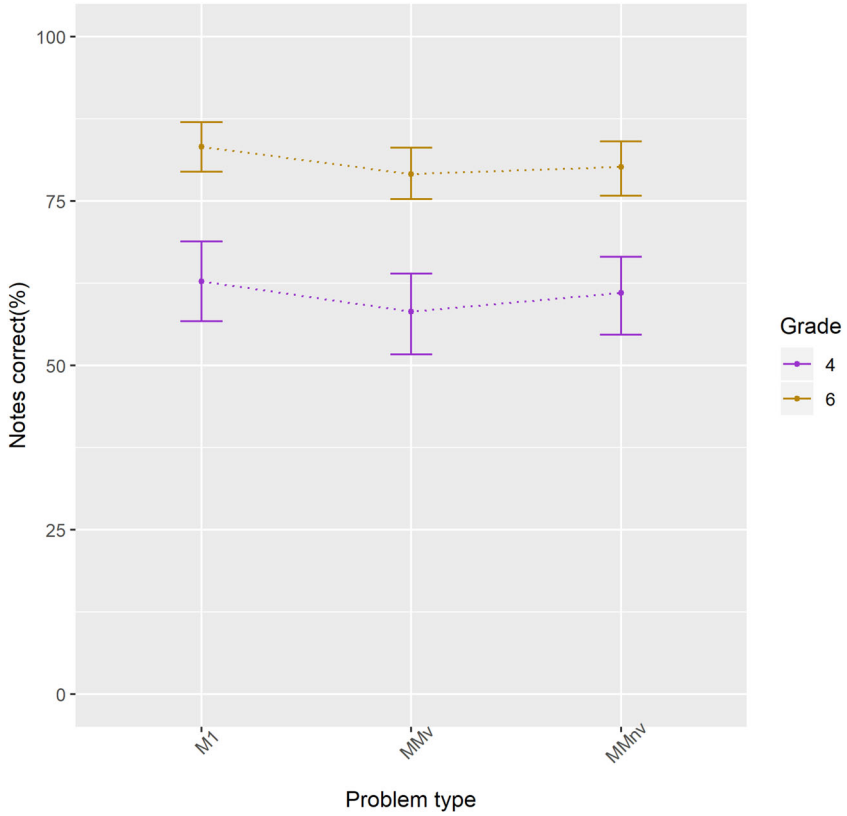

Figure 5. Experiment 1 correct notes in percentages for the notes group when notes taken, by grade and problem type.

the difficulty of even constructing one model might have led more participants to the correct 'none of the above' option, a hypothesis we will explore in more detail later. In other words, taking notes seemed to improve performance on the easier problems, whether by facilitating the synthetisation of information, enhancing children's focus and motivation, and/or decreasing the working memory load. At the same time, it appeared to prompt them to disregard multiple possibilities. This was especially true for older children, presumably because note taking skills improve with age.

When plotting the scores, we can clearly see there is a cut-off at 16 for the notes condition. This is because the best performing children scored 16/16 for the simple problems and 0/8 for the ones that required multiple models. There was just one exception of a child that scored more than 16/24, as illustrated in Figure 6.[4] This was also the only participant in

[4]Plots were created using the 'yarrr' package (Phillips, 2017), adapted by means of code from Stephen Politzer-Ahles, available at http://www.mypolyuweb.hk/~sjpolit/Steve_functions.txt

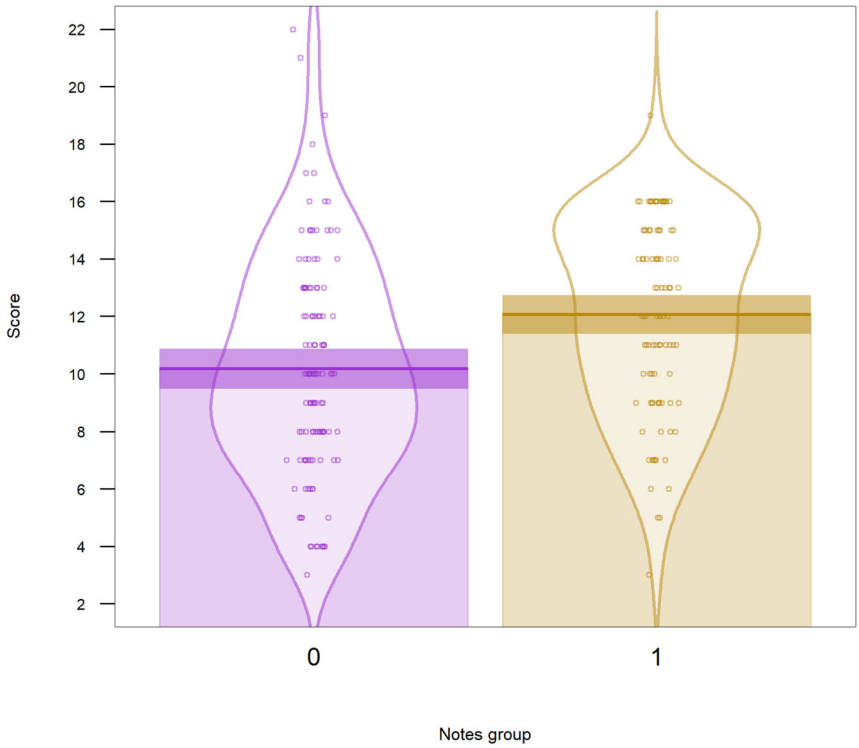

Figure 6. Total scores in Exp 1 by notes groups. The plots show the total scores (out of 24), with 1 point per child. Mean and 95% CI are indicated, as well as plot density by total score.

Experiment 1 that made multiple drawings which represented multiple possibilities. In the group that could not take notes, although their mean score is lower, we see that some children did manage to understand the multiple-model cases and scored above 16/24.

One of our research goals was getting a better understanding of the internal representation of our reasoners. For this reason we provided them only with blank space and no instructions about how they could use it, apart from stating that they could use it in any way they wanted. This way, we wanted to see what kind of notes they would make spontaneously. It turned out that they spontaneously drew analogical representations, with drawings, words or letters representing the items, arranged on paper in the same way as described in the premises. This arguably shows that they also mentally used the same type of representations for solving these problems. Their notes can be interpreted as an external representation of their mental models. The ones that started out by drawing the items often switched to more economical representations as the experiment proceeded. This was likely because they grew tired of drawing unnecessary details but at the

same time can have been functional as for mental reasoning activity visual imagery can impede reasoning (Cf. Knauff & Johnson-Laird, 2002).

Our specific interest was in whether and how they would represent multiple possibilities. Given the results, it will not come as a big surprise that not many children did represent multiple possibilities. Of the 96 children that were in the notes condition, we only identified one who drew representations of multiple possibilities multiple times, with three others drawing them once. The sixth-grader who drew multiple models three times for MMnv problems also correctly answered the corresponding questions, indicating understanding of the multiple possibilities and scoring 19/24, i.e. correctly answering all 16 problems for which construction of one model was sufficient plus three correct multiple-model cases with no valid conclusion.

