Transitive Inferences From Set-Inclusion Relations and Working Memory

Pierre Barrouillet
Université de Bourgogne

Research into the evaluation of transitive inferences from inclusion relations has shown, though there are important intersubject differences, that participants tend to consider the inclusion relation to be symmetrical and that they fail to assume transitivity: the level of correct responses falls for true inferences and increases for false inferences, the greater the number of inferential steps involved (Truth \times Distance interaction). This article puts forth a cognitive load hypothesis to account both for the interaction and for intersubject differences. It was shown that the cognitive load associated with the calculation of the inferences predicted error rates and accounted for the Truth \times Distance interaction. Moreover, the performances were correlated with participants' working memory span.

Research on deductive reasoning in adults has often led to the conclusion that human reasoning does not conform to the rules of logic and is subject to multiple biases (Evans, 1982, 1989). As pointed out by Johnson-Laird and Byrne (1991), "People are rational in principle, but fallible in practice" (p. 19). Numerous theories of reasoning suggest that the limitation of the processing capacities of human participants represents a major source of error (e.g., Braine, 1990; Johnson-Laird & Byrne, 1991; Rips, 1983). The purpose of this article is to determine the extent to which the errors observed in a transitive inference evaluation task known for its difficulty (the set-inclusion task) can be explained in terms of a problem of cognitive load and whether interparticipant differences are due, at least in part, to differences in processing capacity.

The most frequently studied field has been that of the comprehension of the conditional (if \( p \) then \( q \)), in particular using Wason's (1966, 1968) selection task. Although some models point to the implementation of rules of inference (Braine & O'Brien, 1991; Braine & Rumain, 1983; Rips, 1983, 1990), to the manipulation of mental models (Johnson-Laird & Byrne, 1991; Johnson-Laird, Byrne, & Schaecken, 1992), to the use of heuristics (Evans, 1989), or to the mobilization of pragmatic reasoning schemas (Cheng & Holyoak, 1985), all agree that performance by adults is weak.

The difficulties adults experience when trying to reason logically are not limited to the conditional and the selection task alone. Similar conclusions can be drawn concerning other relations and connectors. Of the research that has been conducted, the work into the transitive inferences permitted by the universal quantifier all is particularly interesting. On the one hand, inclusion (all \( A \) are \( B \)) is the formal equivalent in class logic of the conditional in propositional logic. On the other hand, the emergence of the idea of inclusion in children has long been an object of study and the stages of its development have long been known (Barrouillet, 1992, 1993; Bideaud, 1988; Campbell, 1991; Lautrey & Bideaud, 1985; Piaget & Inhelder, 1959; Winer, 1980). In addition, the treatment of conditional statements is simplified by the introduction of meanings relating to class-inclusion relations (George, 1992).

Research involving adult participants has concentrated on the ability of these participants to evaluate the truth of statements of the form all \( A \) are \( D \) or all \( C \) are \( B \) in cases where the premises all \( A \) are \( B \), all \( B \) are \( C \), and all \( C \) are \( D \) have been provided in a text. Thus, Frase (1969) confronted participants with a set-inclusion task in which they were asked to learn the following passage:

"The Fundalas are outcasts 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 set 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 \( E \), respectively, and in the relation all \( A \) are \( C \), the letters \( A \) and \( C \) reflect the order of the sequence \( AC \). Four types of statements may be constructed using the classes presented in the text: (a) true adjacent statements in which the terms of the statement are immediately adjacent in the inclusive hierarchy (i.e., \( AB \), \( BC \), \( CD \), and \( DE \) and are explicit in the text; (b) true remote statements, which can be logically deduced from the preceding statements by transitivity (all \( A \) are \( C \) because all \( A \) are \( B \) and all \( B \) are \( C \)); (c) false adjacent statements obtained by inverting the terms of

I would like to thank Nathalie Gavens, Catherine Bosu, Nathalie Isasa, and Patricia Longère for assistance in conducting the experiments. I gratefully acknowledge the thoughtful comments on drafts of this article by Michel Fayol, Pierre Perruchet, and Stephen Newstead. Finally, I would like to express my gratitude to Tim Pownall.
true adjacent statements when the relation in question is not symmetrical (BA, CB, DC, and ED); and (d) false remote statements obtained by inverting the terms of true remote propositions. Each statement, whether true or false, can be described in terms of the number of inferential steps (the number of premises) necessary for its deduction. There are four statements requiring a single step (the adjacent statements), three requiring two steps (AC, BD, and CE), two involving three steps (AD and BE), and a single four-step statement (AE). The participants were asked to determine if each statement was a valid deduction from the text.

Two phenomena were observed: (a) The participants tended to think of the inclusion relation as symmetrical and frequently accepted false statements as valid deductions (i.e., the participants inferred all B are A from the premise all A are B), and (b) the level of acceptance of statements (whether true or false) decreased as the number of inferential steps required rose. This led to an interaction between the number of inferential steps and truth value. Consequently, an increase in the number of inferential steps led to a fall in the level of correct responses concerning true statements and an increase in the level of correct responses concerning false statements. This is referred to as a step-size effect. The participants therefore displayed a symmetrical (inversion of relation) conception of the relation and failed to assume transitivity. It should be noted that the conversion of the relation (all A are B therefore all B are A) is similar to the tendency to treat the conditional as a biconditional (however, on this point see also Ruman, Connell, & Braine, 1983). Just as the statement if p then q leads participants to produce the inference if q then p (Geis & Zwicky, 1971), the statement all A are B results in the acceptance of all B are A.

Subsequent experimental testing of this paradigm has confirmed the existence of these two phenomena (Carroll & Kammann, 1977; Griggs, 1976; Griggs & Osterman, 1980; Griggs & Warner, 1982; Mynatt & Smith, 1979; Newstead & Griggs, 1984; Newstead, Keeble, & Manktelow, 1985; Potts, 1976) despite the fact that considerable interparticipant differences have been identified (Griggs & Osterman, 1980; Mynatt & Smith, 1979). Experimenters have proposed two interpretations designed to account for the poor performance displayed in hierarchical inclusion tasks. The first interpretation holds that participants may interpret the relation erroneously (Newstead & Griggs, 1984; Potts, 1976, 1978) and are therefore unable to reason logically (Griggs, 1976). The second interpretation suggests that participants may mobilize an inappropriate reasoning schema (Carroll & Kammann, 1977; Griggs & Warner, 1982).

So far, two hypotheses have been proposed to account for the Truth × Distance interaction. Griggs (1976) suggests that participants are hesitant and not inclined to make transitive inferences on the basis of set-inclusion information. This caution would be greater the larger the number of inferential steps (the greater the "distance") involved and would result in an increase in the number of "false" responses. This increase would be a direct function of the number of inferential steps because, as Griggs suggested, participants would store the premises in an isolated way before calculating inferences. Potts (1976, 1978) has suggested that participants might treat the class-inclusion relation as a similarity relation from a linearly ordered representation of the terms (i.e., A-B-C-D-E). This similarity would diminish as the distance separating the terms in the chain of inclusion increases. Two distant terms would be judged to be less similar than two adjacent terms. Thus, the more inferential steps involved in a statement, the more likely participants are to consider it to be false (see also Newstead, Pollard, & Griggs, 1986). Despite these hypotheses, Evans, Newstead, and Byrne (1993) concluded their review of the literature with the admission that for the moment there is no satisfactory explanation for observed performances in hierarchical inclusion tasks.

