Journal of Experimental Psychology:
Learning, Memory, and Cognition
2000, Vol. 26, No. 1, 187–203

0278-7393/00/$5.00 DOI: 10.1037//0278-7393.26.1.187

On the Relation Between Representations Constructed From Text Comprehension and Transitive Inference Production

José Favrel and Pierre Barrouillet
Université de Bourgogne

Deductive inference production from texts is a process considered to involve either the construction of an integrated mental model or the step-by-step coordination of propositional representations of the sentences. These alternative hypotheses were tested in 3 experiments using a set inclusion task paradigm in which participants had to recall the premises and to evaluate transitive inferences. Contrary to what is known about linear ordering relations, order of recalls and reaction times provide evidence that the encoding of set inclusion relations does not result in an integrated representation. These results suggest that the mental models theory needs to take account of the nature of the relation to be represented if it is to become a general theory of reasoning.

Deductive reasoning is considered to be a process that involves the manipulation and transformation of mental representations that result from comprehension processes (Johnson-Laird, 1993; Rips, 1994). Recent theories have primarily focused on the nature of these processes of transformation. Are we in the presence of rules that are isomorphic to those of formal logic (Braine, 1990; Braine & O'Brien, 1991; Rips, 1983, 1994) or are we dealing with the manipulation of mental models (Barrouillet & Lecas, 1998; Bonatti, 1994a, 1994b; Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991; Johnson-Laird, Byrne, & Schaeken, 1994)?

The mental models theory supposes that the production of deductive inferences does not extend beyond the scope of the comprehension processes (Johnson-Laird & Byrne, 1991). In effect, these processes would result in the construction of an internal model structurally isomorphic to the situation the entire set of premises describes. Inference production would thus consist of the simple extraction of information from a single representation constructed in working memory.

In contrast, for the advocates of mental logic (Braine, 1990; Rips, 1994), the comprehension processes should lead to the construction of propositional representations of each of the available premises. Deductive reasoning would thus consist of the implementation of inference rules. These rules would be activated by the presence in working memory of propositions (e.g., *a or b* and *not a*) and would produce a conclusion (i.e., *b*). This conclusion, associated with another premise, might in its turn serve as an input to an inference rule (Braine, 1990; Rips, 1994). Reasoning that implies multiple premises would therefore necessitate the sequential implementation of inference rules and the production of intermediate conclusions.

Thus, for both theories logical reasoning depends on general comprehension processes, the nature of which remains controversial. This controversy is not peculiar to the psychology of reasoning but is also present in the field of text comprehension studies with the opposition between constructionists and minimalists models. The former hold that, as reading progresses, important information is integrated into a situation model that is very similar to a mental model (e.g., Kintsch's Construction–Integration model, 1988, 1995). The latter considers that readers would not automatically produce inferences in order to obtain a complete representation of the situation described but only inferences that ensure the local coherence of the text and inferences that result from knowledge available in LTM (McKoon & Ratcliff, 1992). Thus, the structure of the representations resulting from comprehension processes is a central problem for text comprehension theories as well as for the psychology of reasoning, because it reveals the nature of the inferences made on-line during reading and constrains the processes involved in subsequent logical inferences production.

A number of recent studies have approached this problem in the field of the psychology of reasoning through the study of propositional reasoning based on multiple premises (Braine et al., 1995; O'Brien, Braine, & Yang, 1994). For example, Braine et al. (1995) asked participants to judge the conclusion *P?* on the basis of the premises *X or E; not X; if E then L; not both L and P.* Participants' production of intermediate inferences, the nature of these inferences, and their order of production suggest that the participants do not construct an integrated representation (i.e., a mental model) of all the available information that might give them direct access to the conclusion that is to be judged. Braine et al. interpreted this fact as evidence against the mental models theory.

The problem raised by O'Brien et al. (1994) and by Braine et al. (1995) regarding the organization of information in memory and its effect on the production of inferences in multiple premise problems was already been discussed

José Favrel and Pierre Barrouillet, LEAD–CNRS Université de Bourgogne, Dijon, France.

José Favrel is now at the Faculté de Psychologie et des Sciences de l'Education cou (FPSE), University de Geneve, Geneva, Switzerland.

Correspondence concerning this article should be addressed to Pierre Barrouillet, LEAD–CNRS, Faculté des Sciences Gabriel, 6 Bld Gabriel, 21000 Dijon, France. Electronic mail may be sent to barouil@satie.u-bourgogne.fr.

during the 1970s and 1980s with reference to set inclusion and linear ordering tasks. These two types of task have the same structure: Participants have to judge transitive inferences on the basis of premises containing inclusion relations (*All As are Bs, All Bs are Cs, All Cs are Ds*) in the case of the set inclusion task, and ordering relationships (e.g., *A is larger than B, B is larger than C, C is larger than D*) in the case of the linear ordering tasks. It is now generally agreed that participants solve the linear ordering tasks by constructing an integrated representation of the entirety of the premises (i.e., a linear array like A–B–C–D) on the basis of which the inferences are thought to be "read" directly (Evans, Newstead, & Byrne, 1993). Today, these results are thought to discount the hypothesis that information is encoded in the form of a set of propositions that reflect the linguistic structure of the premises (Garnham, 1996, p. 41).

In contrast, the problem of the representation that underlies performances in the set inclusion task remains unresolved (Evans et al., 1993). A number of authors have suggested that the transitive inferences permitted by the inclusion relation are produced on the basis of an integrated representation of the premises (e.g., a linear array of the form A–B–C–D; Potts, 1976, 1978), which is close to a mental model, whereas others have refuted this hypothesis (Griggs, 1976; Griggs & Osterman, 1980; Griggs & Warner, 1982). It is therefore not impossible that the resolution of the linear ordering tasks, on the one hand, and the set inclusion task, on the other, is based on different representations and processes, even though they have an identical structure. These differences could account for the fact that participants have better performance in the linear ordering task than in the set inclusion task.

The aim of this article is to study the organization of the representations from which deductions are produced and the possible effect of this organization of information on the activities of deductive inference production. We present three experiments whose aim is to test the hypothesis that participants construct an integrated mental model of the entirety of the information in order to solve the set inclusion task. Even though it has given rise to only a small number of recent studies (see, however, Barrouillet, 1996; Carlson, Lundy, & Yaure, 1992; N'Guyen & Revlin, 1993), the set inclusion task possesses numerous advantages compared with propositional logic problems for the study of the organization of information in memory. It makes it possible to measure the relative difficulty of the inferences in terms of the number of premises they require. Unlike the multiple premise problems, which generally require various inference schemas, it requires only one type of inference of the form *All As are Bs* and *All Bs are Cs;* therefore *All As are Cs,* which is known to be very easily produced by participants (Dickstein, 1978).

The Set Inclusion Task

The set inclusion task was first used by Frase (1969), who asked participants to learn the following text:

> The Fundalas are outcast from other tribes in Central Ugala. It is the custom in this country to get rid of certain types of people. The hill people of central Ugala are farmers. The upper highlands provide excellent soil for cultivation. The farmers of this country are peace loving, which is reflected in their art work. The outcasts of central Ugala are all hill people. There are about fifteen different tribes in this area.

This text describes a series of hierarchical inclusion relations between the classes *Fundalas, Outcasts, Hill people, Farmers* and *Peace loving*. In the following, these five classes are designated by the letters A, B, C, D, and E, and the relation *All As are Cs* is expressed using the terms A and C in the order of the relation, AC. The inclusion relation has two properties. It is transitive; that is, the propositions *All As are Bs* and *All Bs are Cs* make it possible to deduce that *All As are Cs.* It is also antisymmetrical; that is, the proposition *All As are Bs* does not make it possible to deduce with certitude that *All Bs are As.*

Frase (1969) observed two phenomena: (a) The participants tended to consider the inclusion relation to be symmetrical and frequently accepted false propositions as true (i.e., they inferred *All Bs are As* from the premise *All As are Bs*), and (b) the acceptance level for propositions (both true and false) decreased as the number of inferential steps increased (step-size effect), resulting in an interaction between the number of inferential steps and truth value. The increase in the step size caused a fall in the level of correct responses for true propositions but an increase in this level for false propositions (Truth Value × Step Size interaction). The participants thus exhibited a symmetrical (inversion of the relation) and nontransitive (rejection of inferences) conception of the relation.

Many experimental replications of this paradigm (Carrol & Kammann, 1977; Griggs, 1976; Griggs & Osterman, 1980; Griggs & Warner, 1982; Mynatt & Smith, 1979; Newstead & Griggs, 1984; Newstead, Keeble, & Manktelow, 1985; Potts, 1976, 1978) have confirmed these two facts while revealing major individual differences (Griggs & Osterman, 1980; Mynatt & Smith, 1979).

The Explanatory Hypotheses

Potts (1976) suggested that participants construct an ordered linear representation of the terms (A–B–C–D–E) and use this as a basis for the evaluation of the presented conclusion (e.g., *AD* true?). Misled by their real-world knowledge, they process the inclusion relation as a relation of similarity. Because the relation of similarity is not strictly transitive, two distant terms within such a representation would be judged to be less similar than two adjacent terms, which would account for the step-size effect. The symmetry of the relation of similarity would account for the tendency to invert the relations and accept false adjacent propositions. This hypothesis is close to that which holds that participants construct a mental model (Johnson-Laird, 1983).

Griggs (1976; Griggs & Osterman, 1980; Griggs & Warner, 1982) proposed an alternative hypothesis that holds that most participants do not store the terms in a linear array but instead store the propositions presented in the text. The result pattern observed in the set inclusion task would then be due to the fact that participants (a) make conversion errors (*AB* implies *BA*) and (b) might be wary of drawing inferences from chains of universal affirmative propositions (cautiousness hypothesis), thus resulting in the Truth Value × Step Size interaction. Participants would thus be all the more reluctant to produce inferences as the step size increases. However, Griggs and Warner (1982) identified a number of important individual differences.