Analysis of type of mistakes

We also analyzed the type of mistake the participants made. For the multiple-model problems with no valid conclusion, this was encoded in terms of first-fit versus first-free-fit strategy, as introduced in Ragni et al. (2007) and Ragni and Knauff (2013). We will briefly explain this with an example. Consider the following model:

A B

If the next premise is 'C is to the right of A' and you only construct a single model, you can either add C to the right of B or add C in between A and B. The latter strategy is called the 'first-fit' (ff) strategy, while the former is called the 'first-free-fit' (fff) strategy. According to Ragni et al. (2007), reasoners by default construct their preferred model according to the fff-strategy, in which the items that were already present in the model are not altered. In this theory, reasoners put items in adjacent 'cells' and prefer to leave already placed items in their cell and use the first free cell. In the example, you cannot know whether C is to the left or right of B, because both are consistent with the information. However, if you have only constructed one model, you will draw an invalid conclusion. In case you constructed your model according to the ff-strategy, you will conclude that C is to the left of B, while if you adopted the fff-strategy, you will conclude that C is to the right of B. For all MMnv mistakes the children made, we looked whether they made their error in the ff or the fff direction. We performed a multilevel logistical regression, taking into account the individual participants as random intercept. Only for the fourth-graders that could take notes was the result significant, as can be seen in Table 3.

Also for the problems with valid conclusion, the mistakes were analyzed. As the correct solution for these problems was either left or right, the possible mistakes were now categorised as the 'wrong' option (left or right but incorrect) and the 'no valid conclusion (nvc)' option. We performed a multilevel analysis,

Table 3. MMnv mistakes in Experiment 1.

Condition	Estimate	Standard error	Z	p
4th grade, no notes	−0.34	0.18	−1.89	.058
6th grade, no notes	0.18	0.13	1.40	.163
4th grade, notes	0.56	0.28	2.01	.045
6th grade, notes	0.32	0.24	1.30	.193

predicting the type of mistake from grade, notes condition and their interaction, with individual participants as random intercept. There was no significant effect of notes ($\chi^2(1)=0.03$, $p = .870$), which provides no evidence for an alternative explanation, that children in the no notes condition generally answered 'no valid conclusion' more often.[5] There was, however, a significant effect of grade ($\chi^2(1)=10.02$, $p = .002$), with the younger children answering nvc more often. This could partially explain why the younger children scored better at the MMnv problems, viz. because they generally answered nvc more often, potentially because they were less confident about their answer.

Wright and Smailes (2015) argued for the importance of considering gender as an important factor in transitive reasoning. We exploratively performed a separate analysis[6], predicting score from notes, type and gender, with individual participants as random intercept. For the grade six data, there was a significant main effect of gender ($\chi^2(1)=4.52$, $p=.034$), as well as a significant three-way interaction between type, notes and gender ($\chi^2(2)=9.19$, $p=.010$). When notes were not allowed, girls were better at the problems with valid conclusion, while the boys benefitted more from the ability to take notes for the problems with a valid conclusion, as can be seen in Figure 7. For the grade four data, there was no significant main effect of gender ($\chi^2(1)=0.23$, $p=.635$), as can be seen in Figure 8. No interactions with gender were significant either. Apparently, this gender difference develops until after the age of nine. We will not speculate here about the nature of this difference.

Experiment 2

Experiment 1 demonstrated that the children showed a strong preference for only considering one possible state of affairs. When they could take notes, they improved at this strategy, even in the cases where it was the wrong strategy. A question that naturally arises from this result, is how we can stimulate them to consider multiple possibilities. And when we succeed at this, it would be interesting to see whether the children in the notes condition can also use their notes to their benefit for the MMnv problems.

[5]We thank an anonymous reviewer for this suggestion.

[6]Including 'gender' as a predictor in the main model proved to be too demanding to obtain a good fit.

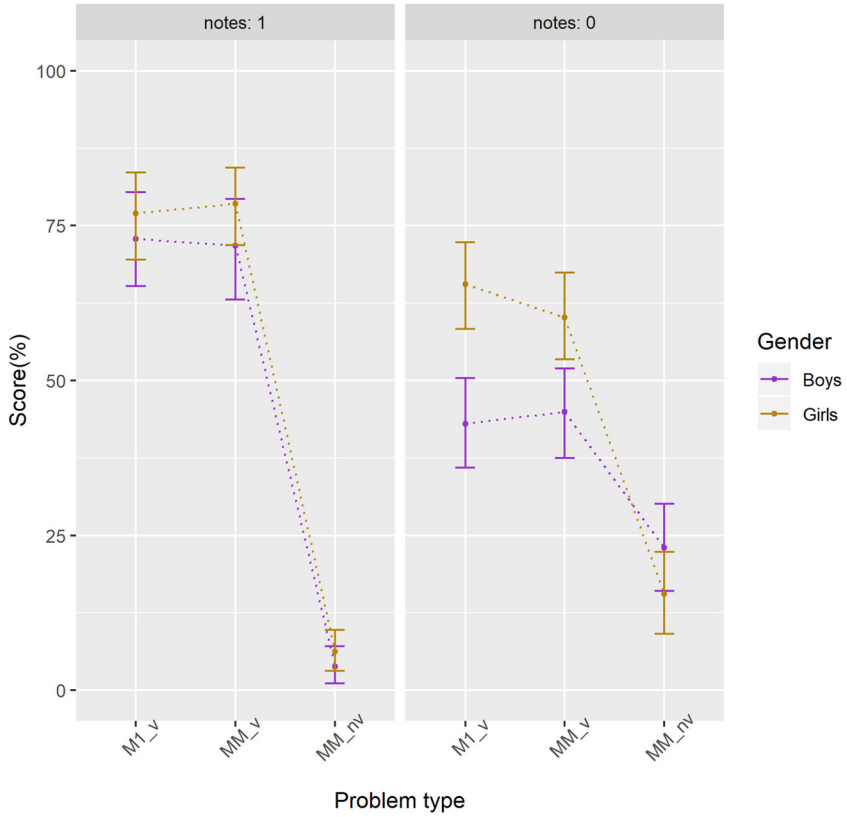

Figure 7. Experiment 1 mean scores in percentages for the sixth-graders by gender, notes group, and problem type.

In Experiment 2 we tried a very simple approach to reach this goal: we added a multiple possibilities example to the example that the instructor presented before the experiment. This was a relatively minimal adaptation, but now at least there was an example case in which it was shown that an object could be either left or right of another object given certain information, whereas the example in Experiment 1 only concerned a case with a definite answer.

Method

Participant info

We tested 132 children. There were 62 boys and 70 girls, 63 sixth-graders ($M = 11.42$ years, $SD = 0.25$) and 69 third-graders ($M = 8.29$ years, $SD = 0.25$). All data was again collected at schools in Flanders, different from those of Experiment 1. All children's data was anonymized before processing. The ethical approval and informed consent were identical to those of Experiment 1. Among the third-graders, 34 children were assigned to the control condition