However, to the best of our knowledge, one hypothesis has not yet been explored: The Truth × Distance interaction may be due to a problem associated with working memory (WM) load and the limitation of cognitive resources. This cognitive load hypothesis appears to be compatible both with theories that adhere to the field of mental logic and with the theory of mental models.

In terms of the field of mental logic and the explanation of reasoning processes by means of rules of inference, Carlson, Lundy, and Yaure (1992) have suggested adapting Rips's (1983) ANDS (a natural deduction system) model to account for the resolution of hierarchical inclusion tasks. They suggested that the production of transitive inferences on the basis of inclusion relations might be considered as a process of constructing and evaluating argument structures within limited-capacity WM (e.g., Baddley, 1986). According to Rips, deductive reasoning is made possible by a set of rules of inference that are applied to the premises stored in WM. According to this model, WM would contain two components: (a) an assertion tree containing the stored premises and deduced statements and (b) a goal tree storing the statement to be evaluated and the subgoals set during the deductive process. Rips, like Carlson et al. (1992), considered that two factors might be responsible for errors: (a) the probabilistic way in which the rules are applied and (b) the WM load, which is a function of the number of statements stored in the assertion and subgoal trees. The frequency of errors would increase as a function of the size of these trees and the number of rules to be applied (or the number of applications of the same rule).

The suggestions made by Rips (1983) and by Carlson et al. (1992) are compatible with the facts previously observed in the set-inclusion task. The number of errors increases as the number of inferential steps involved in the true statement grows and as the number of inferential steps involved in the false statement falls. Now, the increase in the number of inferential steps necessary for the verification of a statement results in (a) an increase in the number of premises to be taken into account and, therefore, an extension of the statement and goal trees and (b) an increase in the number of times the rules of inference are applied. Thus, the more inferential steps a
statement involves, the greater the cognitive load resulting from its calculation. This would explain why performance falls as the number of inferential steps involved in the true statement (which is more difficult to calculate) increases but improves as the number of inferential steps involved in the false statement (which has less chance of being accepted) increases.

Similarly, Johnson-Laird and Byrne (1991) considered that reasoning becomes increasingly difficult as the number of mental models that need to be constructed and manipulated for the resolution of the problem increases. If participants construct a mental model for each of the premises necessary for deduction, the cognitive load increases with the number of inferential steps involved in the statement to be evaluated. Thus, the two theories result in the same predictions. The cognitive load increases with the inferential distance. The greater this distance, the worse the chance that the inference will be produced.

This article examines the implications of this cognitive load hypothesis for the Truth × Distance interaction and for the existence of interparticipant differences. Indeed, Johnson-Laird and Byrne (1991) have suggested that these differences are due to different processing capacities in WM. If the difficulties participants experience in performing the set-inclusion task are partly due to the cognitive load involved in the evaluation of the statements, then the performances would depend on participants’ WM capacity.

The aim of the first part of this article (Experiments 1 and 2) is twofold. First, a forced-choice paradigm was used to establish the extent to which the Truth × Distance interaction can be attributed to difficulties in calculating inferences or, alternatively, to decision processes that operate independently of logical abilities or processing capacities. In this paradigm, participants were presented with problems containing two statements, one true and one false, and were required to identify the true one. According to the cognitive load hypothesis, the cognitive load associated with the resolution of a problem is a good indicator of its difficulty. Thus, a model for the evaluation of cognitive load involved in each problem is presented. It is assumed that the cognitive load depends on the number of inferential steps associated with both the true and false statements presented. This model should predict the relative difficulty of the problems presented and might, therefore, account for the Truth × Distance interaction.

The second part (Experiment 3) deals with the problem of interparticipant differences. The cognitive load hypothesis predicts that the interparticipant differences observed in reasoning tasks are due to differences in processing capacity. The participants performed a set-inclusion task and two tasks designed to assess their WM capacity. The cognitive load hypothesis enables us to make two predictions. First, the higher the participant’s WM capacity, the better his or her performance in the set-inclusion task. Second, the influence of the step-size effect on the rate of correct responses should be weaker, the higher the participant’s WM capacity. Indeed, participants with a high WM capacity should be less sensitive than others to the increase in step size and, therefore, to the increase in the cognitive load involved in the calculation of inferences.

Finally, we compare this hypothesis with Nguyen and Revlin’s (1993) proposal concerning interparticipant variability in the interpretation of the inclusion relation.

The Use of a Forced-Choice Paradigm

Carlson et al. (1992) have emphasized the paradox, for a theory of reasoning, of the necessity of accounting both for deductive abilities (Johnson-Laird, 1983; Rips, 1983) and for errors in tasks that are considered to be simple. Here, this paradox is all the greater in view of the fact that the inclusion relation structures much of the semantic memory of adult participants (Nelson, 1988; Rosch, 1978), that it has been considered to be an elementary logical relationship (Piaget & Inhelder, 1959), and that it is mastered at around the age of 10 or 11 years (Barrouillet, 1992; Bideaud, 1988; Houdé, 1989; Markman, 1978).

The results of the experiments cited above were all obtained using the same paradigm (true–false verification), in which participants evaluate the truth of statements presented one by one. However, the Truth × Distance interaction may be due to the fact that participants are more likely to evaluate statements as false the greater the number of inferential steps they involve, irrespective of whether these statements are true. The increase in this type of response has sometimes been interpreted as a failure on the part of the participants to consider the transitivity of the relation (Griggs, 1976). However, this hypothesis also predicts that an increase in the number of inferential steps will lead to a corresponding increase in the inexactitude of responses. The ability to discriminate between true and false statements should therefore decrease as the number of inferential steps increases. Now, it is possible to distinguish between the level of discrimination between true and false and the rate of the response false produced by the participants. For example, signal-detection theory (SDT) allows one to differentiate between the ability to distinguish noise from signal (for our purposes, between true and false statements), expressed as the value d’, and the decision criterion specified by participants, expressed as β, which determines the rate of the response false.

We have applied this type of analysis to the results obtained by Griggs (Griggs, 1976, Experiments 3 and 6, control groups). The frequency of correct responses was .832, .744, .690, and .723 for true statements involving 1, 2, 3, and 4 inferential steps, respectively, and .462, .585, .633, and .718 in the case of false statements. Although the values of d’, which measures the ability to discriminate between true and false, remained practically constant whatever the number of inferential steps (.87, .87, .84, and 1.17 for Step Sizes 1, 2, 3, and 4, respectively), the value of β increased with inferential distance (.63, .83, .94, and .99, respectively). Thus, an increase in the number of inferential steps would not appear to affect the ability to discriminate between true and false. It would, however, appear to modify the decision criterion applied by participants (who gave the response false with increasing frequency). In fact, the
frequency of the response false was .31, .42, .47, and .50 for 1, 2, 3, and 4 inferential steps, respectively.

These results support Griggs’s (1976) hypothesis: Participants are cautious about accepting remote statements. However, the fact that the ability to discriminate between true and false remains constant whatever the number of inferential steps involved in the statement to be evaluated suggests that the participants calculated the inferences and took account of the transitivity of the relation. This is compatible with the cognitive load hypothesis: Participants do indeed take account of the transitivity of the relation, but the increase in the number of inferential steps, and therefore in the cognitive load, leads to a greater number of false responses. However, given the use of the true–false verification paradigm, the Truth × Distance interaction might simply be due to a modification of the decision criterion (8).

