What do External Representations Tell about Mental Models? An Exploratory Study in Deductive Reasoning

Barbara Bazzanella (B.Bazzanella@studenti.unitn.it)
University of Trento
LASDEC experimental laboratory, Venezia

Paolo Bouquet (Paolo.Bouquet@unitn.it)
Department of Information and Communication Technology, University of Trento

Massimo Warglien (warglien@unive.it)
Department of Business Economics, Ca' Foscari University of Venezia,
LASDEC experimental laboratory, Venezia

Abstract

We present an exploratory study in deductive reasoning based on the experimental elicitation of external representations of the premises as a tool to investigate how individuals build models of the premises and how they use them in reasoning. The goal of our research is not to check whether external, explicit models are useful as an heuristic support in deductive tasks, but to devise a methodology which allows to make as explicit as possible the mental models that people build when executing deductive tasks. We show that while the number and the completeness of models constructed is not significantly associated to deductive performance, the quality of models constructed does matter. Further analysis of external representations lends support to some predictions of the theory of mental models regard which models are more likely to be constructed. In addition, the qualitative analysis of data gave us some interesting information on typical errors patterns, which we clustered in three main categories: model editing, model integration, and modalization.

Introduction

The theory of mental models (Johnson-Laird, 1983), or TMM for short, postulates that individuals build "small scale" representations of reality and use them to reason, decide, or build expectations. Furthermore, such models are supposed to reflect the structure of what they represent. In the domain of deductive thinking, the TMM claims that both common successes and fallacies of human deductive performance can be explained on the ground of how individuals construct mental models, and of the limitations in their capacity to build such models (Johnson-Laird & Byrne, 1991; Johnson-Laird, Byrne, and Schaeken, 1992).

In essence, the TMM postulates that individuals facing a deductive reasoning task will build models of the premises and try to formulate a conclusion that is true in these models, testing the validity of their conclusions by trying to construct a counter-example. However, the individual capacity to generate models of the premises is limited (possibly due to working memory constraints), and individuals may generate incomplete representations of the premises, failing to perform appropriate inferences. Consequently, it is predicted that deductive tasks that involve a larger number of mental models should be more "difficult", i.e. trigger more frequently erroneous conclusions. Furthermore, not all models are equally likely to be generated. The TMM submits that individuals will try to economize working memory usage by constructing models of the premises that represent what is true, not what is false. Furthermore, whenever negative assertions are explicit in the premises, individuals do not represent falsity of the assertion, but instead they directly represent a negative assertion, which in turn can be true or false.

Since mental models are internal representations which are not directly accessible to observers, most research within the TMM tradition has been based on indirect experimental methodologies, that observe only the conclusions drawn by subjects on the ground of given premises, and compare rates of deductive success with the number of models needed to draw the correct inference. To our knowledge, within the TMM research field, only a few experiments have tried to elicit explicit, material representations of the premises from subjects to investigate their effects on deductive performance. Furthermore, the elicitation of external representations has been used to study the heuristic valence of external representations rather than to test the assumptions of the TMM.

In this paper, we explore the experimental elicitation of external representations of the premises as a tool to investigate how individuals build models of the premises and how they use them in reasoning. While we are aware that explicit, "external" representations may differ from mental models, we assume that the two levels of representation are not entirely unrelated. In particular, we hypothesize that representations which are harder to construct as mental models are also harder to generate as external representations. As a consequence, we expect that difficulties in generating the external models needed to draw correct inferences should be reflected in failures of the deductive performance of subjects. We are also aware that external representations may act as heuristic facilitators of

reasoning. However, since the reasoning tasks we explore are rather homogenous, we expect that such facilitation effects, whenever present, should not subvert the relative difficulty of inferences (as confirmed by a control experiment reported below). We will address a few research questions that pertain to central assumptions of the TMM. 1) Does the quantity of models constructed affect deductive performance? 2) Does the quality of models constructed affect deductive performance? 3) Which types of models are easier to generate? Furthermore, our experimental methodology will allow us to investigate additional aspects of the process of model editing.

Experimental design

Unlike other past work with external model (Zhang and Norman, 1994; Kirsh and Maglio, 1996; Bauer and Johnson-Laird, 1993; Bucciarelli and Johnson-Laird, 1999), the goal of our research is not to check whether external, explicit models are useful as an heuristic support in deductive tasks, but to devise a methodology which allows us to make as explicit as possible the mental models people build in their mind when executing deductive tasks.