While retaining Griggs's (1976) idea that participants store the propositions presented in the text, Barrouillet (1996) proposed a slightly divergent hypothesis. The interaction might be due to a problem of cognitive load (cognitive load hypothesis) while individual differences might arise from differences in processing capacity. According to the cognitive load hypothesis, inferences are calculated by means of the step-by-step coordination of the premises (e.g., for AE, AB–BC ⇒ AC, AC–CD ⇒ AD, and AD–DE ⇒ AE) within working memory that is of limited capacity. Thus the increase in the step size (i.e., the number of premises to be coordinated and intermediate inferences to be produced) would result in an increase in the cognitive load associated with the calculation of a proposition. This would explain why the rejection level increases with step size for both true and false propositions.

In Barrouillet (1996), the participants' working memory capacity, evaluated with the Alphabet recoding task and Daneman and Carpenter's (1980) reading span, proved to be a good predictor of the scores in the reasoning task ($r = .41$, $n = 72$). An analysis of the pattern of correlations between the working memory tasks and the set inclusion task revealed that the reading span, which has a known relationship with reading comprehension performance (Daneman & Carpenter, 1980, 1983), was a good predictor of the tendency to reject the symmetry of the inclusion relation ($r = .31$, $n = 72$). High-span participants accepted false adjacent propositions less frequently. In contrast, the alphabet recoding score was linked to variations in performance as a function of step size in the reasoning task ($r = .31$, $n = 72$). The participants who obtained better scores were less sensitive to increases in the step size. The partial correlations indicated that alphabet recoding did not predict the tendency to accept the symmetry of the relation and that the reading span was independent of sensitivity to step size.

Thus solving the set inclusion task seems to involve two types of distinct and relatively independent processes. One, associated with the comprehension of the text and the organization of the information it contains, would be responsible for the tendency to reject the symmetry of the relation, whereas the other would have the role of calculating inferences. The effectiveness of these two processes would depend on working memory capacity.

However, the cognitive load hypothesis leaves two questions unresolved. The first concerns the reasons why the reading span should be principally related to the tendency to reject the symmetry of the relation. It might be supposed that the presence of extensive reading comprehension capacities would make it possible to construct an integrated mental model of the information contained in the text (Oakhill, 1996). This mental model might take the form of the linear array suggested by Potts (1976). However, such a representation should also facilitate the production of inferences, and reading span should then be highly correlated with sensitivity to the increase in step size. This was not observed. Thus the reason why reading span is first and foremost a predictor of the tendency to reject false adjacent propositions remains to be explained.

The second question concerns the link between difficulties in inference production and cognitive load. Many models of reasoning suggest that the production of inferences is demanding (Braine, 1990; Johnson-Laird & Byrne, 1991; Rips, 1983, 1994). If we consider the set inclusion task, similar predictions derive from both the cognitive load hypothesis and Potts' (1976) model. In effect, both models suggest that a proposition will be rejected all the more frequently, the more distant its terms are in the inclusive chain. What is more, Potts' model is compatible with the fact that the participants possessing the greatest cognitive resources also achieve the best performances in the set inclusion task because the construction of the linear array could depend on the participant's cognitive resources.

The Present Experiments

The two questions left unanswered by the cognitive load hypothesis thus point to the same problem: the memory organization of the information present in the premises and the effect of this organization on inference production. The purpose of the three experiments presented below was to determine (a) how participants organize information in memory when solving the set inclusion task, (b) whether an interindividual variability is observable, and (c) the effect of these various types of information organization on inference production. Experiment 1 used a classic set inclusion task that was either preceded or followed by a task requiring participants to recall the premises presented in the text. If some or all of the participants encode information in a linear array (Potts, 1976), information recall should tend to respect the logical order of the inclusion relation (i.e., AB–BC–CD–DE) even when premises are not presented in the logical order. Griggs's (1976) hypothesis and the cognitive load hypothesis predict that recall will respect the order in which the premises appear in the text. Experiment 2 compared recall from set inclusion texts with recall from linear ordering relation texts (e.g., *John is taller than Mark*), which are known to induce the construction of a linear array (Evans et al., 1993). These latter texts should therefore induce logically ordered recalls. Griggs's hypothesis and the cognitive load hypothesis predict more frequent logically ordered recalls from linear ordering texts than from set inclusion texts. Experiment 3 studied the modification of the reaction times (RTs) as a function of step size in the set inclusion task. If participants store the terms in a linear array, the RTs should fall as the step size increases as is observed in the processing of linear ordering relations (Potts, 1976). If participants store atomic propositions and use a step-by-step strategy to calculate inferences (Barrouillet, 1996; Griggs & Osterman, 1980), the RTs should increase as step size grows.

Experiment 1

The aim of this experiment was to determine how participants organize the information present in the text (e.g., the *Fundala* text) with a view to the subsequent production of inferences. It is quite likely that this organization differs depending on the participants' aims. As part of a preexperiment, a group of 32 first-year psychology students were asked to memorize four set inclusion texts in which the order of appearance of the premises did not follow the logical order (e.g., BC, DE, AB, CD). The participants were then asked to recall the premises but not to perform any evaluation or inference production task. Under these conditions, the premises were overwhelmingly recalled in the order in which they appeared in the texts. Of the fully correct recalls of the four premises (67% of recalls), only 6% reestablished the logical order (i.e., AB–BC–CD–DE), whereas 71% reflected the order of presentation in the text. In addition, the number of logically ordered recalls did not increase as successive texts were presented even though all of them had the same structure. This last point indicates that the repetition of the task did not modify the strategies mobilized by the participants.

At the very least, these results indicate that the structure of the set inclusion texts is not in itself sufficient to induce participants to reorganize the information spontaneously. It is possible that this structure is not perceived. If this is not the case, then participants may not make use of it either because they are unable to reorganize the information and therefore store it in the presented order or because the constraints of the recall task do not make this reorganization necessary. Learning in the light of the future use of the information should make it possible to distinguish between these two possibilities. Therefore, the participants of the current experiment were told that they would have to perform two tasks, namely a recall task and a task requiring them to evaluate the inferences that could be produced using the information presented in the text.

Furthermore, we presented the recall task either before or after the proposition evaluation task. Indeed, Kintsch (1986) reported that children's recall of arithmetical word problems differs depending on whether this recall is performed before or after resolution. Recalls prior to problem solving conformed to the surface structure of the story problem text. In most cases they were correct, independently of the difficulty of the problem and the participants' problem-solving performance. In contrast, recalls following resolution reflected what the children had understood of the text. They depended on the difficulty of the problems and were correlated with the problem-solving performance. According to Kintsch, "before" recalls reflect the structure of the propositional representation of the text (text base), whereas "after" recalls are mediated by the mental model constructed by the participant and are therefore linked to problem-solving performance.

As far as the set inclusion task is concerned, the linear array postulated by Potts (1976) is akin to the mental models described by Johnson-Laird (1983) or to the situation model postulated by Kintsch (1988; Kintsch & Welsch, 1991). The construction of a mental model that integrates all the information should result in (a) the recall of information in the logical order, at least when recall follows the evaluation task (see Kintsch, 1986) and (b) a small step-size effect during the evaluation task, because the conclusions are directly accessible in such a representation. In contrast, both Griggs's (1976) hypothesis and the cognitive load hypothesis (Barrouillet, 1996) suggest that the premises are stored atomically in LTM and are then retrieved and coordinated in transient representations constructed in working memory in order to permit the production of inferences. These hypotheses predict that recalls will respect the order in which the premises appeared in the text, irrespectively of whether these recalls take place before or after the evaluation task. Because the production of inferences depends primarily on participants' working memory capacities, the step-size effect should be largely independent of the prior organization of the atomic propositions stored in LTM.

Method

Participants. Sixty-four undergraduate students of the Université de Bourgogne took part in the experiment. They were randomly distributed into two groups depending on the recall conditions (before and after the reasoning task).

Material. The texts presented to the participants referred to four different contents (iron bars, fictitious tribes, pullovers, and cars, see Appendix A) and presented a hierarchical inclusion relation between five classes using propositions of the type *All As are Bs* (AB), *All Bs are Cs* (BC), *All Cs are Ds* (CD), and *All Ds are Es* (DE). In the experimental texts the four propositions were not presented in the logical order of inclusion. We used four permutations to avoid the immediate succession of a term in two consecutive propositions (i.e., AB, CD, BC, DE; BC, DE, AB, CD; CD, AB, DE, BC; or, finally, DE, BC, AB, CD). Each permutation was applied to each content. In addition, for each content, the assignment of terms (e.g., for the contents "iron bars": *black, hollow, bent, long,* and *damaged*) to the five classes A, B, C, D, and E followed one of two orders. For half of the participants, the terms *black, hollow, bent, long,* and *damaged* corresponded to the classes A, B, C, D, and E respectively, whereas for the other half the same terms corresponded to the classes E, D, C, B, and A respectively. Thirty-two experimental texts were therefore obtained by combining the contents (4), the permutations of propositions (4), and the order of the terms (2). We inserted additional material between the propositions in order to make the texts more realistic.

The order of presentation of the contents, permutations, and orders of terms were counterbalanced across the participants. Each participant studied four texts (one per content), each representing a different permutation.

After learning each of the texts, each participant was presented with a notebook containing 10 propositions corresponding to the 10 logical conclusions permitted by the premises (i.e., AB, BC, CD, DE, AC, BD, CE, AD, BE, and AE). However, 5 of them were inverted (e.g., CB or EA) and represented invalid conclusions. Taken across the four studied texts, each type of conclusion was presented to each participant twice in its valid form (e.g., AB) and twice in its invalid form (e.g., BA). Each participant therefore evaluated 40 conclusions, 20 of which were valid and 20 invalid. The order of presentation of the 10 conclusions in the notebooks was random.

Procedure. We instructed groups of 8 participants to learn the information contained in the text presented to them in order to be able to judge the validity of conclusions relating to the texts that we presented subsequently. The experimenter emphasized the importance of learning the texts correctly and that participants could do this in their own time. When the participants considered that they had learned the information, the text was removed and they performed first the recall task and then the evaluation task (recall-before condition) or the evaluation task followed by the recall task (recall-after condition). In the recall task, the participants had to state in writing "all and only that information which, in the text, took the form *All the... are....*" In the evaluation task they received the following instructions:

> You must judge the logical validity of the conclusions presented in this notebook. If the proposition was present in the text or can be logically deduced from it, reply "True." In contrast, if the proposition was not present in the text and cannot be logically deduced from it, reply "False." You must judge the propositions in the order in which they are presented in the notebook without going back.

In both experimental conditions, we removed all the material used in the first task before starting the second task.