One way of avoiding variations in β is to present a true and a false statement simultaneously and ask participants to decide which is the true statement. Two true statements involving a different number of inferential steps should, when compared with the equivalent false statements, be chosen with equal frequencies if the participants are able to calculate the inferences. We have used this type of double forced-choice paradigm in an earlier experiment (Barrouillet, 1989). Participants read a text containing three statements (AB, BC, and CD) and were asked to select the true statement from 36 pairs of statements containing one true statement (AB, BC, CD, AC, BD, and AD) and one false statement (BA, CB, DC, CA, DB, and DA). One may distinguish between four types of problems depending on the nature of the juxtaposed statements: true adjacent versus false adjacent (TA/FA), true adjacent versus false remote (TA/FR), true remote versus false adjacent (TR/FA), and true remote versus false remote (TR/FR). Two alternative hypotheses are possible. If, as the results of earlier studies would suggest, participants possess a symmetrical and intransitive conception of the inclusion relation, then they should choose the FA statement more frequently than the TR statement in TR/FA problems. In contrast, if participants are able to calculate the inferences, they should tend to choose the TR statement. The participants were children between 12 and 14 years of age. The results show that although the 12-year-old children opted for the FA statement, this preference disappeared at age 14 without, however, being replaced by a tendency to select the TR statement.

In Experiment 1 we replicated the earlier experiment but used adult participants. First of all, the use of the forced-choice paradigm should make it possible to establish the existence of the Truth × Distance interaction clearly and independently of any variation in the decision criterion. Secondly, the cognitive load hypothesis, together with the development observed between ages 12 and 14 and the results of the SDT analysis, suggests that adult participants should be able to calculate inferences. Therefore, they should choose TR rather than FA statements when faced with TR/FA type problems. At the same time, the ability to distinguish between true and false statements whatever the remote distance should lead to a level of correct responses higher than the random response level in problems juxtaposing two remote statements (TR/FR).

Experiment 1

Method

Participants. Thirty-six psychology students at the Université de Bourgogne aged between 18 years, 4 months and 23 years, 1 month (M = 19 years, 7 months) took part in the experiment in return for the award of course credits.

Materials and procedure. The participants were asked to learn a text containing three statements presented out of sequence (i.e., not in logical order, see Appendix A). As in the Fundalar text, extraneous information was inserted between the statements. This learning phase lasted approximately 10–15 min. When the participants considered that they had memorized the information contained in the text, they were given a book presenting one pair of statements per page, one being true and one false (e.g., All the square counters are varnished (all the varnished counters are red)). Thirty-six problems were formed from the six true statements resulting from the premises (three adjacent: AB, BC, and CD; three remote: AC, BD, and AD), and the six false statements were constructed by inverting these terms (BA, CB, DC, CA, DB, and DA). The 36 problems were subdivided into 9 TA/FA problems (e.g., AB/BC or CD/DB), 9 TR/FA problems (e.g., AC/CB or BD/DA), 9 TR/FR problems (e.g., AC/DB or BD/DA), and 9 TA/FR problems (e.g., AB/CD or CA/DB). The problems were presented in a random order that was the same for all participants and in which no statement was allowed to occur in 2 successive problems. In half of the problems the true statement was located at the top of the page. In the other half it was situated at the bottom. For any given problem, the true statement was situated at the top of the page for half of the participants and at the bottom of the page for the other half.

The participants were told that each problem contained one true statement, that is to say a statement that was contained in the text or that could be logically deduced from it, and one false statement, that is to say a statement that was not contained in the text and could not be logically deduced from it. They were asked to tick the box corresponding to the true statement. To make sure that the participants did not go back over their work, the experiment was performed in passes of only 2 participants. Because of the difficulty of the task and its duration (20–40 min), a single text was presented.

Results

A total of 36 problems were presented. The mean level of correct responses was 25.56 (71%). The level of success for the TA/FA, TA/FR, TR/FA, and TR/FR problems was 76%, 85%, 55%, and 69%, respectively (see Table 1). The mean number of correct responses to the TR/FR type problems (6.19 out of 9) was significantly higher than random response level, t(35) = 4, p < .01.

As predicted by our hypothesis, the participants appeared able to produce inferences and distinguish between true and false inferences in much the same way as when the problem juxtaposes two adjacent statements (mean level of success for TA/FA problems = 6.81). The difference in success levels between TA/FA- and TR/FR-type problems was not significant. Despite this, the level of correct responses obtained for TR/FA problems was no higher than the threshold for random responses (M = 4.92 out of 9), t(35) < 1. Our hypothesis that TR statements would be preferred to FA statements was
Discussion simply consist of recognizing sentences drawn from the text, storing the premises. In fact, although TA/FA problems can distinguish between true and false inferences. Second, the paradigm the level of correct responses for true statements decreased as inferential distance increased, our results show that adults are able to calculate these inferences because they can distinguish between true and false inferences. Second, the tendency to accept the symmetry of the inclusion relation persisted even when the problem juxtaposed two adjacent statements (TA/FA problems, 24% errors). This latter point is possible that the results obtained are due to the specific order in which the premises appeared in the text (AB-CD-BC) or to effects of a semantic nature relating to the chosen terms. The aim of Experiment 2 was to replicate the main facts identified in the preceding experiment while systematically controlling the effect of the order of premises as well as semantic effects by modifying the content of the premises. In line with the results observed for Experiment 1, the level of success in TR/FR problems should exceed random response level. When the true and false statements both involve one, two, or three inferential steps, the levels of success should be identical and higher than random response level. Finally, the level of correct responses should increase as the number of inferential steps involved in the true statement grows and diminish as the number of inferential steps involved in the true statement increases.

Experiment 2

Method

Participants. Forty-eight students in their first year of psychology at the Université de Bourgogne took part in this experiment. None of them had taken part in the earlier experiment.

Materials and procedure. The texts that were presented to the participants contained, as in the earlier experiment, three statements describing inclusion relations between four classes designated by A, B, C, and D (see Appendix A). The three statements linked adjacent terms in the inclusion hierarchy (all A are B, all B are C, and all C are D). To control for the effects of order, six types of texts were constructed in which the order of appearance of the premises was varied systematically. To control for the semantic effects, the four

Table 1

Mean Numbers of Correct Responses and Standard Deviations (in Parentheses) by Type of Problem in the Reasoning Task for the Three Experiments

<table>
<thead>
<tr>
<th>Experiment</th>
<th>n</th>
<th>TA/FA</th>
<th>TA/FR</th>
<th>TR/FA</th>
<th>TR/FR</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36</td>
<td>6.81 (1.80)</td>
<td>7.64 (1.55)</td>
<td>4.92 (2.73)</td>
<td>6.19 (2.51)</td>
<td>25.56 (7.44)</td>
</tr>
<tr>
<td>2</td>
<td>48</td>
<td>6.81 (2.00)</td>
<td>7.33 (1.97)</td>
<td>5.96 (2.49)</td>
<td>6.69 (2.51)</td>
<td>26.79 (7.44)</td>
</tr>
<tr>
<td>3</td>
<td>72</td>
<td>6.67 (1.88)</td>
<td>7.72 (1.51)</td>
<td>5.26 (2.70)</td>
<td>6.75 (1.92)</td>
<td>26.4 (6.77)</td>
</tr>
</tbody>
</table>

Note. The maximum number of correct responses per cell was nine. TA = true adjacent; FA = false adjacent; TR = true remote; FR = false remote.

*Not different from chance level (p = .05).
terms associated with the Classes A, B, C, and D (in this experiment, square, red, small, and pierced) were combined in four different permutations so that each term occupied each rank in the hierarchy once and the same succession of two terms was not found in any two sequences. Thus, each of the four permutations contained three premises or adjacent statements that were different in each case. Twenty-four texts were formed by multiplying the six sequences of the statements AB, BC, and CD by the four permutations of the terms square, red, small, and pierced.