To this end, we wanted to design an experimental setting in which subject are provided with materials which do not limit their freedom to build models, and do not suggest/imply shortcuts towards the correct solution(s). After considering various options, we decided to focus on a simple scenario, where problems concern features of a human face (*e.g.* the color of hair, or the presence of a beard), and the material consists of a set of physical objects representing the empty shape of a face and different types of hair, eyes, mouth etc. (Figure 1).

Models can be built by assembling objects from such a set. Intuitively, given a premise like "Stefano has black hair", we expected our subjects to build a model from the empty face and an instance of black hair (it is important to notice that each element, e.g. black hair, was provided in several instances). Since a premise is typically consistent with multiple models, subjects were explicitly told that they can build more than one model for each premise.

There were two test conditions: a baseline treatment, in which the new methodology is used; and a control treatment, where the standard methodology was adopted.

In the **baseline treatment**, there were two phases (see http://dit.unitn.it/~bouquet/mental-models for more details on the experimental procedure).

First, subjects were presented with a first set of premises, and asked to construct all possible models of such premises. After, subjects were presented a new set of premises, and asked to update the previously built set of models (or to build new models) to take into account the new premise(s). Finally, subjects were asked to say what follows from the premises. Subjects were told that they would be rewarded in proportion to correct models assembled in the representation of the premises, and for each correct answer.

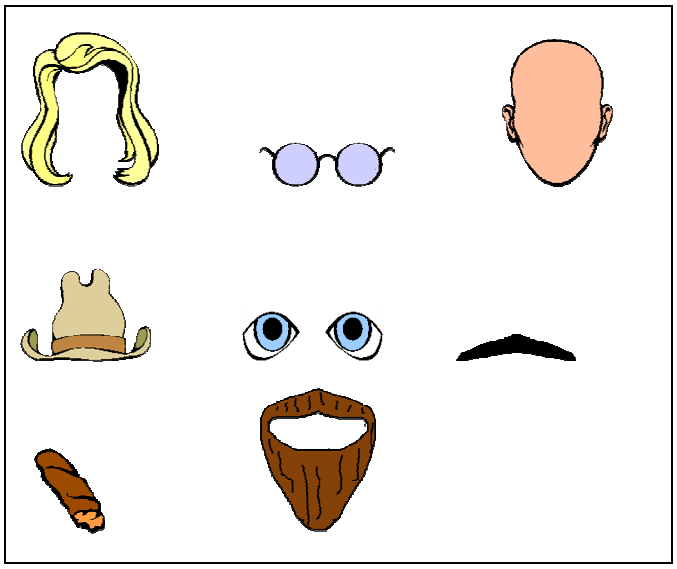

Figure 1: An example of visual materials for assembling external representations

The procedure was repeated for the five problems reported in Figure 2, where we wrote in bold the conclusions and in italics the premises presented after the models of the first premise(s) had been represented by the subject. Each subject performed the task individually, in a quiet room, and each experimental session was entirely video-recorded.

1) If Laura has red hair, then she wears glasses.
Laura doesn't wear glasses.

Laura doesn't have red hair.

2) Either Stefano doesn't have black hair or else he wears a hat - but not both.
Stefano has black hair.

Stefano wears a hat.

3) Antonio wears a hat or he has moustaches - or both.
Antonio doesn't wear a hat.

Antonio has moustaches.

4) If Mary wears glasses then she has green eyes.
Mary wears glasses or else she has green eyes – but not both.

Mary doesn't wear glasses and she has green eyes.

5) Either Giorgio has blue eyes or else he smokes a cigar – but not both.
If Giorgio has brown hair then he has blue eyes.
Giorgio has blue eyes.

Giorgio doesn't smoke a cigar.

Figure 2 The five problems of the experiment.

In the **control treatment**, subjects were presented the same five problems and asked to draw a conclusion without having to construct any external representation. Individuals performed the task individually in a quiet room, but received no monetary reward.

Results

As a first step, we controlled for the effects of our experimental procedure on subjects' inferential success. We checked whether constructing external representations would significantly alter the difficulty of each task as compared to classical experimental treatments with no external representations elicitation. Furthermore, we aimed at controlling whether the baseline treatment would subvert the relative difficulty of inferences, i.e. the difficulty ranking of the five different premise sets. This second test is especially relevant, since the TMM gives a central importance to the difficulty ranking of problems as a source of empirical validation of the theory. If the baseline treatment would alter the relative difficulty of the tasks, its informativeness on the TMM would be hardly defendable. Our expectation was that on the one hand there should be a facilitating effect (due to short term memory constraints mitigation), but on the other hand there should be no change in the relative difficulty of the tasks, given the homogeneity of the tasks themselves.