Results

Recall task. The mean level of correct responses for the recall-before and recall-after conditions combined was .824. We performed a 2 (conditions) × 4 (premises: AB, BC, CD, and DE) × 4 (rank order of presentation of the texts: from 1 to 4) ANOVA (analysis of variance) with repeated measures on the last two factors on the rate of correct recalls. The rate of correct recall increased with the rank of presentation and varied with the type of premise, but the recall-before (.846) and recall-after (.803) conditions did not differ significantly, $F(1, 62) < 1$, and there was no significant interaction. In brief, the quality of recall was not affected by the time at which it took place (before or after) with reference to the evaluation task.

We predicted that recalls should respect the order of the premises in the texts rather than the logical order. Out of a total of 256 recalls, 41 (16%) respected the logical order (i.e., AB, BC, CD, DE), whereas 109 followed the order in which they appeared in the texts, and 106 respected neither of these orders. We calculated the frequency with which the logical order was reestablished by establishing a ratio between the number of logical links reestablished during recall (e.g. AB–BC, BC–CD, or CD–DE) and the number of links that could be reestablished (varying from 1 to 3 depending on whether the participants recalled 2, 3, or 4 premises). As the atomic storage hypothesis predicted, this frequency did not differ for the recall-before (.31) and the recall-after (.28) conditions, $F(1, 62) < 1$. Thus the time of recall, before or after the task, had no effect on either the quality or the organization of recalls. The frequency with which the logical order was reestablished tended to vary with the rank order (.244, .326, .267, and .344 for ranks 1, 2, 3, and 4 respectively), but this was not significant, $F(3, 186) = 1.85, p > .05, MSE = 0.095$.

In sum, the participants organized the information more efficiently than in the preexperiment in which the memorization of the texts was not motivated by any future need to use the information. This change of strategy suggests that certain participants perceived the need to organize the information in order to evaluate the inferences, thus justifying the selected paradigm. However, this organization strategy was difficult to implement as the stagnation of the reestablishment level between ranks 2 and 4 testifies. These results suggest that, in the majority of cases, the participants stored the premises atomically and that, unlike the behavior observed by Kintsch (1986) in connection with arithmetical problems, performing calculations on the premises (recall-after condition) did not lead to their integration into a complete representation.

However, we observed a high level of interindividual variability in the level of reestablishment of the logical order. It was possible to distinguish between three groups of participants: (a) a group of 19 participants who did not reestablish the logical order (no reestablishment, 10 participants in the recall-before condition, and 9 in the recall-after condition), (b) a group of 31 who reestablished it only infrequently (between one and four reestablishments out of all the four recalls with an average of at most one reestablishment per recall, 14 participants in the recall-before and 17 in the recall-after condition), and (c) a group of 14 frequent reestablishers; (FR; five reestablishments or more, 8 in the recall-before and 6 in the recall-after condition). If the recalls of the frequent reestablisher participants reflect at least the partial integration of a logically ordered representation, these participants should achieve better performances than the others in the evaluation task.

Relations between the order of recall and performance in the evaluation task. We performed a 2 (experimental condition: recall before or after) × 3 (rate of reestablishment of the logical order) × 2 (truth value: true or false) × 4 (step size: from 1 to 4) ANOVA with repeated measures on the last two factors for the level of correct responses in the evaluation task. In conformity with the literature, the step size interacted with the truth value, $F(3, 174) = 35.25, p < .001, MSE = 0.174$. As the step size increased, the rate of correct responses decreased for true propositions (.929, .778, .700, and .618 for 1, 2, 3, and 4 inferential steps, respectively) and increased for false propositions (.506, .563, .659, and .697, respectively). The performances in the recall-before (.681) and recall-after (.699) conditions did not differ significantly, $F(1, 58) = 1.10, p > .05, MSE = 0.661$. In contrast, the participants who frequently reestablished the logical order made more correct evaluations (.813) than those who reestablished this order infrequently (.653) or never (.618), $F(2, 58) = 8.46, p < .01, MSE = 0.661$. However, the level of reestablishment did not interact either with the experimental condition, $F(2, 58) = 1.46, p > .05, MSE = 0.661$, or with the Truth Value × Step Size interaction, $F(3, 174) < 1, MSE = .174$.

Thus all the participants, irrespectively of the way they organized the premises during encoding, were equally sensitive to the step-size effect even though the participants who frequently reestablished the logical order obtained better results than the others.

In order to account for these better results, we calculated two performance quality indexes for each participant: an index for the rejection of the symmetry of the relation (Sym.

Index) and a transitivity index (Trans. Index). Because the participants exhibited a strong tendency to accept adjacent propositions independently of their truth value, we simply calculated the Sym. Index by counting the number of correct responses for false adjacent propositions (i.e. "false," scores between 0 and 8). A high level of correct responses for false adjacent propositions (i.e., rejection) reveals a tendency to reject the symmetry of the relation. The Trans. Index evaluated participants' ability to make use of the transitivity of the relation by calculating the value given by the interaction between the truth value and the linear trend observed for the step size for each participant. This equates to adding the inverse effect of the step size on true propositions (reduction in the correct response level) and false propositions (increase in this level). The value of the Trans. Index was high when the increase in the step size resulted in a sharp decrease in the number of correct responses for true propositions and a strong increase in this level for false propositions. The Trans. Index is therefore an indicator of the effect of step size and the difficulty of assuming the transitivity of the relation.

The number of reestablishments performed by the participants during recall was correlated with the Sym. Index, $r = .493, p < .001$, as well as with the Trans. Index, although to a lesser extent, $r = -.273, p = .029$. What is more, these correlations remained significant even when the quality of recall (number of correctly recalled premises) was partialed out: $r = .441, p < .001$, for the Sym. Index and $r = -.256$, $p = .043$, for the Trans. Index. Thus, the greater the extent to which the participants reestablished the logical order of the premises during recall, the greater their tendency to reject the symmetry of the relation (positive correlation with the Sym. Index) and make use of its transitivity (negative correlation with the Trans. Index). However, the tendency to reestablish the logical order during recall was more strongly associated with the correct identification of false adjacent propositions (Sym. Index) than with the use of transitivity (Trans. Index), suggesting that the better performances achieved by the frequent reestablishers were primarily due to the more successful rejection of false propositions.

Discussion

The results of this experiment have revealed three points. First, when participants are motivated to learn the premises through the need to use them in the future (reasoning task), logical order recalls are more frequent (16%) than when learning has no particular purpose (6% in the preexperiment), but they are still infrequent. Second, the calculation of inferences does not appear to result in the reorganization of information in memory, because there is no difference between recalls in the recall-before and recall-after conditions. Finally, the reestablishment of the logical order of the premises results in better performance in the evaluation task, primarily by blocking the acceptance of the symmetry of the relation. These three facts are entirely compatible with the cognitive load hypothesis.

The construction of an integrated mental model should result in logically ordered recalls and small step-size effect. The cognitive load hypothesis predicts that recalls will respect the order in which the premises appeared in the text and a step-size effect largely independent of the prior organization of the atomic propositions stored in LTM.

The rarity of logically ordered recalls, whether before or after the evaluation task, is more compatible with the cognitive load hypothesis than with the hypothesis of the construction of an integrated representation. However, even though the levels for the reestablishment of the logical order in the recalls primarily predicted the rejection of symmetry during the evaluation task ($r = .493$), it was not independent of the step-size effect ($r = -.273$). It therefore remains possible that participants who reestablish the logical order and who achieve the best performance in the evaluation task have integrated the premises into the ordered, integrated representation postulated by Potts (1978). This point was the object of our Experiment 3.

Moreover, Experiment 1 simply allows us to propose that the effects suggested by the hypothesis of the integration of premises within a logically ordered representation are not observed: The majority of participants did not reestablish the transitive order of the inclusion relations in their recalls, and the production of inferences did not increase the level of reestablishment. These results do not provide a decisive confirmation of the hypothesis of the atomic storage of premises, because they might also be due to the fact that the order of recall is not a sufficiently sensitive index of the organization of information in memory. A better test would be to compare the recall of premises from set inclusion texts with the recall of premises from linear ordering texts that are known to induce the construction of an integrated and logically ordered representation.

Experiment 2

The data from the literature suggests that linear ordering relations (e.g., larger than) are integrated at an early stage in a linear, logically ordered representation (Evans et al., 1993; Johnson-Laird & Byrne, 1991). We suggested that the order of the recalls provides us with a relevant index of the organization of the premises in memory and that set inclusion texts did not induce an integrated representation. This interpretation can therefore be strengthened if we can confirm that linear ordering relations give rise to logically ordered recalls, unlike the inclusion relations whose recall rarely respects the logical order (see Experiment 1).

We presented two texts containing set inclusion relations and two texts containing linear ordering relations to participants and asked them to perform a premise recall task followed by a conclusion assessment task. Two participant groups were formed as a function of the order of presentation of the texts: either set inclusion followed by linear ordering or the contrary. If the order of recall of the premises is a reliable index of their organization in memory, participants' recalls should reestablish the transitive order of the linear ordering relations more frequently than that of the inclusion relations.

Method

Participants. Sixty-four undergraduate students at the Université de Bourgogne and the Université de Genève took part in the experiment. The participants were randomly distributed between the two experimental conditions (set inclusion texts then linear ordering texts vs. the reverse order of presentation).

Material. The texts presented to the participants were based on four narratives relating to fictitious tribes, cars, skyscrapers, and basketball players (Appendix B). Two of the narratives contained four set inclusion relations, and the other two presented four linear ordering relations. As in Experiment 1, the order in which the premises appeared in the texts did not correspond to the transitive order of the inclusion relations (i.e., *All As are Bs, All Bs are Cs,* etc.) or of the linear ordering relations (*A is larger than B, B is larger than C,* etc.). We used four permutations of the order of appearance of the premises (i.e., AB, CD, BC, DE; BC, DE, CD, AB; CD, AB, DE, BC; and DE, BC, AB, CD) and applied these to each of the four narratives. In all, we formed 16 texts (8 presenting inclusion relations and 8 presenting linear ordering relations).

Half of the participants (32) studied the two set inclusion texts first followed by the two linear ordering texts (presentation Order 1), whereas the other half studied the texts in the reverse order (Order 2). The order of presentation of the narratives and the permutations of the premises were counterbalanced between the participants.