Each of these 24 texts was presented to 2 participants. The procedure was identical to that used in the earlier experiments.

Results

We observed no effect of the permutation of terms or of the order of presentation of the statements in the text. There was also no effect of interaction on the level of success. The success levels for the four types of problems, TA/FA, TA/FR, TR/FA, and TR/FR, were 76%, 81%, 66%, and 74%, respectively (see Table 1). This again reflects the pattern of results obtained in Experiment 1. The mean number of correct responses differed from the level for random response for all four types of problems, t(47) = 8.03, 9.96, 4.05, and 6.04, respectively (p < .001 in all cases). This was not the case in the preceding experiment. The participants in this experiment, therefore, systematically chose the true rather than the false statement, even in problems of the TR/FA type and thereby testified to a transitive, nonsymmetrical conception of the inclusion relation. However, four of the six sequences of presentation of the statements may have contributed to this result in that they contained two consecutive statements that could be linked by transitivity (e.g., Sequence AB-BC-CD or CD-AB-BC), whereas the sequence selected in Experiment 1 avoided this type of arrangement (AB-CD-BC).

A 2 (true statement type: adjacent or remote) x 2 (false statement type: adjacent or remote) ANOVA with repeated measures on the two variables revealed a significant effect of the nature of the true statement, F(1, 47) = 10.58, p < .01, MSE = 2.55, and of the false statement, F(1, 47) = 5.68, p < .01, MSE = 3.30, p < .05. The interaction was not significant. These results reproduced those obtained in Experiment 1. Similarly, the number of errors increased with the step size of the true statement (21% errors for one step, 29% for two steps, and 32% for three steps, F(2, 94) = 6.14, p < .01, MSE = 0.817, and diminished as the step size of the false statement increased (29% errors for one step, 24% for two steps, and 19% for three steps, F(2, 94) = 5.60, p < .01, MSE = 0.819. Thus, even when participants achieved greater success and performed better than the threshold for random response, effects related to step size persisted. Moreover, the manipulation of the experimental material made it possible to demonstrate that the level of success remained more or less constant when the problems mobilized true and false statements containing the same number of inferential steps, as predicted by an analysis in terms of SDT. When the inferential distance was 1, 2, and 3 steps for both the true and false statements, the levels of success were 76%, 75.5% and 75%, respectively. These results show that the effects obtained in the earlier experiment were not due to a particular order of presentation of statements in the texts or to any semantic effect and prove that the difficulty participants experience in evaluating statements is linked to the number of inferential steps involved.

Discussion of Experiments 1 and 2

The use of the forced-choice paradigm has confirmed two of the main phenomena that had been revealed by earlier research. First, this paradigm confirms the existence of the Truth x Distance interaction by showing that the phenomenon does not depend solely on decision processes. The level of choice of true statements is an inverse function of the number of inferential steps they contain, even when these statements are opposed to the same set of false statements. In contrast, the level of rejection of false statements is a direct function of the number of inferential steps they contain, even when they are opposed to the same set of true statements. This provides confirmation of the difficulties adult participants experience when attempting to take account of the transitivity of the relation and to produce inferences. Second, the tendency to accept the symmetry of the relation persists because the problems that involve two adjacent statements cause high failure levels.

Moreover, the forced-choice paradigm confirms what an analysis in terms of SDT suggested. Participants are able to produce inferences because (a) in both experiments, the success level is significantly higher than the random response rate when a TR statement is opposed to an FR statement and (b) the success level remains approximately the same when the two statements involve the same number of inferential steps, whatever this number is (see Experiment 2).

Evaluation of the Cognitive Load Linked to Problem Solving

The results of the first two experiments have confirmed the existence of a Truth x Distance interaction effect. According to our hypothesis, the cognitive load associated with the calculation of a statement would depend on the number of inferential steps (Inf). We hypothesized that the probability P of a participant accepting a statement is an inverse function of the cognitive load and therefore of Inf: The greater the load, the lower this probability. We postulated that this relation is linear: The storage of a premise, or an additional mental model, would reduce by a constant amount the pool of attentional resources available for processing the problem. Of course, the ANDS (Rips, 1983) model takes account of both the number of statements stored and the probability that the rule of inference will be applied. However, we may be justified in thinking that this probability is particularly high in the case of the transitive inference considered here and that it only has a mild effect on the probability of a correct response. In fact, Dickstein (1978) observed correct responses 95% of the time in the resolution of syllogisms of the type all As are Bs, all Bs are Cs, therefore all As are Cs, which are akin to those we used.

The probability that the true statement will be chosen within the forced-choice paradigm is given by

\[ P(c) = \frac{P(T)}{P(T) + P(F)} \]

where P(c) represents the probability of a correct choice, P(T)
represents the probability of acceptance of the true statement, and \( P(F) \) represents the probability of acceptance of the false statement. When \( P(T) \) is equal to \( P(F) \), the two statements appear to be equally probable and the participant makes a random response, \( P(c) = .50 \). As \( P(T) \) tends toward zero, the false statement tends to be accepted and \( P(c) \) tends toward zero. As \( P(F) \) tends toward zero, \( P(c) \) tends toward one. If in Equation 1 we replace each probability by the terms of which it is a function, probability \( P(c) \) varies as an inverse function of a parameter \( L \) (for load) which is given by

\[
L = \text{Inf}_T/(\text{Inf}_T + \text{Inf}_F),
\]

where \( \text{Inf}_T \) and \( \text{Inf}_F \) represent the number of inferential steps involved in the true and false statements, respectively. The value of the parameter \( L \), which may vary between 0 and 1, was calculated for each problem. When \( L \) is close to 1, the problem is difficult. This occurs when the cognitive load associated with the calculation of the false statement (\( \text{Inf}_F \)) is low and that associated with the true statement is high. Thus the true statement, which is difficult to calculate, tends to be rejected and the false statement accepted. This is generally the case for TR/FA-type problems. In contrast, when the value of \( L \) approaches 0, the problem is easy. This occurs when the cognitive load associated with the calculation of the true statement is low and that associated with the false statement is high. The true statement is easy to calculate and is consequently accepted. This is generally the case for TA/FR-type problems. Given this analysis, the value of the parameter \( L \) is constant (.50) when the true and false statements involve the same number of inferential steps. Thus, problems that juxtapose two adjacent statements, two 2-step inferences or two 3-step inferences should all involve the same degree of difficulty (see Experiment 2).

Linear regression was performed using the parameter \( L \) as the independent variable and, as the dependent variable, the error frequency of each of the 36 problems for the entire set of participants in Experiments 1 and 2 (84 participants). Here, the value of \( L \) varies between .25 (\( \text{Inf}_T = 1, \text{Inf}_F = 3 \)) and .75 (\( \text{Inf}_T = 3, \text{Inf}_F = 1 \)). The linear tendency was significant, \( F(1,34) = 87.73, p < .001 \), and explained 72% of the variance observed in the error frequencies of the 36 problems. The error frequency increased with the value of the parameter \( L \) (see below for a similar analysis using a larger number of participants; see also Figure 1).