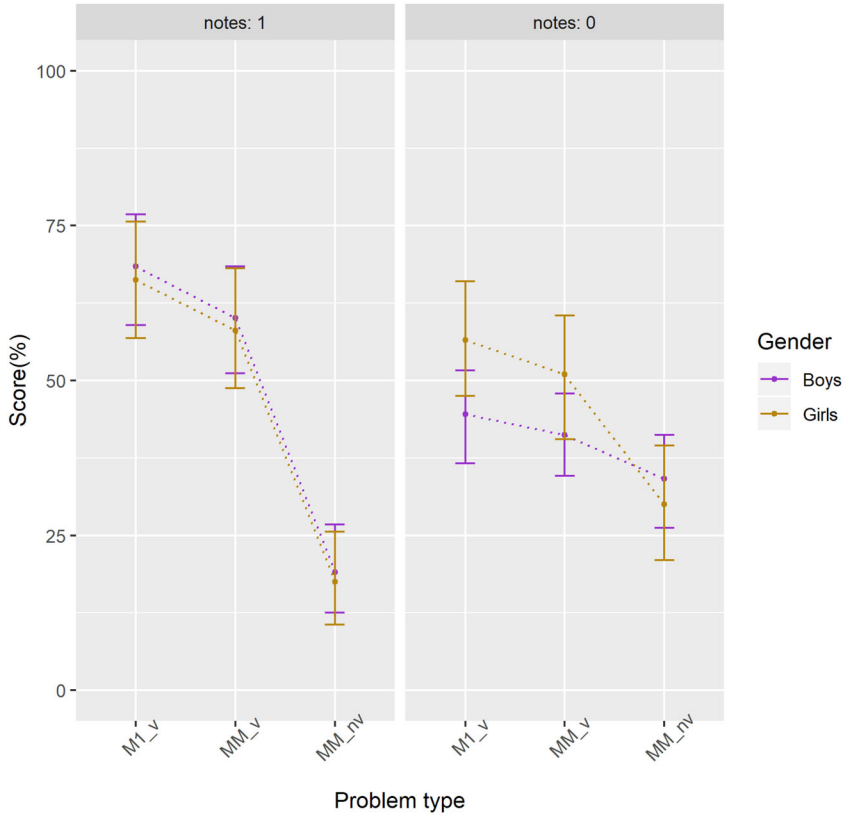

Figure 8. Experiment 1 mean scores in percentages for the fourth-graders by gender, notes group, and problem type.

and 35 to the experimental condition. Among the sixth-graders, 31 and 32 children were in the control and experimental condition, respectively.

Procedure and material

The procedure and material were identical to that of Experiment 1, with the already explained exception that the instructor now also showed the children an example with multiple possibilities.

Results and discussion

Effect of type, notes and grade on accuracy

The same analysis as in Experiment 1 was performed. The R^2 of the entire model was .43 (based on the theoretical variance) and .38 (based on the observation-level variance). If we only look at the fixed effects, the R^2 estimates were .35 and .31, respectively. Again, there was a significant three-way interaction between notes, grade, and problem type ($\chi^2(2) = 13.28$, $p = .001$, see Figure 9).

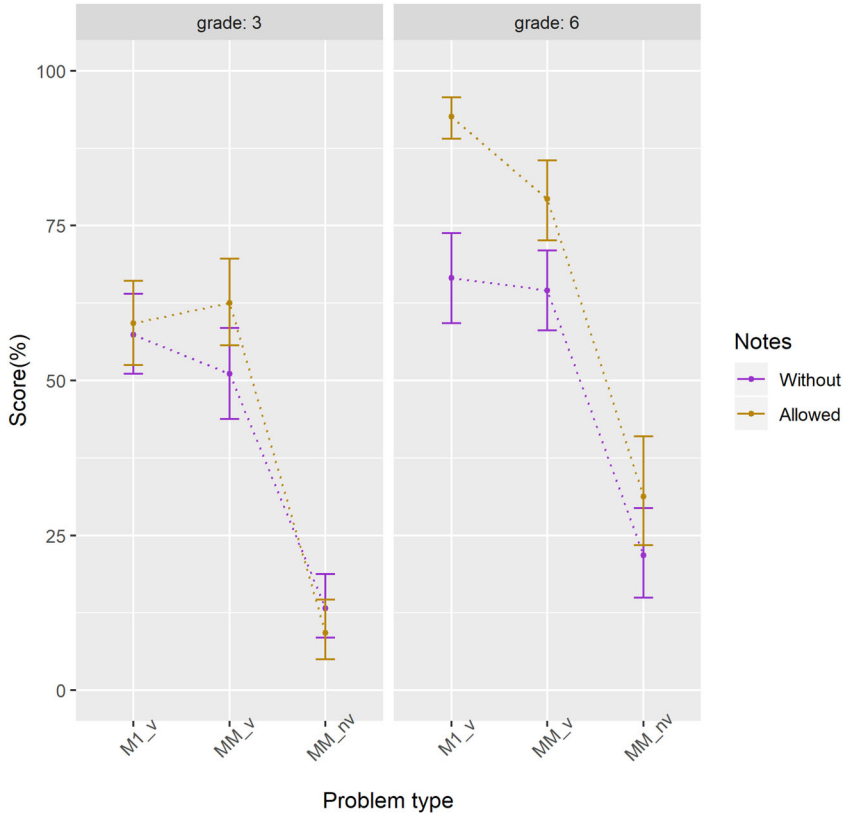

Figure 9. Experiment 2 mean scores in percentages by notes group, grade and problem type.

It does seem that the explicit mentioning of the cases with multiple possibilities had its effect. For Grade 6 participants taking notes resulted in significantly better results for the problems where one model was sufficient, as in Experiment 1. More important, however, is the effect of our manipulation on the MMnv problems. In Experiment 1 those taking notes performed significantly worse at the MMnv problems, while in the current experiment this was not the case. In fact, there was a slight benefit of taking notes for those problems as well now. For the Grade 3 participants, there was only a significant difference between the notes and no notes conditions for the MMv problems. See Tables 4 and 5 for more details on the effect of note taking. Showing children the multiple possibilities actually wiped out the benefits of notetaking in third-graders, presumably because the mechanism of taking notes to represent mental models was too difficult for them, requiring some level of meta-cognition that they do not yet master and therefore confusing them. Kennedy and Lodge (2016) showed that confusion can be associated with blockages or impasses in the learning process

Table 4. Effect of note taking for each problem type by grade combination in Experiment 2.

Problem type	Grade	Estimate	Standard error	Z	p
M1_v	6	1.95	0.33	5.92	< .001
MM_v	6	0.81	0.27	2.97	.003
MM_nv	6	0.54	0.27	1.98	.048
M1_v	3	0.10	0.24	0.42	.675
MM_v	3	0.52	0.24	2.15	.031
MM_nv	3	−0.47	0.32	−1.46	.144

(see also D'Mello et al., 2014). Of course, in our experiment, learning and testing happened at the same time. In the framework of desirable difficulties (see e.g., Bjork & Bjork, 2011), it has been shown that difficulties can reliably enhance learning in the long run. Overcoming misconceptions can precede the development of a more sophisticated understanding of the topic area. This might partly explain the difference between our youngest groups in Experiment 1 and 2. It might also be interesting to see what happens if the same participant would be retested some time later.

The cut-off for the highest scores at 16/24 for the notes group, that was observed in Experiment 1, now was less outspoken, with multiple children providing correct answers to the MMnv problems and achieving a higher total score, as can be observed in Figure 10.