For each of the texts presented, we constructed a notebook containing 10 conclusions relating to the premises. The conclusions corresponded to the 10 logical conclusions permitted by a series of four set inclusion relations (i.e., AB, BC, CD, DE, AC, BD, CE, AD, BE, AE) or four linear ordering relations (i.e., $A > B$, $B > C$, $C > D$, $A > C$, $B > D$, $C > E$, $A > D$, $B > E$, $A > E$). Five of these conclusions were presented in their valid form and the other five were inverted (e.g., BA, CA, EB). The truth values of the conclusions were counterbalanced in such a way that each of them appeared an equal number of times in its valid and in its inverted form. The order of presentation of the different conclusions in the notebooks was random.

Procedure. The experiment was conducted with groups of 6 participants and lasted 1 hr. First of all, the participants had to learn the information present in the text in order to be able to judge the logical validity of the conclusions relating to the premises in this text. The experimenter emphasized the importance of learning the texts thoroughly. The participants were given as much time as they needed to learn the texts (this learning phase lasted between 8 and 12 min). When the participants thought that they had remembered all the information, the texts were taken away from them and they had to recall all, and only, that information that was presented in the form *All . . . are . . .* present in the set inclusion texts, or in the form *. . . is higher than . . .* in the linear ordering texts. After this recall task, the participants had to judge the logical validity of the conclusions presented in the corresponding notebooks without, at any stage, going back over their answers. This procedure was repeated for all four texts.

Results

Recall task. The mean level of correct recall of the premises over all the conditions was .908. We performed a 2 (order of presentation: set inclusion–linear ordering vs. linear ordering–set inclusion) × 2 (type of relation: set inclusion vs. linear ordering relations) × 4 (type of premise: AB, BC, CD, and DE) ANOVA with the last two factors as within-subject factors on the rate of correctly recalled premises. The order of presentation of the relations had no effect on the correct recall level, $F(1, 62) < 1$, and did not interact with any of the other variables studied. The level of correct recalls of the set inclusion relations (.87) was lower than that of the linear ordering relations (.95), $F(1, 62) = 10.80, p < .01, MSE = 0.13$.

We predicted that linear ordering texts should induce more logically ordered recalls than set inclusion texts. For each of the 256 premise recalls, we calculated the frequency with which the logical order was reestablished (see Experiment 1, Table 1). An ANOVA on the frequency of reestablishment of the logical order with the order of presentation of the relations as between-subject factor and the type of relation as within-subject factor revealed that the mean level of reestablishment was higher for the linear ordering relations (.37) than for the set inclusion relations (.19), $F(1, 62) = 14.61, p < .001, MSE = 0.069$. The frequency of reestablishment in Order 1 (set inclusion texts and then linear ordering texts) was less (.19) than that obtained for reverse (Order 2: .37), $F(1, 62) = 5.05, p < .05, MSE = 0.20$. These two factors did not interact, $F(1, 62) = 0.316, p = .58, MSE = 0.07$. The rate of reestablishments was higher for both types of relation in the Order 2 than in the Order 1 (see Table 1). This last result suggests that to perform the set inclusion task first induced a nonreestablishment strategy for both types of relation, whereas to perform the linear ordering task first tended to induce a reestablishment strategy.

Transfer of encoding strategies from one type of text to the other. To analyze these effects of strategy transfer, we compared the reestablishment frequencies for the set inclusion texts and linear ordering texts as a function of whether they were presented first or second. When presented first, the reestablishment level was higher for linear ordering relations (.474) than for set inclusion relations (.12; Newman-Keuls, $p < .001$, see Table 1). The participants therefore had a tendency to spontaneously reestablish the linear ordering relations but not the set inclusion relations. Next, the reestablishment level for set inclusion relations was slightly lower when they were presented in first (.12) than in second position (.27; $p = .06$), whereas the reestablishment level for linear ordering relations was greater when presented in first (.474) rather than in second position (.27; $p = .008$). Finally, even when presented in second position, the reestablishment level of the set inclusion relations (.27) remained considerably lower than that of the linear ordering relations previously studied (.474; $p = .003$). These comparisons suggest that the reestablishment strategy used for the linear

Table 1
Frequencies of Reestablishment of the Logical Order of the Premises as a Function of the Type of Relation (Set Inclusion Relation [SIR] vs. Linear Ordering Relation [LOR]) and Their Order of Presentation (Experiment 2)

	Task	
Order	1	2
1	SIR: .12	LOR: .27
2	LOR: .47	SIR: .27

ordering relations is transferred, even though with difficulty, to the set inclusion relations, whereas the strategy of nonreestablishment for the set inclusion relations is transferred to the linear ordering relations.

In summary, the hypothesis that the order of recalls represents a sufficiently sensitive index of the organization of the premises in memory is strengthened by the relatively great propensity shown by participants to reestablish the logical order of the linear ordering relations compared with their reticence to reestablish the logical order of the set inclusion relations. Moreover, the fact that the prior use of a reestablishment strategy (when processing the linear ordering relations) does not result in a huge increase in the number of set inclusion relations reestablishments supports the idea that there is a difficulty at the level of logical organization (or integration) that is specific to the set inclusion relations.

Relations between the order of recall and performance in the evaluation task. The mean level of correct responses for the evaluation task was .80 (Table 2). We calculated a 2 (order of presentation) × 2 (type of relation) × 2 (truth value of the conclusions: true vs. false) × 4 (step size: 1, 2, 3, and 4) ANOVA, with repeated measures on the last three factors for the levels of correct judgments. This analysis revealed two main phenomena. First, the level of correct responses was lower for the set inclusion relations (.72) than for linear ordering relations (.88), $F(1, 62) = 37.5, p < .001$, $MSE = 0.17$, especially for adjacent propositions (.76 and .92, respectively), $t(63) = 7.34, p < .001$. Second, the Truth Value × Step Size interaction differed as a function of the type of relation. As observed in Experiment 1, this interaction was highly significant for set inclusion relations, $F(3, 186) = 23.99, p < .001, MSE = 0.08$, and explained a large part of the variance ($R^2 = .279$). The increase in the number of inferential steps resulted in a strong reduction in the level of correct judgments for valid conclusions (.91, .71, .64, and .59 for Step Sizes 1, 2, 3, and 4, respectively) and an increase in the level of correct judgments for invalid conclusions (.61, .68, .78, and .86, respectively). For linear ordering relations, this interaction was also significant, $F(3, 186) = 3.15, p = .026, MSE = 0.04$, but only explained a little more than 5% of the variance ($R^2 = .048$). The level of correct judgments decreased slightly for the valid conclusions (.91, .86, .80, and .81 for Step Sizes 1, 2, 3, and 4, respectively) and did not vary in a systematic way for invalid conclusions (.93, .88, .95, and .91). Thus, the effect of inferential distance on the evaluation of valid and invalid conclusions was weaker for the linear ordering than for the set inclusion relations, hence the double interaction Type of Relation × Truth Value × Step Size, $F(3, 186) = 11.96, p < .001$, $MSE = 0.06$.

Table 2
Frequencies of Correct Responses to True and False Propositions (Evaluation Task, Experiment 2)

	Step size			
Truth value	1	2	3	4
SIR				
True	.91	.71	.64	.59
False	.61	.68	.78	.86
LOR				
True	.91	.86	.80	.81
False	.93	.88	.95	.91

Note. SIR = set inclusion relation; LOR = linear ordering relation.

These results are compatible with the hypothesis that there is a link between the organization of information in memory and the production of inferences: The facts that the linear ordering relations are evaluated better, and are less sensitive to the increase of the step size, than the set inclusion relations would be due to their being more frequently integrated in a logically ordered representation when encoded in memory. In contrast, the evaluation of the set inclusion relations seems to be performed on the basis of an atomic representation of the premises. This nonintegration and the step-by-step processing that results from it would have two consequences. First, the conservation of the premises in memory would be more costly, and this might explain why the recall of the premises and the judgments of adjacent premises are worse for the set inclusion than for the linear ordering relations. Second, the increase in the inferential distance of the conclusions might result in an increase in working memory load and a reduction in the likelihood that participants will produce an inference that corresponds to the conclusion to be evaluated.

This interpretation has to be mitigated by two facts. First, even if the linear ordering relations were spontaneously reestablished near four times more than set inclusion relations, they were not reestablished systematically by all the participants. Second, the mean level of correct judgments for linear ordering relations did not replicate the increase with step size classically reported. We thus studied the relations between the reestablishment of the premises and the performances in the evaluation task in a more precise way.

Relations between the order of recall and performances in the evaluation task. For both types of relation, the frequency of reestablishment of the logical order of the premises was correlated with the level of correct judgments of adjacent propositions, $r(64) = .27$ and $.29$ for set inclusion and linear ordering relations, respectively, $ps < .05$, and inferences, $r = .24, p = .05$, for the set inclusion relations and $.51, p < .001$, for the linear ordering relations. However, when the effect relating to the number of correct recalls was partialed out, only the correlation between frequency of reestablishment and level of correct judgments of the linear ordering inferences remained significant, $r(61) = .455$, $\mathrm{p} < .001$. This partial correlation was greater than the partial correlation between the frequency of reestablishment and the level of correct judgments of the set inclusion inferences (.19), suggesting that only the reestablishments of the linear orderings correspond to an integration of the premises in memory that facilitates the production of inferences.

De facto, the participants who reestablished the logical order of the linear ordering relations more than two times out of three ($n = 18$) showed an increasing accuracy with step size (correct judgment rates were equal to .96, 1, 1, and 1 for valid conclusions and .97, 1, 1, and 1 for invalid conclusions), whereas the participants who reestablished the logical order of the set inclusion relations at the same rate ($n = 7$) showed a Step Size × Truth Value interaction (correct judgment rates of .93, .81, .78, and .86 for valid conclusions and .93, .76, 1, and 1 for invalid conclusions). These results thus suggest a qualitative difference between the linear ordering and set inclusion reestablishments, in addition to the quantitative difference previously described.