The suggested interpretation is based on a post hoc analysis of the results. However, this is a powerful model because there is no weighting of the parameters \( \text{Inf}_T \) and \( \text{Inf}_F \). Furthermore, it conforms to Rips's (1983) ANDS model and to the application of this model proposed by Carlson et al. (1992) as well as to Johnson-Laird and Byrne's (1991) theory of mental models. The difficulty of the problems therefore seems to be closely linked to the cognitive load associated with each of the two statements presented. This hypothesis leads us to predict the same success levels as would be expected using Potts's (1976) hypothesis concerning the treatment of the inclusion relation as a relation of similarity. However, our hypothesis does differ in one point. Potts (1976) suggested that judgments of similarity result from participants' tendency to order the categories mentioned in the text within a single dimension (similar to a linear sequence) on the basis of which they calculate the similarity. In our experiments, if distance effects are to appear, participants should be capable of establishing the correct logical sequence of categories and should therefore use transitivity correctly. Inferential distance effects would therefore be linked to an erroneous interpretation of a representation, in itself correct, of the premises contained in the text. According to this hypothesis, interparticipant differences would result from differences in the interpretation of the relation. In contrast, we hypothesize that they are due to differences of a functional order that relate to the ability to perform the calculations necessary for the deduction of the inferences. Thus, the two hypotheses lead to different predictions of the links between performance and processing capacity. Our hypothesis predicts that performance observed in the reasoning task will be correlated with participants' processing capacities, whereas the hypothesis that inferential distance effects are due to a problem in the interpretation of the relation predicts nothing of the sort.

At the beginning of this article, we suggested that the hypothesis that poor performance in reasoning tasks may be due to a limitation of processing capacity requires us to establish (a) that the cognitive load associated with the resolution of a problem is a good indicator of its difficulty and (b) that the interparticipant differences observed in reasoning tasks are due to differences in processing capacity. The data we have obtained is compatible with the first point. Analysis of the second point is the object of Experiment 3.

**Experiment 3**

The aim of this experiment was to verify the existence of a link between performances obtained in a deductive reasoning task and the participant's processing capacities. This capacity was evaluated using two tests designed in accordance with Baddeley's (1986, 1990) definition of WM. These tests used different contents. Research does, of course, exist that suggests that the performances obtained in tasks investigating WM capacity are independent of their contents (Turner & Engle, 1989). However, to prevent any possibility of a correlation between performances resulting from the specific contents of the task, we asked our participants to perform the so-called alphabet recoding (AR) task used by Kyllonen and Christal.
(1990) and Woltz (1988) and the reading span (RS) task devised by Daneman and Carpenter (1980, 1983). According to Kyllonen and Christal, the AR task is one of the most accurate measures of WM capacity. The RS task was also used in view of the fact that the reasoning task that our participants were asked to perform was based on premises presented in writing within a narrative text. Daneman and Carpenter (1980, 1983) have shown that this span is linked to the understanding of the text. In the first task, participants had to use an arithmetic operation (addition or subtraction) to transform a series of letters while considering the alphabetical order as a numerical order (e.g., $A + 1 = B, A + 2 = C$, etc.; the series $MBE$, when acted on by the operator $+ 2$, becomes $ODG$). In the second task, participants read aloud a series of sentences and had to remember the last word of each one in the correct sequence.

To perform these tasks, participants must store and maintain information in WM while performing another operation (calculation or reading). Baddeley’s (1990) model of WM holds that the maintenance of information in the articulatory loop while other operations are being performed by the central processor should rapidly lead to cognitive overload resulting from limitations to participants’ processing capacities. Success levels in these tasks are thus generally considered to be an indicator of WM capacity. A close link between WM capacity and performances in reasoning tasks has been observed by Kyllonen and Christal (1990). However, the reasoning task that most closely resembled ours appeared to be only loosely linked to WM capacity in the experiment conducted by these authors (correlation between syllogism-solving task and AR task: $r = .15$, $N = 392$; Kyllonen & Christal, 1990, p. 418). The aim of this experiment was, therefore, to verify that the performances observed for the deductive reasoning task used in the earlier experiments are linked to participants’ WM capacity as evaluated using the tasks presented above.

**Method**

**Participants.** Seventy-two first-year psychology students at the Université de Bourgogne took part in this experiment. None of them had participated in the earlier experiments.

**Materials and procedure.** The participants were asked to perform three tasks: a deductive reasoning task identical to that used in Experiments 1 and 2 and two tasks designed to evaluate their WM capacity (the AR task and the RS task). In the deductive reasoning task, the participants studied the same text as that used in Experiment 2 and had to resolve the 36 problems in which a true statement was juxtaposed with a false one. One week later, the participants were asked to perform the two tasks designed to evaluate their WM capacity.

**AR task.** Twenty-four random three-letter sets were formed and allocated at random to six operations ($- 3, - 2, - 1, + 1, + 2, + 3$), each operation receiving 4 sets. Certain letter sets were discarded, namely sets in which the result of the operation contained a letter from the initial set (e.g., $AFC + 2 = CHE$), sets containing two consecutive letters (e.g., $APB$), and sets that could not be calculated (e.g., $SPZ + 1$). The task was presented on-screen in a Hypercard environment. The letters appeared successively on screen for 1 s, followed by the operation (e.g., $+ 2$) for 2 s. When the participants saw a ? appear on the screen they had to perform the mental calculation and write their responses in a notebook. The participants were told that they had to calculate the entire response before writing it down (i.e., they could not convert the first letter and write down the result, then convert the second, etc.) Sets for which participants paused for more than 1 s while writing their responses were marked 0. The 24 sets were preceded by a 6-set training session. The pass was taken in the participants’ own time with participants pressing the mouse button to display the next problem. The sequence of set presentation was the same for all of the participants. The participants were given a break after the 12th set. One point was awarded for each correctly converted set (scores could range from 0 to 24).

**RS task.** Sixty sentences with unrelated contents containing between 13 and 15 words were formed and grouped into sets of 2, 3, 4, 5, or 6 sentences (three sets of each). The sentences were presented one by one in a Hypercard environment and read aloud by the participants who had to try to remember the final word. When the final word of the sentence was read, the experimenter clicked the mouse to replace the sentence on the screen with the following one. When the word Recall appeared, the participant had to say aloud the last word of each sentence in the order of appearance. The number of sentences per set increased progressively (i.e., three sets of 2 sentences, followed by three sets of 3 sentences). In order not to submit participants to a sequence of failures, sets consisting of $n$ sentences were not presented unless the participant had succeeded in at least one set consisting of $n - 1$ sentences. The test phase was preceded by six trial attempts. The score awarded to the participant equalled the total number of sentences contained in sets that were recalled in their entirety and in the correct order (e.g., a participant who recalled three 2-sentence sets, two 3-sentence sets, and one 4-sentence set would receive a score of $6 + 6 + 4 = 16$).

The order in which the two tasks were presented was alternated.

**Results**

The performances obtained in the reasoning task were comparable to those obtained in the preceding experiments. The levels of correct responses for problems of the type TA/FA, TA/FR, TR/FA, and TR/FR were 74%, 86%, 59%, and 75%, respectively (see Table 1). The mean number of correct responses differed from the level for random response for all four types of problems, $f(71) = 9.80, 18.08, 2.40$, and 9.99, respectively, $p < .05$ in all cases.

The mean number of correct responses for the AR task was $10.8$ ($SD = 4.9$). The mean number of correctly recalled sentences in the RS task was 14.0 ($SD = 5.2$). The correlation between the scores obtained in these two WM tasks was $.34$, $p < .01$. The number of correct responses in the reasoning task was correlated with both the AR score ($r = .31, p < .01$) and the RS score ($r = .35, p < .01$). These two scores (AR and RS) were standardized and then averaged ($M = 100, SD = 15$) to calculate a global WM score. The correlation between this score and the number of correct responses in the reasoning task was significant ($r = .41, p < .001$; see Table 2). This correlation was significant for the TR/FA problems ($r = .42, p < .001$), which were the most difficult; for the TR/FR problems ($r = .37, p < .01$); and for the TA/FA problems ($r = .31, p < .01$), but not for the TA/FR problems ($r = .20, p > .05$), which were also the easiest.