More detailed analysis of the note taking

The note-taking behaviour in Experiment 2 showed rougly similar patterns to Experiment 1, but with notable differences. Notes were taken in 70% of the cases, with a mere 56% for the third-graders and 85% for the sixth-graders. The distribution of amount of notes taken can be seen in Figure 11. Of the notes taken, 85% were correct. Third-graders took correct notes in 76% of the cases, while sixth-graders did so in 91% of the cases. Of these correct notes, 98% were complete (third-graders 97% and sixth-graders 99%). Of the 747 occasions in which notes were actually taken for multiple model problems, now 76 (i.e. 10%, versus 0.49% in Experiment 1) showed information that we interpreted as representing multiple models. Again, the comparison between the cases in which notes were taken and those in which notes were not taken showed a similar pattern as when comparing the cases in which notes were allowed with those in which notes were not allowed, as can be observed in Figure 12. Analyses are again based on multilevel models with individual participants as random intercept. Whether the notes taken were correct or not significantly predicted the score for the problems ($\chi^2(1) = 51.53$, $p < .001$). The interaction with problem type was not significant this time ($\chi^2(2) = 5.62$, $p = .060$). Sixth-graders had significantly more correct notes than third-graders ($\chi^2(1) = 9.19$, $p = .002$), as pictured in Figure 13. Whether notes were taken or not did not

Table 5. Differential effect of note taking across grade for each problem type in Experiment 2.

Problem type	Estimate	Standard error	Z	p
M1_v	1.85	0.41	4.52	< .001
MM_v	0.29	0.36	0.79	.430
MM_nv	1.02	0.42	2.39	.017

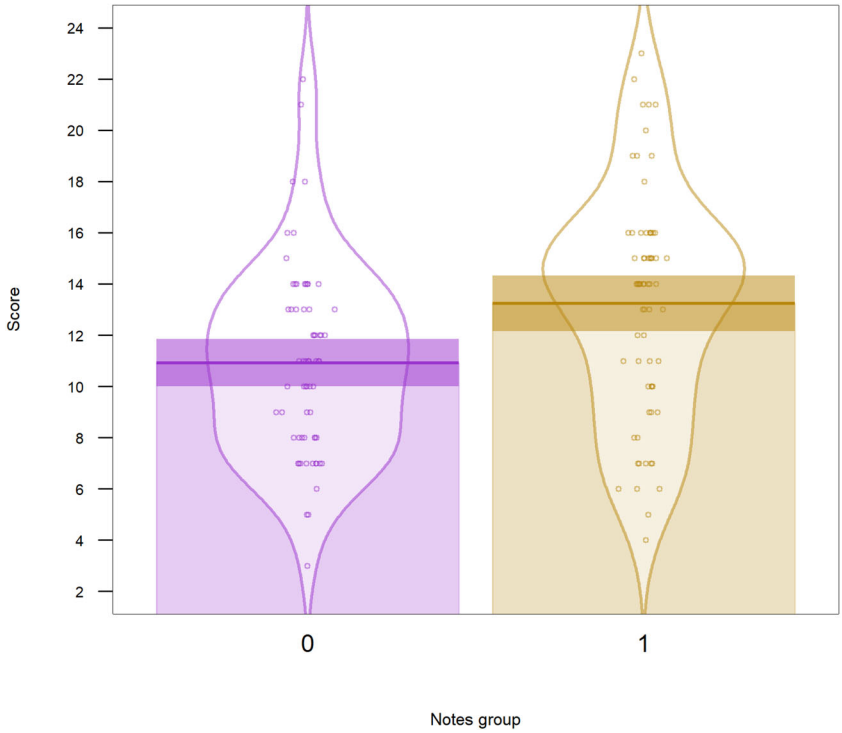

Figure 10. Total scores in Exp 2 by notes groups.

depend on problem type ($\chi^2(2) = 0.77$, $p = .681$), but did depend on problem order ($\chi^2(1) = 36.45$, $p < .001$), with less notes taken on later problems. Items were again mostly represented by words, followed by drawings, followed by a combination of both and exceptionally initial letters or symbols, as can be seen in Figure 14.

The improved scores for Experiment 2 are not just random noise. When combining the data of the sixth-graders from both experiments, we see a significant interaction of experiment with problem type ($\chi^2(2) = 9.14$, $p = .010$). The contrasts in Table 6 show that sixth-graders in Experiment 2 score significantly higher for each problem type, compared to those in Experiment 1.

The notes they made were very similar to those in Experiment 1: analogical representations of the spatially ordered items. Out of 67 participants

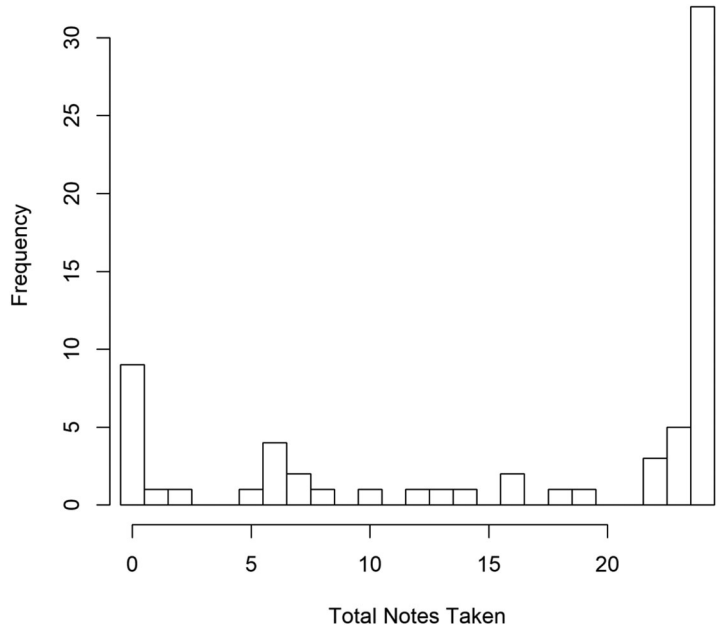

Figure 11. Frequency distribution of total amount of notes taken in experiment 2.

in the notes condition, nine drew representations of multiple possibilities. Remarkably, all but one did not draw multiple models but a single model in which they indicated in some way that there were multiple possibilities, for example by using arrows, arcs or representing an item twice. This way of representing multiple possibilities in a single model has been coined an 'isomeric model' in Schaeken et al. (2007). With so few children representing multiple possibilities, we cannot draw any conclusions here, but it may be worthwhile to investigate whether reasoners, and children specifically, prefer these isomeric representations over multiple models.

Analysis of type of mistakes

Looking at the MMnv mistakes, we see, in Table 7, that there were no significant preferences for either fff or ff mistakes, when taking into account the random effect of the individual participants.

As for Experiment 1, we analyzed the mistakes for the valid conclusion problems separately, with a multilevel analysis, predicting the type of mistake from grade, notes condition and their interaction, with individual participants as random intercept. There were no significant effects, neither of notes ($\chi^2(1)=2.90$, $p = .089$) nor grade ($\chi^2(1)=0.00$, $p = .951$), excluding an alternative explanation in terms of systematically more 'nvc' answers.