Discussion

This experiment established first the validity of the recall paradigm. When the linear ordering relations are substituted for the inclusion relations, the level of reestablishment of the logical order increases, in accordance with the generally accepted hypothesis that the linear ordering relations give rise to a spatially defined, integrated representation of the premises. The analysis of the transfer effects also suggests that the strategy of integrating information in memory used for the linear ordering relations is difficult to transfer to the set inclusion texts. This fact runs counter to Mynatt and Smith's (1979, see also Griggs & Warner, 1982) hypothesis, which supposes that the difficulties related to the set inclusion task are primarily due to the fact that the appropriate representational schema (i.e., a linear array) is available but not used. Even if the tasks immediately succeed one other (linear ordering before set inclusion relations) this is not enough to cause the transfer of the optimum strategy.

Second, it seems that the reestablishment of the linear ordering relations in the recalls corresponds to an integration of the premises in memory that facilitates the production of inferences. In contrast, the reestablishment of the logical order of the set inclusion relations does not facilitate the evaluation of inferences. This reestablishment would thus not provide an alternative to the procedure that consists of the step-by-step production of inferences in order to evaluate conclusions. The reestablishment of the logical order of the set inclusion relation might then correspond to a reordering of the premises that would nevertheless be stored relatively independently of one another in memory.

As a consequence, three types of representation might be constructed by the end of the comprehension process: an atomic representation of the premises that are not coordinated with one another; an atomic representation of the premises that nevertheless establishes the links between them on the basis of the terms they share; and, finally, an integrated representation that specifies the relations between all the terms present in the premises. A study of the reaction times should allow us to determine the nature of the constructed representation. Experiment 3, which made use of just such a paradigm, was therefore intended to determine (a) the structure of the representation of the set inclusion texts constructed at the end of the comprehension process, (b) the existence of an interindividual variability relative to the organization of the premises in memory, and (c) the effect of any such variability on inference production.

Experiment 3

It should be possible to decide between the competing hypotheses by measuring the RTs at each step-size value. The "atomic storage of premises" hypothesis (Barrouillet, 1996; Griggs & Osterman, 1980) holds that inferences should be produced step by step. The more steps involved in the proposition to be evaluated, the greater the number of premises that have to be retrieved into memory and, consequently, the greater the number of steps to be performed in the calculation. These can only result in an increase in the RTs. If participants store not the premises but the terms that they contain in an integrated representation (Potts, 1976, 1978), the RTs should become shorter as the number of steps involved in the inference increases, as observed in the case of inferences produced on the basis of linear ordering relations (e.g., "larger than," Potts, 1972, 1974).

Potts (1976) confirmed this latter prediction. However, Griggs and Osterman (1980) failed to replicate this result. Griggs and Osterman suggested that this difference in results may be caused by the samples taken and were the result of individual differences. It is possible that only the participants who achieve the best performances construct an integrated representation, thus explaining Potts's results (1976), whereas the others have to operate a step-by-step calculation of the conclusions.

In this experiment, participants were presented with a set inclusion task in which the RTs between the presentation of the proposition to be evaluated and the production of the response were recorded. If the participants who achieve the best performances integrate the information into a linear, ordered representation of the terms, they should, in accordance with Potts's (1976, 1978) observations, exhibit an RT pattern that falls with the increase in step size. In contrast, the cognitive load hypothesis predicts that RTs should increase with step size for all participants, even those who achieve the best performances.

Method

Participants. Thirty-two undergraduate students at the Université de Bourgogne took part in the experiment. None of them had participated in Experiments 1 and 2.

Material. The texts used were the same as for Experiment 1. However, we retained only one order for matching the terms to the values A, B, C, D, and E. This resulted in 16 texts (4 contents × 4 permutations of the order in which the premises appear). Each participant studied four texts and saw each content and each permutation. Following each text, the participants were presented with 20 propositions for evaluation: 10 valid propositions (AB, BC, CD, DE, AC, BD, CE, AD, BE, and AE) and 10 invalid propositions obtained by inverting the terms. These propositions were presented one by one on screen in a random order using software that also recorded the type of response and the RTs.

Procedure. The procedure was the same as that used for the evaluation task in Experiment 1. The participants had to give their responses (true or false) by pressing one of two keys labeled on the computer keyboard.

Results

Proportion correct data. The correct response level (68%) was identical to that observed in Experiment 1 (69%). As the step size increased, the level of correct responses fell for the true propositions and increased for the false propositions (Table 3). This Truth Value × Step Size interaction was significant, $F(3, 93) = 30.84, p < .001, MSE = 0.34$.

RT data. As far as RTs are concerned, if the participants store the terms in a linear array, then RTs should fall as step size increases, whereas the hypothesis of an atomic storage of premises and a step-by-step calculation predicts that RTs should increase with the step size. The RTs were recorded for all the responses, whether correct or not. For all participants, we disregarded any RTs that differed by more than two standard deviations from the mean of all RTs. This was the case in less than 5% of the recorded RTs. We then calculated a participant-specific mean RT for each step-size value (from 1 to 4) and for each response type (i.e., correct detections, correct rejections, false alarms, and omissions). We analyzed only the RTs for the correct responses.

The mean RTs for correct responses to true propositions varied significantly as a function of the step size, $F(3, 93) = 15.67, p < .001, MSE$[1] $= 7.11$ (Table 4).

The verification of true adjacent propositions ($M = 3{,}690$ ms) was faster than for two-step inferences ($M = 6{,}022$ ms, Newman–Keuls, $p < .001$), which was in turn faster than for three-step inferences ($M = 8{,}238$ ms, $p = .004$). The verification of four-step inferences (i.e., AE, $M = 6{,}366$ ms) was faster than for three-step inferences ($p = .006$), an effect known as the end-term effect. If we disregard the end-term effect on the AE inference, the calculation of an additional inferential step took slightly more than 2 s. This confirms our hypothesis of the cognitively costly calculation of inferences on the basis of atomically stored information that is coordinated in working memory. In contrast, although the mean RTs for the rejection of false propositions increased with the number of propositions, this effect was not significant ($F < 1$). This suggests that the strategies used for the evaluation of these propositions were different from those used to verify the true propositions.

A better test of our hypothesis would be to analyze the RTs for the correct responses for each type of true proposition presented (i.e., AB, BC, CD, and DE in the case of adjacent propositions, AC, BD, and CE for two-step inferences, AD and BE in the case of three-step inferences, and AE for four-step inferences, see Figure 1). Only the participants who produced at least one correct response for each proposition were considered in this analysis (29 participants out of 32).

As we predicted, the RTs increased as a function of the step size within each chain of propositions (i.e., AB–AC–AD–AE for A, BC–BD–BE for B, and CD–CE for C). Propositions of the form AX took longer to verify the more inferential steps they included, $F(1, 28) = 11.75, p < .01$, $MSE = 12.50$ for the linear trend. The same effect was observed for propositions of the form BX, $F(1, 28) = 14.73$, $p < .001$, $MSE = 30.7$, as well as for propositions of the form CX, $F(1, 28) = 17.48, p < .001, MSE = 4.93$. These results confirm and reinforce those derived from the global analysis of the RTs as far as the step-size effect is concerned. Within each inferential chain, the verification time for a proposition is a function of the number of inferential steps involved.

This analysis also revealed a phenomena that could shed some light on the processes involved in inference production and the resulting end-term effect: given a constant step size, the verification of AX propositions was faster than that of other propositions, except in the case of adjacent propositions. The RTs for the adjacent propositions AB, BC, CD, and DE ($M = 3{,}534$, 3,745, 3,946, 3,447 ms, respectively) did not differ. In contrast, as far as the two-step inferences were concerned, the proposition AC ($M = 5{,}381$ ms) was verified more rapidly than the propositions BD ($M = 6{,}492$ ms) and CE ($M = 6{,}384$ ms), $F(1, 28) = 4.37, p < .05$, $MSE = 4.95$. In the case of three-step inferences, AD ($M = 7{,}062$) was verified faster than BE ($M = 9{,}326$ ms), $F(1, 28) = 6.10, p < .02, MSE = 12.20$. Thus the considerable end-term effect observed for the mean RTs (8,238 ms at Step Size 3 as against 6,366 ms at Step Size 4, see Table 4) was due to the conjunction of two phenomena. The first corresponds to a reduction in the RTs in the AX chain when passing from AD ($M = 7{,}062$ ms) to AE ($M = 6{,}322$ ms). However, this difference was not significant, $F(1, 28) = 2.20, p = .15, MSE = 3.60$. The second relates to the fact that the global end-term effect is obtained by simply comparing the four-step inference AE with the three-step inference AD as well as with the BE inference. Now, the BE inference took the longest time to verify ($M = 9{,}326$ ms), and, more generally, BX inferences took longer to verify than AX inferences.

We hypothesize that these phenomena could result from the necessity to select the relevant information and, possibly, inhibit the irrelevant information in order to verify inferences (Conway & Engle, 1994). When an inference is calculated, we can suppose that the presence of interfering propositions in WM has to be inhibited if the calculation is to be successful. The first term (B) of BX type inferences activates the premise AB, which is of no use for calculation (similarly, CX propositions activate BC). In contrast, in the case of AX type propositions, the term A can only activate the single premise AB, which is always required for calculation. Thus AX inferences would not activate any premise that interferes with the term on which the calculation has to be based (i.e., A). The cognitive cost and the time

Table 3
Frequencies of Correct Responses to True and False Propositions as a Function of Step Size in Experiment 3

	Step size			
Truth value	1	2	3	4
True	.893	.755	.691	.672
False	.490	.591	.660	.734

[1] For convenience, the *MSE*s on RTs are given in s^2.

Table 4
Mean Reaction Times (in Milliseconds) and Standard Deviations for Correct Responses on True and False Propositions as a Function of Step Size in Experiment 3

	Step size							
	1		2		3		4	
Truth value	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
True	3,690	1,395	6,022	2,759	8,238	6,207	6,366	4,341
False	5,811	2,468	6,097	2,676	6,273	2,837	6,496	4,200

required to inhibit irrelevant information would therefore be greater in the case of BX and CX type propositions than for AX type propositions, thus explaining why the latter are verified more quickly.