Those participants whose WM score was greater than the group mean by more than two-thirds of the standard deviation (i.e., more than 110, 18 participants) were considered as possessing a high WM capacity, whereas those whose score was lower than the group mean by more than two-thirds of the standard deviation (i.e., less than 90, 18 participants) were considered to have low WM capacity. The other 36 partici-
The effect of WM capacity was greater \(p < .001\). Note. TA = true adjacent; FA = false adjacent; FR = false remote; TR = true remote.

Another prediction concerned the link between WM capacity and problem type. The mean number of correct responses in the reasoning task was 30.2 (out of 36) for the high WM capacity participants, 26.6 for the medium WM capacity participants, and only 22.2 for the low WM capacity participants. More important, the interaction between WM capacity and problem type was significant, \(F(6, 26) = 3.94, p < .01, MSE = 1.63\). The effect of WM capacity was greater in the TR/FA problems (the most difficult) than in the TA/FR problems, which were the easiest (see Table 3).

As predicted by our hypothesis, the performances obtained in the reasoning task were thus linked with the participants' WM capacity, although this variable is not in itself sufficient to account for the observed performance (the \(R^2\) value was .16). Another prediction concerned the link between WM capacity and the strength of the step-size effect. This strength is evaluated in our paradigm by the slope of the regression line of the error frequency in the 36 problems on the values of L. A strong step-size effect results in a steep slope. Consequently, according to the cognitive load hypothesis, the lower the WM capacities, the steeper the slope of the regression line. This linear regression was computed for each of the three groups of participants classified in terms of WM capacity (high, medium, and low). The values of the slope were .346, .697, and 1.089 for the high, medium, and low WM capacity groups, respectively (see Table 3). Thus, performance was increasingly sensitive to the growth in the number of inferential steps, the lower the participants' WM capacity was.

Thus, the results confirmed the two predictions derived from the cognitive load hypothesis concerning the interparticipant differences in the reasoning task: (a) The higher the WM capacity of a given participant, the better his or her performance in the set-inclusion task, and (b) the higher the participant’s WM capacity, the weaker the influence of the step-size effect on the rate of correct responses.

Let us recall that our main hypothesis predicted (a) that the cognitive load associated with the resolution of a problem is a good indicator of its difficulty and (b) that interparticipant differences observed in the reasoning tasks are due to differences in processing capacity. To investigate the first point, a linear regression was performed using the value \(L\) and the error frequency in the 36 problems of all of the 156 participants to take part in the three experiments. This linear regression was significant: The value \(L\) explained 83% of the variance observed in the error frequency (see Figure 1).

An analysis in the light of the WM tasks (see Table 2) reveals that the AR scores were correlated with performance in problems requiring the calculation of inferences (i.e., TR/FA, \(r = .36, p < .01\), and TR/FR, \(r = .29, p < .02\)). The RS scores were correlated with performance in problems containing an FA or a TR statement (i.e., TR/FA, \(r = .33\), TR/FR, \(r = .32\), and TA/FA, \(r = .31\)).

It therefore appears that the AR score is a good predictor of the ability to produce inferences, whereas the RS takes account of participants' ability to store premises correctly, an ability that enables them to reject FA statements. The first of these tasks therefore seems to be associated with the implementation of transitivity, but the second could be linked to a rejection of the symmetry of the relation. This is confirmed if a transitivity index and an index of asymmetry are calculated for each participant’s comprehension of the relation. The first was calculated by evaluating the gap between the number of successes in problems involving TA statements (TA/FA and TA/FR) and the success level in problems involving TR statements (TR/FA and TR/FR). The value obtained indicates the extent to which the participant's performance deteriorates when changing from the evaluation of premises to the evaluation of true inferences. The asymmetry index was evaluated by counting the number of successful responses to

<table>
<thead>
<tr>
<th>Type of problem</th>
<th>TA/FA</th>
<th>TA/FR</th>
<th>TR/FA</th>
<th>TR/FR</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA/FA</td>
<td>6.83</td>
<td>7.86</td>
<td>5.44</td>
<td>6.47</td>
</tr>
<tr>
<td>TR/FA</td>
<td>6.11</td>
<td>7.11</td>
<td>3.39</td>
<td>6.11</td>
</tr>
</tbody>
</table>

Table 3

Mean Number of Correct Responses in Reasoning Task for Experiment 3 as a Function of Type of Problem and Working Memory (WM) Capacity Level, With Slopes and the \(R^2\) Values of Regression Lines From L Values to Error Rates for Each Group

<table>
<thead>
<tr>
<th>WM capacity level</th>
<th>Type of problem</th>
<th>Regression line</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TA/FA</td>
<td>TA/FR</td>
</tr>
<tr>
<td>High</td>
<td>7.39</td>
<td>8.06</td>
</tr>
<tr>
<td>Medium</td>
<td>6.83</td>
<td>7.86</td>
</tr>
<tr>
<td>Low</td>
<td>5.61</td>
<td>7.11</td>
</tr>
</tbody>
</table>

Note. The maximum number of correct responses per cell was nine. TA = true adjacent; FA = false adjacent; FR = false remote; TR = true remote.
problems that juxtapose a premise with its converse (AB/BA, BC/CB, and CD/DC). These two indices proved to be statistically independent \((r = .13, p > .25)\). The transitivity index was correlated with the scores obtained in the AR task \((r = -.31, p < .01)\) but not with the RS scores \((r = -.20, p > .10)\). In contrast, the asymmetry index was correlated with RS \((r = .31, p < .01)\) but not with the AR score \((r = .04)\). The tests of semipartial correlation coefficients (Cohen & Cohen, 1983) indicated that the transitivity index depended on AR scores, \(t(69) = 2.29, p < .05\), but not on RS, \(t = 0.87\), whereas the asymmetry index depended on RS, \(t = 2.79, p < .01\), but not on AR scores, \(t = 0.64\).

These results suggest that the difficulty of the task relates to two dimensions (i.e., transitivity and asymmetry) that are independent of one another and that could be managed by two distinct sets of cognitive processes. The first is devoted to the storage and maintenance of the information presented in the text (i.e., nature of the terms linked by the statements and direction of each relation), whereas the second would appear to be used for the calculation of inferences on the basis of this information. These facts confirm the hypothesis of a link between performance in the reasoning task and WM capacity. Furthermore, they reveal that the tendency, or lack of it, to consider the relation as symmetrical is partly linked to information storage capacity. The ability to produce inferences on the basis of this information is linked to the performance obtained in the AR task. These conclusions provide support for the hypotheses that can be derived from applying the models proposed by Carlson et al. (1992) or Johnson-Laird and Byrne (1991) to the task used here. In fact, both of these models presuppose a stage at which premises are stored (or mental models constructed) and a stage at which the premises are coordinated through the application either of a rule (Carlson et al., 1992) or of a heuristic for the coordination of the initial models (Johnson-Laird & Byrne, 1991). These facts also lend support to the hypothesis proposed by Griggs and Osterman (1980), according to which participants store premises one by one and then perform their calculations.