Finally, for Experiment 2, there was no significant main effect of gender, nor for the sixth-graders ($\chi^2(1)=0.15$, $p=.696$), nor for the third-graders

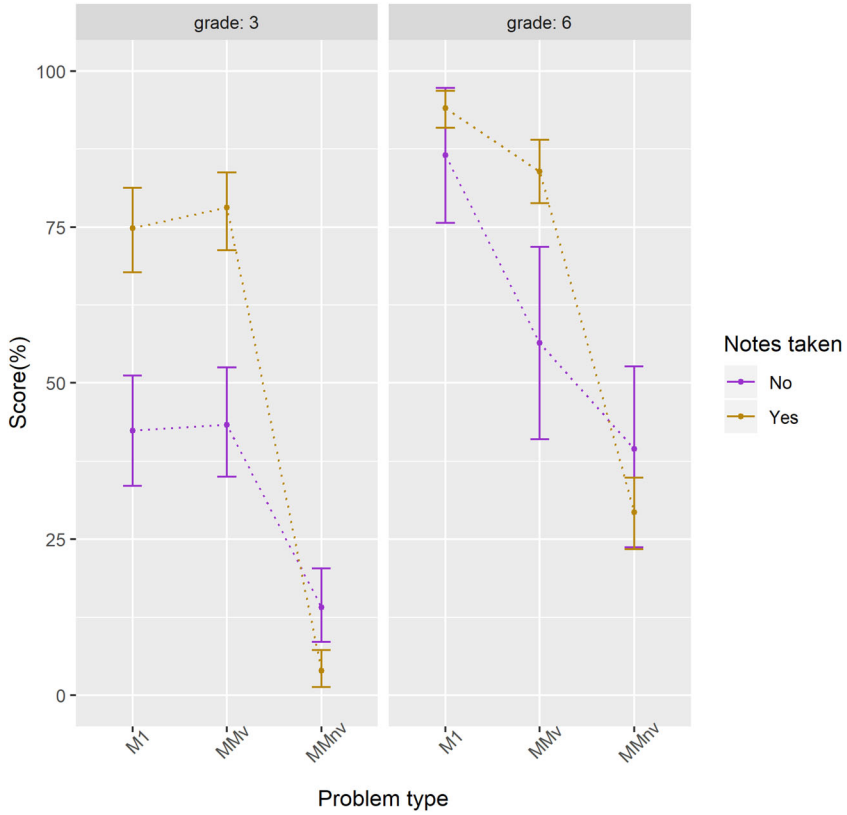

Figure 12. Experiment 2 mean scores in percentages for the notes group, by notes taken, grade and problem type.

($\chi^2(1)=0.02$, $p=.895$), as can be seen in Figures 15 and 16, respectively. There was, however, a significant interaction between type and gender for the third-graders ($\chi^2(2)=20.40$, $p=<.001$). The three-way interaction type-notes-gender for the third-graders ($\chi^2(2)=5.14$, $p=.076$) and for the sixth-graders ($\chi^2(2)=5.86$, $p=.053$) approached significance.

General discussion

The first research goal we defined, was to specify the effect of working memory on relational reasoning. For problems that could be solved by means of a single mental model, we found what was expected: alleviating the weight on working memory results in better performance. For problems where considering multiple models is required, the situation is more complicated. The commonly accepted explanation of why people do not vary upon their preferred mental model, is the principle of parsimony, which is explained in terms of limited working memory capacity (Goodwin &

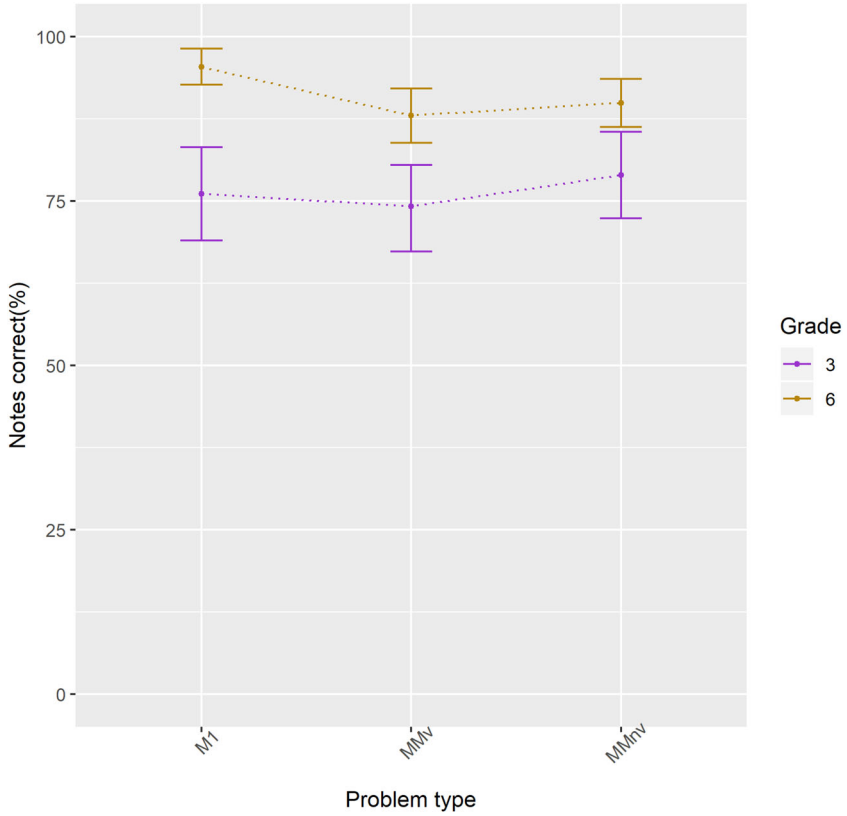

Figure 13. Experiment 2 correct notes in percentages for the notes group when notes taken, by grade and problem type.

Johnson-Laird, 2005; Jahn et al., 2007; Markovits & Barrouillet, 2002). In light of our results, it is worth reconsidering this explanation. By allowing our participants to use paper and pencil, we provided them with a mechanism to overcome the limitations of working memory. This had the expected result for the reasoning problems in which varying the preferred model was not required: the children who could take notes performed significantly better. The best explanation for this indeed seems to be in terms of working memory capacity. Now, if the reason why people do not vary their preferred model is limited working memory capacity, we would especially expect an improvement for those problems in which varying their preferred model is required. However, we saw a reversed effect (Experiment 1) or only a slight improvement (Experiment 2). This suggests that working memory capacity in itself cannot be the sole motivation why reasoners choose to refrain from constructing multiple models.

A first plausible alternative explanation is insufficient inhibitory control. Once one mental representation, consistent with the premises, is

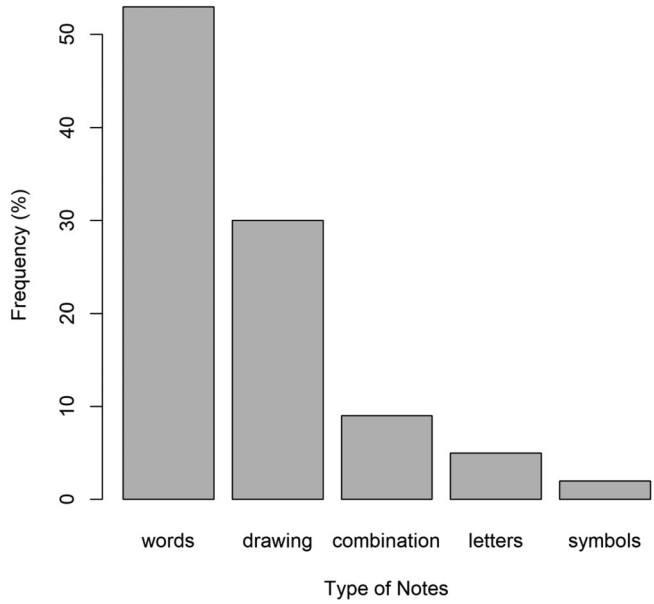

Figure 14. Type of notes taken in experiment 2.

Table 6. Differential effect of experiment for each problem type for sixth-graders.