Similarly, the calculation of the four-step inference AE is distinguished by the fact that it mobilizes all the stored premises (i.e., AB, BC, CD, and DE). This calculation does not therefore require participants to select between relevant and irrelevant information or to inhibit the latter. This results in a saving of cognitive resources when compared with the three-step inferences that, although they require one less inferential step, also demand the inhibition of irrelevant information. Thus, the end-term effect could result from a step-by-step calculation of inferences provided that selection processes of information in working memory are needed to perform the task. An alternative hypothesis would be that some participants used an heuristic process (e.g., all the propositions beginning with A or ending by E are true). However, such an heuristic process could not account for the step-size effect observed with the AX type propositions.

Individual differences. The results obtained for the mean RTs (true propositions) applied to all the participants irrespectively of their performance level. However, it could be that the participants who achieved the best performance exhibited an RT pattern that falls with the increase in step size, as suggested by Potts's (1976) and Griggs's and Osterman's (1980) hypotheses. Participants for whom the number of correct responses was lower than the mean value by more than two thirds of one standard deviation (i.e., score = 47 for $M = 54.72$ correct responses out of 80) were considered to be low-level participants ($n = 12$), whereas participants for whom the number of correct responses exceeded the mean value by more than two thirds of one standard deviation (i.e., score > 62) were considered to be high-level participants ($n = 8$). The other participants were considered to be medium-level participants (i.e., score between 47 and 62, $n = 12$). The change in mean RT as a function of the step size (correct responses to true propositions) for each of the three groups (i.e., low, medium, and high level) is presented in Figure 2.

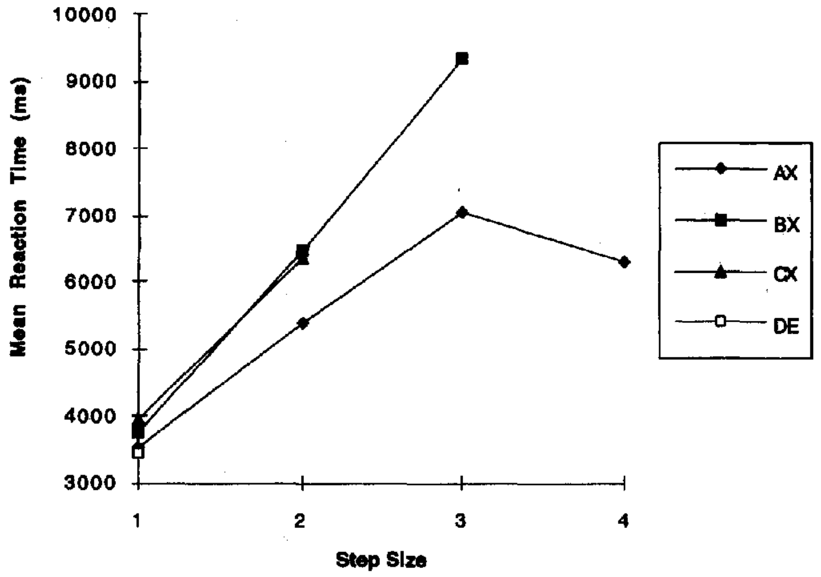

Figure 1. Mean reaction times for the verification of each of the 10 true propositions as a function of their initial term and the number of inferential steps (Experiment 3). AX refers to AB, AC, AD, and AE for 1, 2, 3, and 4 inferential steps, respectively; BX refers to BC, BD, and BE for 1, 2, and 3 inferential steps, CX refers to CD and CE for 1 and 2 inferential steps. Only one proposition has D as its initial term: DE, 1 inferential step.

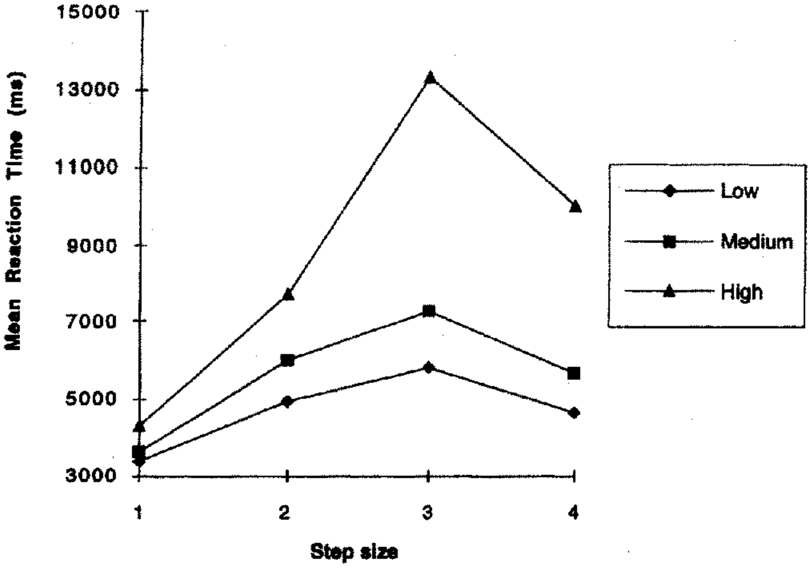

Figure 2. Mean reaction times for the verification of true propositions as a function of the number of inferential steps (step size) and the performance levels of the participants (Experiment 3).

The ANOVA for each group revealed that the step size had a significant effect on the RTs at each level of success: low level, $F(3, 33) = 4.71, p < .01, MSE = 2.57$; medium level, $F(3, 33) = 10.69, p < .001, MSE = 2.60$; high level, $F(3, 21) = 6.86, p < .01, MSE = 17.00$. A 3 (success level: low, medium, high) × 4 (step size) ANOVA with repeated measures on the last factor for the mean RTs at each inferential step value revealed a level effect, $F(2, 29) = 4.41, p = .02, MSE = 39.10$, the step-size effect noted above as well as an interaction between levels and step size, $F(3, 87) = 3.70$, $p = .003$, $MSE = 6.06$. The high-level participants took longer ($M = 8{,}831$ ms) than the medium-level participants ($M = 5{,}641$ ms) and the low-level participants ($M = 4{,}682$ ms) to verify the true propositions (Newman-Keul's test: $p = .03$ and .01, respectively). At the same time, the step-size effect was greater, the higher the participants' level of performance. Thus, the difference in RT between Step Sizes 1 and 3 was greater for high-level participants (from 4,288 to 13,340 ms; i.e., 9,052 ms) than for medium-level participants (from 3,603 to 7,269 ms, 3,666 ms) and low-level participants (from 3,377 to 5,805 ms, 2,428 ms), $F(1, 29) = 9.49, p < .01, MSE = 11.40$. In contrast, there was no difference in this effect between the medium- and low-level participants, $F < 1$.

For each participant, we calculated a linear regression between the step size of the accepted true propositions (the number of which varied between 20 and 40) and the recorded RT. This regression was significant ($p < .05$) for 19 of the 32 participants. For only one of them was the slope negative (−104 ms). However, the regression was not significant. Thus for none of the participants did we observe a reduction in RTs as the step size increased, as the hypothesis of the use of an ordered representation of the terms would suggest. The participants with the best performances were those for whom the increase in step size had the greatest effect on RT. The high-level participants had a mean slope of 2,210 ms, whereas this was 1,380 ms for medium-level participants and only 811 ms for the low-level participants, $F(2, 29) = 3.72, p = .04, MSE = 1.26$.

In summary, the analysis of RTs for correct responses to true propositions revealed three important facts. First, in accordance with our hypothesis, propositions took longer to verify the more inferential steps they involved (except AE, which took shorter to verify than AD). Second, the participants who achieved the best performances were also the slowest. Finally, the step-size effect on the RTs increased with participants' success in the task. This final point totally contradicts the hypothesis of the early integration of the premises in an ordered linear representation from which participants would "read" the inferences. In effect, participants who manage to construct such a representation should also perform best in the verification task. Consequently, if the problem is resolved on the basis of such a representation, the RTs of the participants exhibiting the best performance should decrease as the step size increases, as has been observed in the processing of linear series. In contrast, the RTs of participants who were not able to construct this representation, and who therefore achieved the worst performances, should be highly affected by the step size. This is the complete opposite of what we observed. Whatever the level of success, the verification time for the propositions increased with the step size, and this effect was all the greater, the more successful the participants were in the task.

It is possible that these differences between the slopes as a function of performance level, as well as the fact that the participants exhibiting the weakest performances are also the fastest to verify adjacent propositions, may be due to a speed–accuracy trade-off. However, the possible existence of such a trade-off does not diminish the significance of our main result in any way. The high-level participants had the steepest slopes, whereas the hypothesis of the early construction of an integrated representation predicts that the slopes relating to these participants should have the opposite sign.

Discussion

These results thus confirmed Barrouillet's (1996) hypothesis. When participants have to evaluate an inferential proposition, they perform a cognitively costly calculation for the step-by-step integration of atomically stored premises. The RTs analysis suggests that this strategy is universal when the premises are identical to those used here. In the light of these results, it is practically impossible that any of our participants might have reasoned on the basis of a previously integrated representation (e.g., in the form of a linear representation of the terms). In contrast, we can be certain that the participants who achieve the best performances use a computational strategy, with resolution taking longer the greater the number of inferential steps involved in the proposition for evaluation.

General Discussion

The results of the three experiments presented here suggest that the information contained in the set inclusion texts are stored in memory in an atomic way and are coordinated on a step-by-step basis in working memory when an inference has to be produced or evaluated. These results thus indicate that the processes involved in the comprehension and storage of information in the set inclusion task do not include the spontaneous, on-line production of logical inferences. They therefore differ from the results obtained by Lea (1995) when testing Braine et al.'s model (1984). Lea suggested that logical inferences (i.e., disjunctive syllogism) might be produced on-line even if they are not necessary for the establishment of the global coherence of the text. This contrast in results suggests that if participants use a transitive inference rule that is authorized by the universal quantifier "All" when solving our task (see Rips, 1994), then this rule is of a very different nature from those described by Braine et al. (1984).

Thus the status of the representation of information that supports the production of inferences in the set inclusion task appears to be closer to what Kintsch (1988) terms a *propositional text base* and Graesser, Swamer, Bagget, & Sell (1996) an *explicit text base* than an integrated, ordered representation of the mental model type. The results of the three experiments presented here, together with those obtained by Barrouillet (1996), suggest that inferences are produced through the retrieval and coordination of premises that are stored atomically, probably in a propositional form.