The data provided by our experiment suggest that symmetry errors are due not only to errors of interpretation but also to storage difficulties (i.e., first stage). The ability to reject FA statements is predicted by the RS, which is a good predictor of comprehension in reading (Daneman & Carpenter, 1983). These facts suggest that poor readers would be prone to accept FA statements because of the nontransitivity of the relation and accept FA statements because of the nontransitive and symmetrical interpretation of the relation, that is, by the interpretation that can be derived from this rule. Participants who use this rule would judge the true and false statements correctly, irrespective of whether they are adjacent and specified in the text, or remote.

Rule 2, known as the aggregate interpretation, does not include the possibility of transitive inferences and accepts the symmetry of the relation. This rule therefore corresponds to a wholly incorrect interpretation of the inclusion relation. Participants who use this rule would understand all A are B and all B are C as \(A = B\) and \(B = C\). This type of interpretative rule allows participants to produce transitive inferences, and, therefore, to evaluate TR statements correctly, but will also lead to the acceptance of false statements, whether adjacent or remote, because the relation is interpreted as symmetrical (i.e., for these participants, all A are B means all A are B, and all B are A, because \(A = B\)).

Rule 3, termed the aggregate interpretation, does not include the possibility of transitive inferences and accepts the symmetry of the relation. This rule therefore corresponds to a wholly incorrect interpretation of the inclusion relation. Participants who use this rule would represent all A are B and all B are C as \(A = B\) and \(B = C\) with \(B_1 B_2\). They see no apparent connection between one category pair and another and no relation holds between A and C. Given such a rule, remote relations, whether true or not, are seen as describing unlinked categories. A participant using this rule would reject the remote statements (true or false) because of the nontransitivity of the relation and accept FA statements because of the perceived symmetry of the relation. It may be noted that the Truth × Distance interaction is usually explained in terms of a nontransitive and symmetrical interpretation of the relation, that is, by the interpretation that can be derived from this rule.

Thus, the hypothesis of the use of varied interpretative rules situates interparticipant differences at the level of the comprehension of the relation in the same way as that proposed by Potts (1976) or Newstead and Griggs (1984). In fact, the hypothesis proposed by Nguyen and Revlin (1993) is that the use of Rule 3 by the majority of participants in earlier experiments (in particular those that use the Fundalas text as
the basis for deductions) is due to the fact that the denomination of the categories does not allow participants to identify a common superordinate category. The participants would therefore interpret the relations that are presented in the text as referring to distinct and unrelated categories (fundalas, outcasts, hill people, farmers, and peace lovers). In contrast, when the text contains convergent categories that clearly refer to a single superordinate set (e.g., officers in the Leptus Party, officers in the Terran Coalition, officers in the Rockite Party), participants should interpret inclusion relations correctly. We turn our attention later to the fact that even when the text explicitly presents categories that refer to a single superordinate category, as is the case here (i.e., the pieces in a game), we once again encounter the main phenomena observed in connection with the Fundalas passage.

In our opinion, the differences observed in performance were caused by the functional aspects of task resolution. In other words, these differences would not imply that participants interpret the relation that confronts them differently. Instead, they may reflect a varying ability to store information and process it using operations that may be common to all participants. If this is the case then there should exist differences in the level of processing ability that correspond to the supposed uses of interpretative rules.

Use of Interpretative Rules or Processing Capacities?

It is possible to establish response patterns corresponding to the supposed use of one of the three above-mentioned rules in the forced-choice task (see Appendix B for the criteria applied). Of the 72 participants in Experiment 3, 76% could be classified as users of one or other of these rules (29 participants for Rule 1, 14 for Rule 2, and 12 for Rule 3). The mean global WM scores were 91.8, 93.7, and 105.4 for the users of Rules 3, 2, and 1, respectively. WM capacities are therefore higher the more closely the applied "rule" corresponds to the properties of the inclusion relation, \( F(2, 26) = 9.4, p < .001, \text{MSE} = 116.43 \). This result may be complemented by a qualitative analysis. The rule (1, 2, or 3) that is thought to be used is linked to the level of WM capacity (high, medium, or low), \( x^2 = 16.23, p < .005 \) (see Table 4). The high WM capacity participants were mainly classified as Rule 1 users (none of them used Rule 3), whereas the low WM capacity participants were mainly classified as Rule 3 users and only 2 of them used Rule 1.

If we are to preserve the compatibility of the hypothesis that different participants use different interpretative rules with the data provided by our experiments, we must assume that the way participants interpret the relation is linked to their WM capacities. In our opinion, these data can be explained more satisfactorily if we consider that the level of success in the reasoning task is a function of the quality of storage of the premises in short-term memory and the ability of participants to perform calculations using these data. The quality of storage and the precision of the calculations would be a function of the participant's WM capacity.

One of the predictions resulting from such a viewpoint is that all of the participants should be affected by the same task parameters, for example by inferential distance. According to our hypothesis, performance should be increasingly sensitive to the effects of inferential distance. The more difficulty participants experience in storing and coordinating the premises, the lower their available WM capacity. In contrast, the hypothesis advanced by Nguyen and Revlin (1993) specifies that the participants who display the poorest performance use an interpretative rule that excludes transitivity (i.e., Rule 3). In consequence, performance is independent of inferential distance. Of course, the two hypotheses agree on the fact that the participants who are thought to use Rule 3 perform worse on remote statements than on adjacent statements when they are true and the opposite when they are false. However, our hypothesis also predicts that participants will perform worse on three-step TR statements than on two-step TR statements, whereas Nguyen and Revlin’s hypothesis makes no such prediction because these participants are not thought to implement transitivity.

This hypothesis was tested on the 19 participants from all three experiments who could be classed as Rule 3 users in accordance with Nguyen and Revlin’s (1993) criteria. As predicted by our hypothesis, these participants committed significantly more errors in problems containing three-step TR statements, such as \( AD \) (\( M = 4.47, \text{maximum} = 6 \)) than in problems containing two-step TR statements, such as \( AC \) or \( BD \) (\( M = 3.63, \text{maximum} = 6 \), \( F(1, 18) = 10.31, \text{MSE} = 0.65, p < .005 \)). Similarly, they committed significantly fewer errors in problems containing three-step FR statements, such as \( DA \) (\( M = 1.37 \)) than in problems containing two-step FR statements, such as \( CA \) or \( DB \) (\( M = 2.37 \), \( F(1, 18) = 16.68, \text{MSE} = 0.57, p < .001 \)). Thus, those participants who, according to Nguyen and Revlin, use interpretative Rule 3 do in fact take account of the transitivity of the relation. The fact that they almost systematically choose the FA statement when this is juxtaposed with a TR statement (in at least 78% of cases) does not mean that they fail to take account of the transitivity of the relation and do not calculate the inferences. Instead it suggests that they experience difficulties in the calculation and that these difficulties are greater the more inferential steps involved in the remote statement that is to be evaluated.

Our hypothesis may be challenged by the hypothesis proposed by Potts (1978), according to which the effects of inferential distance are due to an evaluation of similarity on an ordered, one-dimensional scale (i.e., \( A-B-C-D-E \)). The higher the distance separating the terms, the lower their similarity. Now, we have shown that Rule 3 users are sensitive to the step-size effect. It would then be necessary to assume that the participants who use Rule 3 have constructed such a scale (the statements are presented out of order in the text) and have therefore taken account of the transitivity of the relation. Our

<table>
<thead>
<tr>
<th>WM capacity level</th>
<th>Interpretative rule</th>
<th>Undefined</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Medium</td>
<td>16</td>
<td>6</td>
</tr>
<tr>
<td>Low</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>
interpretation does not exclude effects due to divergent interpretations of the relation. We simply maintain that it is very improbable that the performances observed over a period of two decades in hierarchical inclusion tasks are due to a nontransitive conception of the relation on the part of certain participants. We suggest that all of the participants take account of the transitivity of the relation to produce inferences and that the interparticipant differences are due to differences in the ability to perform the associated calculations.