Table 1 reports the main results of the baseline and the control treatments. While there seems to be an overall facilitation effect, it turns out to be only a partial one. Only tasks 1, 3 and 4 display significant facilitation effects, while task 5 exhibits a reverse effect (although at a weaker significance level). What is more important for us, such facilitation effects do not alter the difficulty ranking of the tasks, which is exactly the same in both treatments (3,1,2,5,4).

Table 1: Correct inferences by treatment.

	Baseline Treatment (N=41†)		Control Treatment (N=44)		signif. (2 pro-portions test)
	Freq.	%	Freq.	%	
1	33	82,5	24	54,54	0.05
2	21	51,21	22	50	
3	37	90,24	28	63,63	0.05
4	8	19,51	2	4,54	0.05
5	9	21,95	18	40,9	0.10

† In task 1, baseline treatment, only 40 subjects are considered due to the ambiguity of an otherwise correct answer, that was dropped)

It is therefore legitimate to look inside the response behavior of subjects in the baseline treatment. The first question is whether there is any association between "external" models constructed by subjects and their success in drawing correct conclusions. If there is any relation between the difficulty to generate mental models of the premises and the difficulty to construct their explicit representation, this should be reflected in the association between external representations and inferential conclusions. We consider two types of representations constructed by subjects: "critical representations" (at least those models needed for drawing the correct inference are represented), and complete representations (all models of the premises have been constructed). For example, in the *ModusTollens* problem (task 1), the model representing "Laura doesn't have red hair and doesn't wear glasses" is sufficient, in conjunction with the second premise, to get the right answer (and thus is the critical representation). The classical three models of the material conditional constitute the complete representation.

Tables 2 and 3 summarize our experimental evidence. Of the five tasks, one have an extreme outcome : task 3 is correctly performed by almost all subjects. Consequently, this task provides little useful statistical information. On the other hand, the remaining four tasks provide intriguing evidence.

Table 2: Critical representations and answers by task.

	Critical representation		No critical Representation		p (Fisher test, 1-sided)
	correct answer	wrong answer	correct answer	wrong answer	
task 1	15	1	16	7	0.061
task 2	14	2	6	18	0.0001
task 3	36	4	1	0	1.00
task 4	1	5	7	28	0.754
task 5	0	5	29	7	0.061

In three such tasks, there is a significant association between critical representations and correct answers (although often significance is weak, at the 10% level).

However, no significant association can be found for complete representations, with the exception of task 2. Furthermore, no correlation was found in any one task between the number of models and correct answers. While this appears to provide (partial) support to the hypothesis that semantic factors affect the deductive performance, the results appear in striking contrast with the conventional TMM emphasis on semantic incompleteness as source of error, and on the numerousness of models as the main explanatory factor of deductive performance. Instead, the accent is on the ability to generate the "appropriate" models

– something that might be only loosely related to computational bottlenecks, and point instead at the process of "editing" a model (see the discussion in the next section). The weakness of a "computational bottleneck" explanation is also demonstrated by the fact that in some tasks often individuals do not construct too few models, but instead construct too many of them. For example in the case 4 nine subjects out of thirtythree build up too many models but drew a erroneous conclusion.

Table 3: Complete representations and answers by task.

	compl. repr.		incomp. repr.		p
	correct answer	wrong answer	correct answer	wrong answer	(Fisher test, 1-sided)
task 1	7	1	25	7	0.487
task 2	13	1	7	19	0.00007
task 3	28	4	9	0	1.00
task 4	1	5	7	28	0.356
task 5	9	25	0	7	0.150

Obviously, one may plausibly argue that in the TMM the numerosity of models is only a proxy to the more subtle process of incomplete model construction. Our methodology allows to capture additional insights into the process of model construction. Some of these appear in clear agreement with the main TMM assumptions.

The TMM predicts that models which reflect what is explicitly asserted in the premises and contain no negation are the most likely to be constructed. Our data clearly support such an hypothesis. In four tasks out of five (1,3,4,5), there are such models (Table 4, type I). In all four cases, they are significantly more frequent than any other model ($p<.05$).