We hypothesize that the reading of the text leads to the construction of a propositional text base that contains the propositions that can be directly derived from the text. As suggested by many models of textual comprehension (Kintsch, 1988; McKoon & Ratcliff, 1992; van den Broek, Risden, Fletcher, & Thurlow, 1996) any overlap caused by the repetition of terms in the premises (e.g., BC and CD) might result in the production of simple inferences (i.e., two instances of the same word refer to the same concept) that ensure textual coherence. However, this coherence appears to be difficult to establish. Because the premises are presented out of order, these overlaps often involve two propositions that are distant from one another in the presented text. Only participants with good reading abilities (e.g., with a high reading span) would be able to organize the propositions in the text base sufficiently well for the premises containing a common term to be linked in long-term memory (Singer, Andrusiak, Reisdorf, & Black, 1992).

This activity would result in high-performance participants identifying the logical order of the premises and encoding them in memory in this order. The results of Experiment 3 suggest that this encoding does not correspond to the formation of an integrated mental model of the entirety of the information that might permit the direct reading of inferences, as is probably the case with relations of order such as "higher than." However, the establishment of connections between the propositions and the encoding of the premises in the logical order (AB–BC–CD–DE) intrinsically ensure the directionality (i.e., B → C vs. C → B) of each of the stored relations because such a representation fixes the order of the terms. This would explain the fact that in Experiment 1, the participants who reestablished the logical order on recall were those who were least inclined to accept the symmetry of the inclusion relation.

The establishment of connections between premises distributed throughout a text would require a high reading span. In effect, a premise that has already been processed must be sufficiently activated in working memory for information undergoing processing to be associated with it (Just & Carpenter, 1992). Thus Daneman and Carpenter (1983) showed that the reading span, among other things, is a good predictor of the resolution of anaphoric relations. In consequence, participants with a high reading span would be better able to organize the propositions optimally in the text base and would be the least inclined to accept the symmetry of the relation (Barrouillet, 1996).

Nevertheless, the impossibility of constructing an integrated mental model would have two consequences. First, the inferences that assure the coherence of the text would provide a minimalist representation "from which strategic inferences could be constructed by retrieval operations" (McKoon and Ratcliff, 1992, p. 440). Second, the atomic storage of the premises would make the permanent maintenance in working memory of all the information supplied by the text impossible. Inferences would therefore be calculated step by step by means of the retrieval from LTM of the propositions activated by the terms of the inference to be judged. The efficiency of calculation would primarily depend on the participant's ability to retrieve and maintain in WM a large number of propositions and intermediate conclusions in cases where calculation requires multiple inferential steps, and this would be relatively independent of the way in which these propositions are organized in LTM. When the propositions are interconnected in the text base, these connections would ensure the directionality of the relations (see above) and might also facilitate the calculation of the inferences. In effect, connections established between the atomic propositions in the text base might facilitate the successive retrieval of various items of information that have to be coordinated in working memory through a process of spreading activation (Anderson, 1983, 1993). This would explain why, in Experiment 1, the reestablishment level on recall slightly but significantly ($r = -.273$) reduced the effect of step size on performance in the evaluation task.

This would also explain the relative independence of the storage and calculation processes observed by Barrouillet (1996). Recall that the reading span was a good predictor of the tendency to reject the symmetry of the relation in the set inclusion task, whereas the alphabet recoding performance predicted the ability to produce inferences. As already explained, the organization of the propositions in the text base would be dependent on participants' reading capacities and would ensure the directionality of the stored relations. Because reading comprehension abilities are correctly evaluated by the reading span, it is not surprising that this span was related to the rejection of symmetry. The alphabet recoding task requires that participants (a) maintain the letters for transformation in memory, (b) perform frequent retrievals of the alphabetical chain from LTM, and (c) maintain the results in working memory. The processes involved in this task are very close to those required for inference production in the set inclusion task (SIT). It is therefore not surprising that alphabet recoding performance was related to the ability to produce inferences.

The question of whether calculations themselves are performed on the basis of rules of inference (Rips, 1994) or the construction of transient mental models (Johnson-Laird et al., 1994) remains open. According to Lea (1995), the rules postulated by Braine et al. (1984) should be applied on-line during reading. Our results suggest that if such syntactic rules are implemented in the set inclusion task, they are manifestly of a different nature. Similarly, the implementation of the syntactic rules proposed by Rips (1994) for the "All" quantifier could not seem to be particularly compatible with the temporal costs observed in Experiment 2 (increase of 3.4 s between 1 and 2 inferential steps and 5.7 s from 2 to 3 steps!). However, this extra time may result from the difficulty of retrieving and selecting propositions that match the activation conditions of the rule. This increase could point to the construction of transient mental models. Such models have been evoked by Johnson-Laird et al. (1994) in their response to O'Brien et al.'s (1994) criticisms. It is true that participants could construct transient mental models from pairs of isolated propositions on the basis of coreference, drawing intermediate conclusions. However, this suggestion weakens the main proposal of the theory, that participants would construct an integrated representation that is isomorphic to the state of affairs described and makes the mental models theory indistinguishable from its concurrents. Moreover, it could be supposed that these transient models should further the construction of an integrated model because each of these transient models constitutes a piece of this integrated model. However, the completion of the set inclusion task did not have any effect on the reestablishment of the logical order when recall-before and recall-after conditions were compared in Experiment 1.

Concluding Comments

In a comparison of the ways in which the mental models theory has been applied to the fields of text comprehension and reasoning, Garnham (1996) suggested that there is no alternative to this theory in the field of textual comprehension. In contrast, he proposed that in the reasoning field the mental models theory does not exclude the possibility that some reasoning might be permitted by the manipulation of mental representations of sentences. The results issuing from the three experiments presented here seem to confirm this point of view. They suggest that the resolution of the set inclusion task is not based on the construction of an integrated mental model of all the information supplied, in contrast to the linear ordering tasks, even though these tasks are structurally and logically identical and even though the participants have no additional factual knowledge relating to the topic of the texts. The problem is therefore to determine the reasons for which participants do not process the two types of problem in the same way.

There are three reasons that might explain why participants do not construct integrated mental models in the set inclusion task. The first might be the difficulty that they experience in representing relations of the type *All As are Bs.* Not only is comprehension of the inclusion relation a late occurrence in development (Barrouillet & Poirier, 1997), its representation could be difficult because it brings together sets rather than isolated items as is the case of relations of the type *Mark is taller than Paul.* Even though Johnson-Laird suggested that this difficulty is resolved through the representation of a finite and arbitrary number of tokens (Johnson-Laird, 1983; Johnson-Laird & Bara, 1984), the psychological reality of such representations is still in doubt (Ford, 1994). At the same time, the very ambiguity of statements such as *All As are Bs,* which can signify either the identity or the inclusion of the two sets, might make them difficult to represent as the diversity of graphic illustrations given by N'Guyen and Revlin's (1993) participants indicates, and might therefore prevent the construction of mental models. Mani and Johnson-Laird (1982) indeed observed that when a statement is compatible with more than one model, the construction of a mental model becomes difficult and the information is encoded in a propositional form. In contrast, a relation such as *A is larger than B* could be a lot easier to represent with a mental image for example.

The second reason could be the absence of knowledge in LTM relating to the nature of the inclusion relation itself (e.g., the fact that it is a linear relation that permits the ordering of sets). In the field of text comprehension, the mental models theory specifies that the mental model results from the interaction of a propositional representation of the text (close to the linguistic structure) and inferences based on the participant's general knowledge relating to the situation. In the set inclusion task, it is possible that participants who are not mathematicians do not possess knowledge about the fact that the inclusion relation makes it possible to establish orders within sets in the same way that the "higher than" relation makes it possible to order items. The low level of transfer observed in Experiment 2 from the linear ordering texts toward the set inclusion texts provides evidence in this respect. In the absence of such knowledge, the participants would not appear to implement any strategy for organizing information within an integrated representation that makes the linear relation explicit, unlike what is observed with the "higher than" relation that is frequently used to order objects.

The third reason could be that, unlike linear ordering texts, the key terms of set inclusion texts cannot be aligned along a single underlying dimension (e.g., the size for relations such as *Paul is taller than Mark, Mark is taller than John, . . .*), thus hampering the construction of an integrated representation.[2] However, the mental models theory does not actually permit such distinction because the tokens used to construct mental models for syllogisms would be quite abstract representations and the relation between them is assumed to be represented by their mere ordered chaining in a single model (see Johnson-Laird & Byrne, 1991, p. 119). Thus, the abstractness of these representations should make it possible to integrate tokens referring to classes very different in nature in a single representation.

Thus, the application of the mental models theory to reasoning requires a precise definition of the conditions under which an integrated mental model of the entirety of the supplied information is constructed. This construction does not appear to depend only on the quantity of information to be integrated or the participants' factual knowledge concerning the described situations (see the different results obtained with linear ordering texts and the set inclusion texts) but also on the nature of these relations. As stressed by Stevenson (1996), Johnson-Laird (1983) proposed that there are two kinds of representation of discourse, a superficial propositional representation and a mental model. The former, structurally similar to the linguistic input, represents the meaning of the utterances. The latter, structurally similar to states of affair in the world, represents their reference. Our results suggest that when this reference is difficult to represent, correct reasoning might be based on the propositional representations alone.

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

References

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Anderson, J. R. (1993). *Rules of the mind.* Hillsdale, NJ: Erlbaum.

Barrouillet, P. (1996). Transitive inferences from set inclusion relations and working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 6,* 1408–1422.

Barrouillet, P., & Lecas, J. F. (1998). How can mental models theory account for content effects in conditional reasoning? A developmental perspective. *Cognition, 67,* 209–253.

Barrouillet, P., & Poirier, L. (1997). Comparing and transforming: An application of Piaget's morphisms theory to the development of class inclusion and arithmetic problem solving. *Human Development, 40,* 216–234.

Bonatti, L. (1994a). Propositional reasoning by model? *Psychological Review, 101,* 725–733.

Bonatti, L. (1994b). Why should we abandon the mental logic hypothesis? *Cognition, 50,* 17–39.

Braine, M. D. S. (1990). The natural approach to reasoning. In W. F. Overton (Ed.), *Reasoning, necessity, and logic: Developmental perspectives* (pp. 135–158). Hillsdale, NJ: Erlbaum.

Braine, M. D. S., & O'Brien, D. P. (1991). A theory of if: A lexical entry, reasoning program, and pragmatic principles. *Psychological Review, 98,* 182–203.