Experiment	Problem type	Estimate	Standard error	Z	p
1 – 2	M1_v	−1.05	0.16	−6.52	< .001
1 – 2	MM_v	−0.62	0.15	−4.06	< .001
1 – 2	MM_nv	−1.16	0.17	−6.72	< .001

Table 7. MMnv mistakes in Experiment 2.

Condition	Estimate	Standard error	Z	p
3rd grade, no notes	0.06	0.15	0.39	.695
6th grade, no notes	−0.13	0.28	−0.45	.654
3rd grade, notes	−0.23	0.18	−1.25	.212
6th grade, notes	0.77	0.45	1.71	.086

constructed, inhibitory control is required to not halt calculations and instead look for further possibilities. An additional reason, then, why people do not vary their preferred mental model, is lack of inhibitory control. It could be that the requirements on inhibitory control are even tougher when this first mental model is written out on paper. The children could see a consistent representation of the information in front of them but still had to resist the temptation of thinking 'I found it' and answering the question. This explanation is in line with the effect described by Begolli and Richland (2016), that sequentially presented alternative representations can pose a high burden on the executive functioning system and require participants to inhibit attention effortfully. As working memory capacity and inhibitory control are correlated and both are good predictors of reasoning performance (Handley et al.,

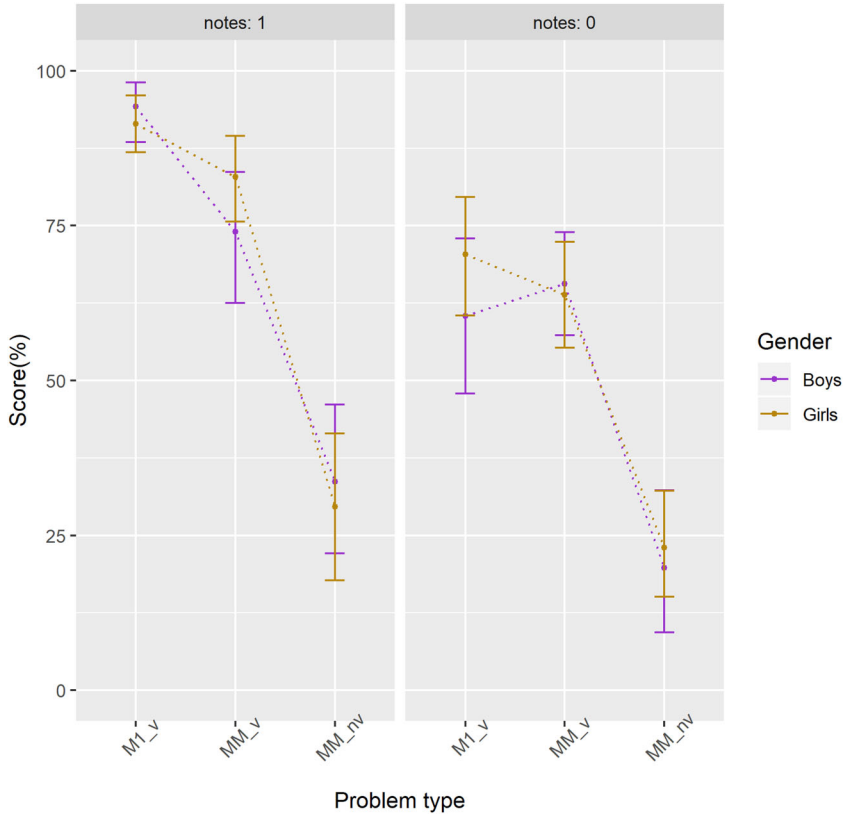

Figure 15. Experiment 2 mean scores in percentages for the sixth-graders by gender, notes group, and problem type.

2004), an explanation in terms of working memory capacity can easily seem plausible when in fact inhibitory control is also part of the explanation. With our manipulation, we may have managed to positively influence working memory capacity without positively influencing inhibitory control, which, in combination with the problem types, sheds more clarity on the role of each of these executive functions. Handley et al. (2004) suggested that inhibitory control was mostly important when dealing with prior knowledge, when beliefs need to be ignored. In our experiments this was not the case, but still inhibitory control seems essential for the specific reasoning aspect of varying preferred models. Its importance for reasoning performance could weigh in heavier than previously assumed.

A second alternative explanation, both compatible with an account in terms of working memory and in terms of inhibitory control, concerns rational thinking dispositions. According to Stanovich and West (1998), a thinking disposition is best seen as a cognitive style that is more malleable, with rational thinking dispositions being those related to reasoning. The

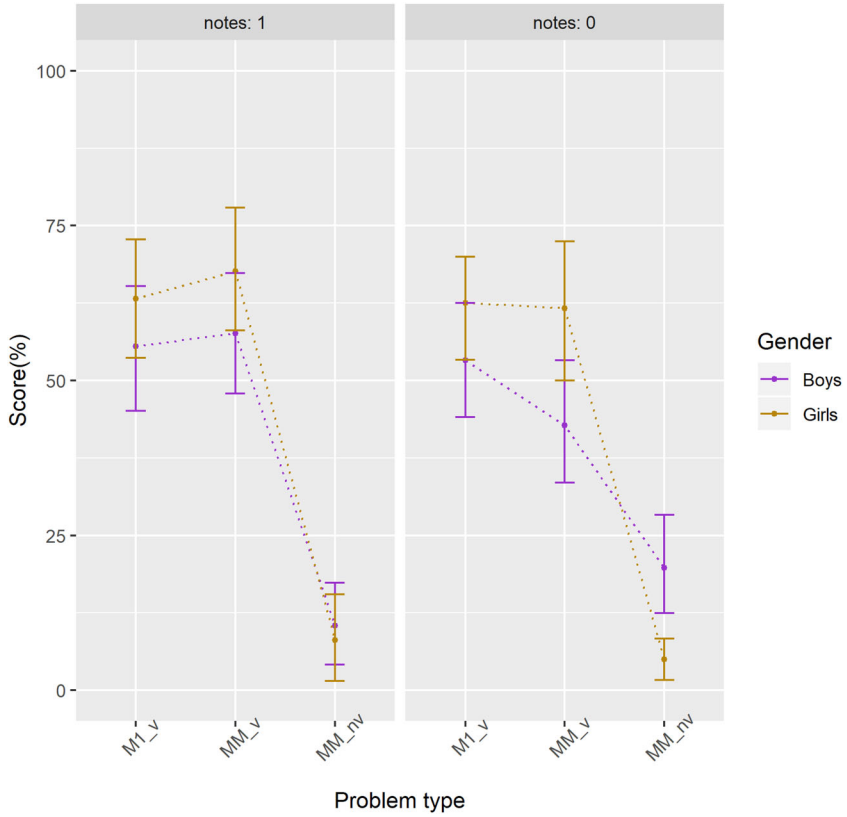

Figure 16. Experiment 2 mean scores in percentages for the third-graders by gender, notes group, and problem type.

advantage of an interpretation in terms of rational thinking dispositions is that it can also explain the difference between the results of Experiment 1 and 2. There is no reason to assume that the capacities of working memory and inhibitory control of the children in Experiment 1 were different than those in Experiment 2. But our slightly more elaborate instruction in Experiment 2, with the explicit mention of multiple possibilities, may have changed their propensity to consider alternatives. That is what is meant by the 'malleability' of thinking dispositions: "Although you cannot improve working memory by instruction, you can tell someone to spend more time on problems before she gives up, and if she is so inclined, she can do what you say" (Baron, 1985, p.15). Rational thinking dispositions come in many different flavours. 'Consideration of alternatives', mentioned by Markovits and Barrouillet (2002), seems particularly apt for capturing the effect we are after, with 'actively open-minded thinking' a more general candidate that could carry the load (Cf. Baron, 1985, 1993; Stanovich & West, 1998; Toplak et al., 2014).