Another interesting regularity is that models which are negations of such explicit, assertive models typically come second (Table 4, type II). In task 1 and 4 there are models which are just the negation of the explicit assertive ones above, and they are significantly ($p<.05$) more frequent than any other model, with the exception of course of the latter.

Up to here we reported a quantitative analysis in order to test the main assumptions of TMM. The aim of this second part of the analysis is to examine some qualitative evidences which our method reveals about the nature of internal representations when individuals not trained in logic carry out deductive problems. In others words, we suggest that external representations can help us to understand why people make systematic errors in deductive reasoning tasks.

In this examination, we extensively used the data from video recording, as they allow us to follow the steps of the representation process. Interestingly enough, we found a regular error pattern which recurs in the control treatment as well; this provides a preliminary evidence that our experimental methodology does not alter in a substantial way the nature of (some) errors.

We identified three main typologies of failures. Table 5 reports the frequency of the four kinds of error in our experiment.

Table 4: Model types and their frequency

Type	Task	Model	Freq.	Signif.
I	1	Laura with red eyes and glasses	40	0.01
I	3	Antonio with hat and moustaches	40	0.05
I	4	Maria with glasses and green eyes	36	0.01
I	5	Giorgio with brown hair and blue eyes	41	0.01
II	1	Laura with no red hair and no glasses	18	0.05
II	4	Mary with no glasses and no green eyes	21	0.01

Significance according to two proportion tests with each other model in the task (type I) or with each other model except type I models (type II)

Model editing. It seems that subjects consider each model of the premises not as an atomic entity, but as a molecular entity which in turn consists of atomic sub-elements. As a consequence, the elimination of atomic elements from a model does not lead necessarily to the elimination of the entire model, but only to its modification. That's why what we call *model editing* this type of error.

We report a paradigmatic example from task 1 (*Modus Tollens*), illustrated in Figure 3. Of the subjects who drew an invalid inference, three out of nine erroneously concluded: "Laura has red hair". If we look at what they did, we discover that they fleshed out only one model of the first premise (a face with red hair and glasses, top left of Figure 3). When presented with the second premise, they removed the glasses from the picture without eliminating the model. After this step of model editing, only one possible "conclusion" was available, namely that "Laura has red hair".

Integration failure. In many cases, we noted that subjects have troubles in integrating the different models of premises. Indeed, they start with constructing the models of each premise, but then seem unable to see have all these models can be integrated, and therefore draw conclusions like: "nothing follows", "there is a contradiction in the premises", and so on. We found a nice example of integration failure in task 4, in which five subjects of the baseline and three in the control treatment concluded that

there are two different Mary's. Our explanation is that subjects could not create a coherent picture in which all premises are true, and therefore used this "trick" to exit the *impasse*. This problem was addressed also in (Bouquet and Warglien, 1999), where the idea lf local mental models was introduced to explain some deductive failures due to a lack of models integration.

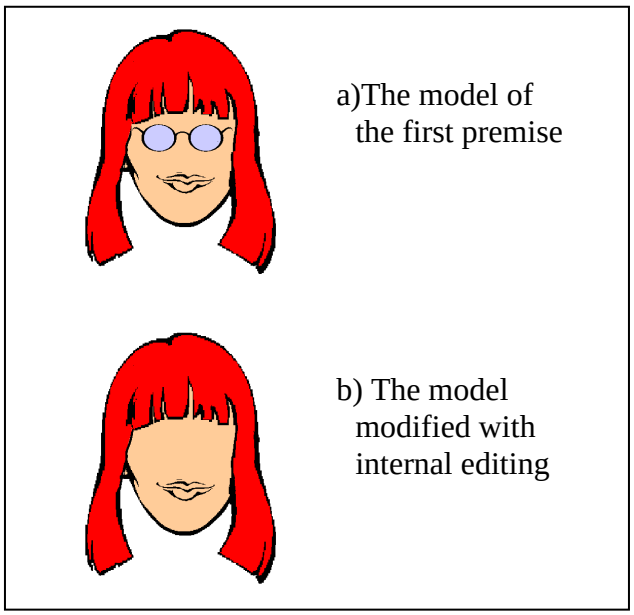

Figure 3: an example of model editing

Modalization. A third kind of error frequently happens with negative premises, like "Either Stefano doesn't have black hair". With such a premise, subjects must find a representation strategy which is different from the positive case. One possible solution would be to build a new model for every available hair color different from black (in our experiment, four hair colors were available); this would be a correct representation, but only under a closed world assumption (namely, that the available colors are all possible colors). However, subjects didn't not adopt this strategy, and in many cases decided to represent the fact that Stefano does not have black hair by creating a model in which Stefano has brown hair, or has no hair at all. Once such a possible model is built, some subjects get stuck in this representation, and reason as if Stefano had brown hair, a fact that is definitely not guaranteed by the premise that Stefano does not have black hair.