Braine, M. D. S., O'Brien, D. P., Noveck, I. A., Samuels, M. C., Lea, R. B., Fisch, S. M., & Yang, Y. (1995). Predicting intermediate and multiple conclusions in propositional logic inference problems: Further evidence for a mental logic. *Journal of Experimental Psychology: General, 124,* 263–292.

Braine, M. D. S., Reiser, B. J., & Rumain, B. (1984). Some empirical justification for a theory of natural propositional logic. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 18, pp. 313–371). San Diego, CA: Academic Press.

Carlson, R. A., Lundy, D. H., & Yaure, R. G. (1992). Syllogistic inference chains in meaningful text. *American Journal of Psychology, 105,* 75–99.

Carroll, M., & Kammann, R. (1977). The dependency of schema formation on type of verbal material: Linear orderings and set inclusions. *Memory & Cognition, 5,* 73–78.

Conway, A. R. A., & Engle, R. W. (1994). Working memory and retrieval: A resource-dependent inhibition model. *Journal of Experimental Psychology: General, 4,* 354–373.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19,* 450–466.

Daneman, M., & Carpenter, P. A. (1983). Individual differences in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 561–584.

Dickstein, L. S. (1978). Error processes in syllogistic reasoning. *Memory & Cognition, 6,* 537–543.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction.* Hillsdale, NJ: Erlbaum.

Ford, M. (1994). Two modes of mental representation and problem solution in syllogistic reasoning. *Cognition, 54,* 1–71.

Frase, L. T. (1969). Structural analysis of the knowledge that results from thinking about text. *Journal of Educational Psychology Monograph, 60,* 1–16.

Garnham, A. (1996). The other side of mental models: Theories of language comprehension. In J. Oakhill & A. Garnham (Eds.), *Mental models in cognitive science* (pp. 35–52). Hove, England: Psychology Press.

Graesser, A. C., Swamer, S. S., Baggett, W. B., & Sell, M. A. (1996). New models of deep comprehension. In B. K. Britton & A. C. Graesser (Eds.), *Models of understanding text* (pp. 1–32). Hillsdale, NJ: Erlbaum.

Griggs, R. A. (1976). Logical processing of set inclusion relations in meaningful text. *Memory & Cognition, 4,* 730–740.

Griggs, R. A., & Osterman, L. J. (1980). Processing artificial set inclusion relations. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 39–52.

Griggs, R. A., & Warner, S. A. (1982). Processing artificial set inclusion relations: Educing the appropriate schema. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 51–65.

Johnson-Laird, P. N. (1983). *Mental models.* Cambridge, England: Cambridge University Press.

Johnson-Laird, P. N. (1993). *Human and machine thinking.* Hillsdale, NJ: Erlbaum.

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition, 16,* 1–62.

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

Johnson-Laird, P. N., Byrne, R. M. J., & Schaeken, W. (1994). Why models rather than rules give a better account of propositional reasoning: A reply to Bonatti and to O'Brien, Braine, and Yang. *Psychological Review, 101,* 734–739.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99,* 122–149.

Kintsch, W. (1986). Learning from text. *Cognition and Instruction, 3*(2), 87–108.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction–integration model. *Psychological Review, 2,* 163–182.

Kintsch, W. (1995). Information accretion and reduction in text processing: Inferences. *Discourse Processes, 16,* 193–202.

Kintsch, W., & Welsch, D. M. (1991). The construction–integration model: A framework for studying memory for text. In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human in honor of Bennett B. Murdock* (pp. 367–385). Hillsdale, NJ: Erlbaum.

Lea, B. (1995). On-line evidence for elaborative logical inference in text. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 6,* 1469–1482.

Mani, K., & Johnson-Laird, P. N. (1982). The mental representation of spatial description. *Memory & Cognition, 10,* 181–187.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 3,* 440–466.

Mynatt, B. T., & Smith, K. H. (1979). Processing of text containing artificial inclusion relations. *Memory & Cognition, 7,* 390–400.

Newstead, S. E., & Griggs, R. A. (1984). Fuzzy quantifiers as an explanation of set inclusion performance. *Psychological Research, 46,* 377–388.

Newstead, S. E., Keeble, S., & Manktelow, K. I. (1985). Children's performance on set-inclusion and linear-ordering relationships. *Bulletin of Psychonomic Society, 23,* 105–108.

N'Guyen, D. B., & Revlin, R. (1993). Transitive inferences from narrative relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1197–1210.

Oakhill, J. (1996). Mental models in children's text comprehension. In J. Oakhill & A. Garnham (Eds.), *Mental models in cognitive science* (pp. 77–94). Hove, England: Psychology Press.

O'Brien, D. P., Braine, M. D. S., & Yang, Y. (1994). Propositional reasoning by mental models? Simple to refute in principle and in practice. *Psychological Review, 4,* 711–724.

Potts, G. R. (1972). Information processing strategies used in the encoding of linear ordering. *Journal of Verbal Learning and Verbal Behavior, 11,* 727–740.

Potts, G. R. (1974). Storing and retrieving information about ordered relationship. *Journal of Experimental Psychology, 103,* 431–439.

Potts, G. R. (1976). Artificial logical relations and their relevance to semantic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 2,* 746–758.

Potts, G. R. (1978). The role of inference in memory for real and artificial information. In R. Revlin & R. E. Mayer (Eds.), *Human reasoning* (pp. 139–161). Washington DC: V. H. Winston & Sons.

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

Rips, L. J. (1994). *The psychology of proof.* Cambridge, MA: MIT Press.

Singer, M., Andrusiak, P., Reisdorf, P., & Black, N. L. (1992). Individual differences in bridging inference processes. *Memory & Cognition, 20,* 539–548.

Stevenson, R. J. (1996). Mental models, propositions, and the comprehension of pronouns. In J. Oakhill & A. Garnham (Eds.), *Mental models in cognitive science* (pp. 53–76). Hove, England: Psychology Press.

van den Broek, P., Risden, K., Fletcher, C. R., & Thurlow, R. (1996). A "landscape" view of reading: Fluctuating patterns of activation and the construction of a stable memory representation. In B. K. Britton & A. C. Graesser (Eds.), *Models of understanding text* (pp. 165–188). Hillsdale, NJ: Erlbaum.

Appendix A

Type of Text Used in Experiment 1 (Translations From the Texts Written in French)

Text 1, Permutation AB, CD, BC, DE

On a path, Paul sees a variety of construction materials, in particular iron rods. He notices that they do not always have the same appearance—not the same size or the same color. He notices that all the black rods are hollow. Some rods are bent and others are straight. He sees that all the bent bars are long. When he lifts them, he notes that all the hollow bars are bent. Because some of the bars have been used so many times, some of them have been damaged and Paul notes that all the long bars are damaged.

Text 2, Permutation BC, DE, AB, CD

A recent ethnological study investigated the way of life of the peoples of Central Ugala. It is the custom in this country to banish certain types of people from the tribes. All the exiles in Central Ugala are mountain-dwellers. The researchers are certain that all the farmers in this country are extremely peaceful, which is reflected in their artistic productions. Among the peoples studied, it appeared that all the Fundalas are exiles from other tribes in Central Ugala. Since the high grounds provide excellent soil for agriculture, all the mountain-dwellers in Central Ugala are farmers. There are approximately fifteen different tribes in this region.

Text 3, Permutation CD, AB, DE, BC

Mister Dupont has just been employed in the vehicle registration office of Casteltown which surveys the characteristics of vehicles registered in the town. He is told that, for environmental reasons, all the municipal vehicles are electric. So that they can be easily recognized, all the vehicles with sirens are red and green. To do his job, Mister Dupont has to remember that all electrical vehicles have

MC in their registration number. Traffic flow in the town is eased by the fact that all the red and green vehicles are municipal vehicles.

Text 4, Permutation DE, BC, AB, CD

As winter approaches, John decides to buy a big pullover. The retailer tells him that, because of the current fashion, all the well-known brands of pullover have polo-necks. Then he shows him that all the pullovers in the shop window are made of pure wool and are therefore well suited to the season. He states that because of the current fashion, all the Jacquard pullovers are in the shop window. He also states that John can make his choice with full confidence because all the pure wool pullovers are well-known brands.

Appendix B

Type of Texts Used in Experiment 2

Set Inclusion Texts

Text 1

A recent ethnological study investigated the way of life of the peoples of Central Ugala. It is a custom in this country to exile certain types of people. All the Fundalas are exiles from other tribes in Central Ugala. Moreover, for reasons of defense and security, all the exiles in Central Ugala are mountain-dwellers. Since the high grounds provide excellent soil for agriculture, all the mountain-dwellers in Central Ugala are farmers. Furthermore, the ethnologists state that all the farmers in this country are peace-loving. There are approximately fifteen different tribes in this region.

Text 2

Mister Dupont has just been employed in the vehicle registration office of Casteltown which surveys the characteristics of vehicles registered in the town. First of all he is told that, so that they can be easily recognized, all the vehicles with sirens are red and green. What is more, since all the red and green vehicles are municipal vehicles, they are not taxed. For environmental reasons, all the municipal vehicles are electric. To do his job, Mister Dupont also has to remember that all electrical vehicles have MC in their registration number.

Linear Ordering Texts

Text 1

On a trip to Paris, John visits the La Défense district, which mainly consists of sky-scrapers. The guide points out that the Seiko building is higher than the Gan building. Then, during a well-earned break on a sunny café terrace, he notices that the Gan building is higher than the IBM building. From the central esplanade, John sees that the IBM building is higher than the Elf building. As he leaves the area, John notices that the Elf building, which is the most recent to go up, is higher than the BNP building. After the visit, John decides to consult a physiotherapist because he's got a stiff neck which is beginning to become painful.

Text 2

In order to make up teams of the same level during a basketball training session, the trainer decides to take account of two factors he knows to be crucial: the height and expertise of the players. First of all, he compares the height of the children who want to play. The trainer has already seen that Lewis is taller than Frank. Now he sees that Frank is taller than George. When he gets them to stand back-to-back, the trainer sees that George is taller than Paul. Finally, he sees that Paul is taller than James who happens to be the most experienced.

Received April 17, 1997
Revision received July 6, 1999
Accepted July 19, 1999 ■