A third alternative or complementary explanation could concern context monitoring (Cf. Chatham et al., 2012). When children are constructing only one model, rather than impulsively aborting further reasoning once one model is constructed, what they may be failing to do is appropriately monitor for cues that would tell them whether another model is possible. The instructions in Experiment 2 may have prompted such monitoring, by highlighting for children that they will need to distinguish between cases with and without valid conclusions.

In summary, differences in working memory load can explain the improved performance on M1 and MMv problems in the notes condition in both experiments, inhibitory control can explain the difficulties with the MMnv problem type and the rational thinking disposition responsible for consideration of alternatives and/or context monitoring can explain the improved results for the MMnv problem type in Experiment 2. Evidently, this interpretation is a line of thought that will require further research before strong claims can be made. More specifically, measuring these variables and seeing how well they predict the score on MMnv problems will be needed.

The second research goal, a more detailed understanding of reasoners' internal representation, can be answered with a much clearer picture. Our young participants spontaneously drew iconic representations of the spatially organised items. It seems plausible that their internal strategy for solving the reasoning problems is also based on such representations, rather than on a logical strategy involving, for example, understanding of the transitivity of the relation 'is to the left of'. Making abstraction of whether the iconic representations were drawn with images, words or letters, we can interpret them as mental models and interpret our results as supporting mental model theory.

Finally, our third research goal was obtaining a developmental perspective. It is clear that children get better at reasoning with one representation as they grow older. Less clear is whether their understanding of multiple models improves a lot at the ages we tested. In Experiment 1, the sixth-graders scored worse than the fourth-graders on the MMnv problems. The sixth-graders in Experiment 2 scored a bit better, but still only at chance level. So the extent to which they understand multiple possibilities is not that clear. One hypothesis is that producing multiple models themselves is still too challenging for them, although they may be capable of understanding the concept when it is explained to them. What *is* clear, is that our third-, fourth- and sixth-graders spontaneously use iconic representations and that a strong bias for ignoring multiple possibilities is present. In Experiment 1, only three out of 96 children drew multiple models. In comparison, Ragni et al. (2007) found that 10% of adult participants drew multiple models in a similar task.

We did not find a significant preference for the fff-strategy, which consists of adding an item to the model while leaving the relative positions of the already represented items unchanged, when taking into account the random effect of individual participants. It could be argued that such preference develops with age, as we found a maximal percentage of 63% fff-mistakes for the sixth-graders in the notes condition, whereas Ragni et al. (2007) found a 78% percentage for adult reasoners. Computationally modelling this as the default behaviour, as done in Ragni and Knauff (2013), may be overstating this effect, at least for modelling the behaviour of young reasoners.

A remarkable result was the (non-significant) preference for the ff-strategy among fourth-graders in the no notes condition in Experiment 1. Ragni et al. (2007) mention a possible linguistic explanation for such preference: 'to the left of' can be interpreted as '*immediately* to the left of'. This interpretation results in the ff-strategy when adding items to the model. Still, it remains unclear why only our fourth-graders in the control condition would prefer this interpretation. The first-free-fit strategy is based upon analogy with writing down items in cells and then looking for the first free cell. Changing items from cell is mentally costly. Although we did not provide our participants with cells, it does not come as a surprise that this strategy was adopted more in the notes condition than in the condition without notes. When writing down items, even on a blank space, each item takes up a certain amount of space. We could regard this as virtual cells. Writing a third item next to two existing items is easier than writing it in between them, especially when there is not much space left. The mental cost as postulated by the preferred mental model theory is as such mirrored to some extent in the cost to squeeze in a third item in between two items that were represented as being next to each other. Sometimes children needed to write such a third item very small or had to erase one of the other ones and write it elsewhere, an action that can be interpreted as changing the item's cell.

One limitation of this study is that we did not measure inhibitory control and open-minded thinking, which would have allowed to check for correlations with varying the preferred model and answering MMnv problems correctly. A second limitation was the third answering option. The third multiple-choice option was 'none of the above'. This is the correct answer if you interpret the question as 'what can you validly conclude from these premises?'. As a descriptive statement, however, it is incorrect: the answer is either left or right, even if one cannot know which is the case based on the information in the premises. In retrospect, this should have been stated more clearly. However, the children were clearly instructed that this was the option to choose if they thought there were multiple possibilities or if they had the feeling there was no correct answer. A third limitation is the

fact that no response times were collected. As the test was conducted on paper, it was not possible to do this per exercise, but maybe total times could at least have been registered. This would have been useful to compare the total time on task spent by the children in both conditions. Longer time spent on problems might be part of the reason why some children were more successful or simply more engaged with the problem. However, practical considerations prompted us to not record response times: measuring times with a chronometer might not be that reliable and switching to a computer version of our experiment might have made the note taking less natural. A fourth limitation was varying the age of the younger children. It would have been better to stick to fourth-graders for Experiment 2 as well, in order to be able to compare consistently with Experiment 1.

Conclusion

We ran two reasoning experiments in which school children of two different age groups each solved 24 reasoning exercises. The control groups just had to choose the correct conclusion from three different options, while the experimental groups additionally had some blank space where they could make helpful notes before choosing their answer. In the first experiment, we observed a strong tendency to construct only a single model, resulting in a much lower score for the multiple-model problems with no valid conclusion, for which taking into account multiple possibilities is required to arrive at the correct answer. Understanding these multiple possibilities proved to be rather challenging for our young participants, even for the oldest ones and with the help of notes and some concise explanation. Taking notes was useful to improve their single-model strategy, which explains why in the notes condition accuracy was higher for the M1 and MMv problems, but significantly *lower* for the MMnv problems. Likewise, sixth-graders had lower scores than fourth-graders on these problems, because they were *better* at applying the preferred model strategy. In an attempt to overcome this preferred model bias, we explicitly showed them a multiple-model example with no valid conclusion in the second experiment, thus subtly explaining the preferred model mistake. This had some beneficial effect, but understanding of the multiple-model cases was still surprisingly low.

Based on these results, we argued that the reason why people do not vary their preferred model, is not *only* because of working memory considerations. Taking notes should be a substantial help on the working memory front, but did not yield very beneficial results when it comes to varying the preferred model. So parsimonious use of working memory cannot be the only reason why participants tend to be satisfied with one mental model, even in cases where multiple ones are possible. With inhibitory control, an

open-minded thinking disposition and context monitoring, we suggested three other possible causes of variability for this trait. To what extent each of these four causes come into play, may well depend on the exact reasoning circumstances. Moreover, as the causes are likely to be correlated, disentangling their respective influences will require experimental care. Our results cast doubt on wether preferred model bias exists in young reasoners, suggest that producing representations of multiple possibilities is by no means evident until the age of twelve, and that the principle of parsimony for mental model construction in itself is not sufficient to explain it.