We call this error a *modalization error* as, from a logical point of view, it shows a confusion between satisfiability (truth in at least one possible model) with validity (truth in all possible models). We classify under this type of error also conclusions formulated with expressions like "it may be that ….", " X or Y" , "it is possible that" and so on.

We are aware that this strategy in representing negation is quite different compared with those reported in others studies (see Mayo, Schul and Burnstein, 2004 for more details about a distinction between "the fusion model" and "the schema-plus-tag model"), but it is important to stress that instances of the same error were found also in the control treatment. This corroborates the hypothesis that this form of error is not artificially produced by our methodology; the use of external representations simply makes available a possible explanation.

Discussion and Conclusions

In this paper we presented the preliminary results of an experiment on deductive reasoning we carried out with a new methodology. The methodology is based on the use of external representations, which should provide some insight on what models people build to solve a problem, and how they manipulate them (*e.g.* to accommodate new premises). A first conclusion is that the proposed methodology seems to provide a valuable source of data, as the comparison with a control experiment on the same test set shows that the relative difficulty of tasks is preserved. What is more interesting is that the new methodology seems to allow researchers to reach a finer granularity in the analysis of data, and to "observe" reasoning strategies which are not apparent in the traditional experimental setting. Examples are the editing of partial models, the representation of negative information, the lack of a coherent integration between models of different premises.

Our future work will aim at investigating these qualitative aspects of external representations, as they seem to provide good explanations of many errors observed in the standard setting.

Table 5: error frequencies

Task	Type error	Baseline Treatment	Control Treatment
1	Editing	3	5
	Integration	4	5
	Modalization	1	8
2	Integration	8	6
	Modalization	5	3
3	Editing	1	-
	Integration	1	-
	Modalization	-	11
4	Editing	4	
	Integration	15	18
5	Integration	19	10
	Modalization		1

References

Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science*, *4*(6), 372-378.

Bouquet, P., & Warglien, M. (1999). Mental models and local models semantics: the problem of information integration. *Proceedings of the European Conference on Cognitive Science (*ECCS'99), (pp. 169-178). University of Siena, Italy.

Cox, R., & Brna, P. (1993). Analytical reasoning with external representations. In R. Cox, M. Petre, P. Brna & J. Lee, Eds. *Proceedings of the AI-ED93 Workshop on Graphical Representations, Reasoning and Communication.* Pp. 33-36- August, Edinburgh.

Johnson-Laird, P.N. (1983). Mental models: *Towards a cognitive science of language, inference and consciousness*. Cambridge, England: Cambridge University Press.

Johnson-Laird, P.N., & Byrne, R. M. J. (1991). *Deduction.* Hillsdale, NJ : Lawrence Erlbaum Associates.

Johnson-Laird, P.N., & Byrne, R. M. J., Schaeken W. (1992). Propositional reasoning by model. *Psychological Review*, 3, 418-439.

Bucciarelli, M., & Johnson-Laird, P.N.(1999). Strategies in syllogistic reasoning. *Cognitive Science*, 23, 247-303.

Kirsh, D., and Maglio, P. (1994). On distinguish epistemic from pragmatic action. *Cognitive Science*, *18*, 513-549.

Mayo, R., Schul, Y., and Burnstein E. (2004). "I am guilty" vs. "I am innocent": Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology.* 40(4), 443-449.

Norman, D. (1993). Cognition in the head and in the world. *Cognitive Science*, 17,1-6.

Radvansky, G. A., Spieler, D. H., & Zacks, R. T. (1993). Mental model organization. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 19, 95-114.

Scaife. M., & Rogers, Y. (1996). External cognition: how do graphical representations work? *International Journal of. Human-ComputerStudies*, 45, 185-213.

Zhang. J., & Norman, D.A. (1994). Representations in distributed cognitive tasks. *Cognitive Science*, 18, 87-122.

Zhang. J. (1997). The nature of external representations in problem solving. *Cognitive Science*, 21, 179-217.