Acknowledgements

Earlier versions of this study were presented at the 'Thinking About the Possible' Summer School, Central European University, Budapest, 9-14 July 2018 and at the Potsdam Research Institute for Early Learning & Educational Action Conference, Potsdam, Germany, 4-5 October 2018. We thank the various researchers that came to discuss our ideas and provided inspiring comments and questions. We would like to offer special thanks to Tessa Wittock and Lore Mivis for collecting the data and to Stef Herregods for his assistance with data processing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Authors' contributions

W.S. and K.D. conceived of the presented idea. W.S. and K.D. developed the theory, conceived and planned the experiments. K.D. and T.H. analysed the data. K.D. took the lead in writing the manuscript. All authors provided critical feedback and helped shape the research, analysis and manuscript. T.H. was a Postdoctoral Fellow of the Research Foundation-Flanders (FWO-Vlaanderen) at the start of this project. K.D.'s funding came from the C1 project BOF (C14/17/043).

Funding

This article was supported by Bijzonder Onderzoeksfonds, Fonds Wetenschappelijk Onderzoek.

References

Ameel, E., Verschueren, N., & Schaeken, W. (2007). The relevance of selecting what's relevant: A dual process approach to transitive reasoning with spatial relations. *Thinking & Reasoning*, *13*(2), 164–187. http://www.tandfonline.com/doi/abs/10.1080/13546780600780671

Andrews, G., & Halford, G. S. (1998). Children's ability to make transitive inferences: The importance of premise integration and structural complexity. *Cognitive Development*, *13*(4), 479–513. https://doi.org/10.1016/S0885-2014(98)90004-1

Bara, B. G., Bucciarelli, M., & Johnson-Laird, P. N. (1995). Development of syllogistic reasoning. *The American Journal of Psychology*, *108*(2), 157–193.

Baron, J. (1985). *Rationality and intelligence*. Cambridge University press.

Baron, J. (1993). Why teach thinking?—An essay. *Applied Psychology: An International Review*, *42*(3), 191–214.

Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science*, *4*(6), 372–378. https://doi.org/10.1111/j.1467-9280.1993.tb00584.x

Begolli, K. N., & Richland, L. E. (2016). Teaching mathematics by comparison: Analog visibility as a double-edged sword. *Journal of Educational Psychology*, *108*(2), 194–213. https://doi.org/10.1037/edu0000056

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher R. W. Pew & J. R. Pomerantzf (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society*. (pp. 59–68). Worth Publishers.

Braine, M. D. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, *85*(1), 1–21.

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, *23*(3), 247–303. https://doi.org/10.1207/s15516709cog2303_1

Chatham, C. H., Claus, E. D., Kim, A., Curran, T., Banich, M. T., & Munakata, Y. (2012). Cognitive control reflects context monitoring, not motoric stopping, in response inhibition. *PloS One.*, *7*(2), e31546. https://doi.org/10.1371/journal.pone.0031546

D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, *29*, 153–170.

Fodor, J. A. (1975). *The language of thought*. Crowell.

Fodor, J. A. (1983). *Modularity of mind*. MIT Press.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, *40*(2), 177–190. https://doi.org/10.1037/0012-1649.40.2.177

Gilhooly, K., Logie, R., Wetherick, N., & Wynn, V. (1992). Working memory and strategies in syllogistic reasoning tasks. *International Journal of Psychology*, *27*(3-4), 148–148.

Goodwin, G. P., & Johnson-Laird, P. N. (2005). Reasoning about relations. *Psychological Review*, *112*(2), 468–493. https://doi.org/10.1037/0033-295X.112.2.468

Handley, S. J., Capon, A., Beveridge, M., Dennis, I., & Evans, J. S. B. (2004). Working memory, inhibitory control and the development of children's reasoning. *Thinking & Reasoning*, *10*(2), 175–195.

Inhelder, J., & Piaget, B. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures*. Routledge & Kegan Paul.

Jahn, G., Knauff, M., & Johnson-Laird, P. N. (2007). Preferred mental models in reasoning about spatial relations. *Memory & Cognition*, *35*(8), 2075–2087. https://doi.org/10.3758/bf03192939

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard university press.

Johnson-Laird, P. N. (2006). *How we reason*. Oxford University Press.

Kennedy, G., & Lodge, J. M. (2016). All roads lead to rome: Tracking students' affect as they overcome misconceptions. In A. P. S. Barker S. Dawson & C. Colvin (Eds.), *Show Me the Learning. Proceedings ASCILITE 2016.* (pp. 318–328). SA.

Klauer, K. C. (1997). Working memory involvement in propositional and spatial reasoning. *Thinking & Reasoning*, *3*(1), 9–47.

Knauff, M., & Johnson-Laird, P. N. (2002). Visual imagery can impede reasoning. *Memory & Cognition*, *30*(3), 363–371. https://doi.org/10.3758/bf03194937

Manktelow, K. (1999). *Reasoning and thinking*. Psychology Press.

Markovits, H., & Barrouillet, P. (2002). The development of conditional reasoning: A mental model account. *Developmental Review*, *22*(1), 5–36. https://doi.org/10.1006/drev.2000.0533

Markovits, H., Dumas, C., & Malfait, N. (1995). Understanding transitivity of a spatial relationship - a developmental analysis. *Journal of Experimental Child Psychology*, *59*(1), 124–141. https://doi.org/10.1006/jecp.1995.1005

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination r2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society, Interface*, *14*(134). http://search.proquest.com/docview/1938853354/

Pears, R., & Bryant, P. (1990). TRANSITIVE inferences by young-children about spatial position. *British Journal of Psychology*, *81*(4), 497–510. https://doi.org/10.1111/j.2044-8295.1990.tb02375.x

Phillips, N. (2017). *Yarrr: A companion to the e-book "yarrr!: The pirate's guide to r"*. https://CRAN.R-project.org/package=yarrr

Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review* , *120*(3), 561–588. https://doi.org/10.1037/a0032460

Ragni, M., Fangmeier, T., Webber, L., & Knauff, M. (2007). Preferred mental models: How and why they are so important in human reasoning with spatial relations. In Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) (Vol. 4387, pp. 175–190).

Rips, L. J. (1983). Cognitive processes in propositional reasoning. *Psychological Review*, *90*(1), 38–71.

Schaeken, W., Van Der Henst, J.-B., & Schroyens, W. (2007). The mental models theory of relational reasoning: Premises' relevance, conclusions' phrasing, and cognitive economy. In W. S. W. Schaeken A. Vandierendonck (Ed.), *The mental models theory of reasoning: Refinements and extensions* (pp. 129–149). Erlbaum.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*(2), 161–188. https://doi.org/10.1037/0096-3445.127.2.161

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Rational thinking and cognitive sophistication: Development, cognitive abilities, and thinking dispositions. *Developmental Psychology*, *50*(4), 1037–1048. https://doi.org/10.1037/a0034910

Van der Henst, J.-B., Yang, Y., & Johnson-Laird, P. N. (2002). Strategies in sentential reasoning. *Cognitive Science*, *26*(4), 425–468. https://doi.org/10.1207/s15516709cog2604_2

Wright, B. C. (2001). Reconceptualizing the transitive inference ability: A framework for existing and future research. *Developmental Review*, *21*(4), 375–422.

Wright, B. C., & Smailes, J. (2015). Factors and processes in children's transitive deductions. *Journal of Cognitive Psychology (Hove, England))*, *27*(8), 967–978. http://www.tandfonline.com/doi/abs/10.1080/20445911.2015.1063641 https://doi.org/10.1080/20445911.2015.1063641