We have seen that the participants categorized as Rule 3 users were sensitive to the effects of inferential distance. The same is true of the 46 participants who may be classified as Rule 1 users if we apply a strict classification criterion (i.e., no more than two errors out of nine whatever the problem type: TA/FA, TA/FR, TR/FA, or TR/FR). The error frequencies for the 36 problems are a function of the value $L$, $F(1, 34) = 4.88, p < .05, R^2 = .13$, whereas these participants commit only 5.1% errors when confronted with TR/FA-type problems. The increase in the number of errors as a function of the cognitive load, which is itself a function of the inferential distance, is similarly observed in the 29 participants who may be classified as Rule 2 users, $F(1, 34) = 32.08, p < .001, R^2 = .49$. Therefore, inferential distance has an effect on error frequencies, whatever the performance level of the participants.

Another way of demonstrating the generality of this effect is to group together the participants of the three experiments as a function of the number of errors committed (Table 5). For all performance levels, the frequency of errors in the 36 problems is a function of the value $L$, whether the participants have committed 5 errors or more than 20.

In other words, whatever the level of performance, or the rule that provides an a posteriori explanation of the participant’s behavior, performances are affected by inferential distance. Whatever hypothesis one adopts to account for this phenomenon, it must exclude the possibility that participants fail to recognize or fail to take account of the transitivity of the relation, given the method of presenting the premises. In fact, there is nothing in the results obtained to indicate that different levels of performance are based on different procedures for task resolution.

### General Discussion

The aim of the series of experiments presented here was to clarify the problems linked with the interpretation of the results obtained in so-called hierarchical inclusion tasks and, in particular, to investigate the existence of an interaction between inferential distance and the truth value of statements (Truth $\times$ Distance interaction). The literature habitually accounts for this phenomenon by supposing that participants interpret the inclusion relation erroneously or make use of an inappropriate schema for the representation of the premises.

We have advanced the hypothesis that all of the observed phenomena can be explained if we assume that the difficulty presented by these problems is a function of the cognitive load associated with their resolution. Interparticipant differences in performance would then be due to differences in participants’ processing capacities. The introduction of the forced-choice paradigm not only confirms the existence of the phenomena observed within the traditional true–false verification paradigm but also makes it possible to demonstrate that participants are able to take account of transitive inferences because, in all of the experiments, the level of correct responses to the TR/FR problems was higher than random response level. Despite this, it would appear that the effects associated with inferential distance persist. Moreover, our paradigm reveals the persistence of errors in connection with TA statements that are often the object of conversion errors.

The hypothesis that the observed errors were due to a problem of cognitive load and processing capacity was evaluated by asking a group of participants to undertake a hierarchical inclusion task and WM span tasks. A link appeared between the two sets of tasks. In general, the participants who achieve the best results are those with the highest WM spans. Moreover, the relative difficulty of reasoning problems is a function of the cognitive load required for their resolution, provided that we accept that this cognitive load is a linear function of the number of premises necessary for the evaluation of a statement.

Our explanatory hypothesis differs from earlier hypotheses in that it assumes that the difficulties encountered by adults

<table>
<thead>
<tr>
<th>$L$ value</th>
<th>Step size</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>.25</td>
<td>1</td>
<td>3</td>
<td>.01</td>
</tr>
<tr>
<td>.33</td>
<td>1</td>
<td>2</td>
<td>.02</td>
</tr>
<tr>
<td>.40</td>
<td>2</td>
<td>3</td>
<td>.02</td>
</tr>
<tr>
<td>.50</td>
<td>1</td>
<td>1</td>
<td>.03</td>
</tr>
<tr>
<td>.50</td>
<td>2</td>
<td>2</td>
<td>.03</td>
</tr>
<tr>
<td>.60</td>
<td>3</td>
<td>3</td>
<td>.00</td>
</tr>
<tr>
<td>.67</td>
<td>2</td>
<td>1</td>
<td>.05</td>
</tr>
<tr>
<td>.75</td>
<td>3</td>
<td>1</td>
<td>.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Performance level (number of errors out of 36)</th>
<th>0-4</th>
<th>5-9</th>
<th>10-14</th>
<th>15-19</th>
<th>&gt;20</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>.25</td>
<td>.01</td>
<td>.09</td>
<td>.14</td>
<td>.18</td>
<td>.44</td>
<td>.12</td>
</tr>
<tr>
<td>.33</td>
<td>.02</td>
<td>.09</td>
<td>.18</td>
<td>.33</td>
<td>.53</td>
<td>.17</td>
</tr>
<tr>
<td>.40</td>
<td>.02</td>
<td>.15</td>
<td>.26</td>
<td>.33</td>
<td>.58</td>
<td>.21</td>
</tr>
<tr>
<td>.50</td>
<td>.03</td>
<td>.20</td>
<td>.31</td>
<td>.44</td>
<td>.54</td>
<td>.25</td>
</tr>
<tr>
<td>.50</td>
<td>.03</td>
<td>.14</td>
<td>.31</td>
<td>.44</td>
<td>.64</td>
<td>.25</td>
</tr>
<tr>
<td>.60</td>
<td>.00</td>
<td>.15</td>
<td>.22</td>
<td>.38</td>
<td>.92</td>
<td>.23</td>
</tr>
<tr>
<td>.67</td>
<td>.05</td>
<td>.24</td>
<td>.53</td>
<td>.67</td>
<td>.77</td>
<td>.38</td>
</tr>
<tr>
<td>.75</td>
<td>.07</td>
<td>.35</td>
<td>.53</td>
<td>.66</td>
<td>.73</td>
<td>.40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$n$</th>
<th>48</th>
<th>27</th>
<th>36</th>
<th>32</th>
<th>13</th>
<th>156</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>.36***</td>
<td>.57***</td>
<td>.75***</td>
<td>.71***</td>
<td>.29***</td>
<td>.83***</td>
</tr>
</tbody>
</table>

***$p < .001$. |
are the result not of conceptual problems (faulty comprehension and interpretation of the inclusion relation) but of functional problems. We believe that we have demonstrated that, contrary to the hypothesis proposed by Nguyen and Revlin (1993), the habitually observed results are not due to a symmetrical, nontransitive interpretation of the relation on the part of a significant number of participants. Nevertheless, the work of these authors is extremely interesting in that they show that a minimal modification to the way the relation is formulated has a major impact on performance. When class-inclusion relations are expressed so as to emphasize a commonality among the categories (e.g., officers in the Leptus Party, officers in the Terran Coalition), participants tend to draw the appropriate inferences.

However, we do not think that the improvements that Nguyen and Revlin (1993) observed are due solely to the use of categories that the participants clearly recognize as convergent. Instead, we believe that the texts we have used enable participants to see that the categories all belong to a single superordinate category (i.e., the pieces in a game). An analysis of the tactics used to assist participants in Nguyen and Revlin's experiment reveals that they presented all inclusion relations by means of the copula in (e.g., "all members in the Leptus Party are in the Terran Coalition," "all members in the Fundala Party are in the Outcast Coalition," "all Lup are in Biv"). It is possible that this use of in induced the participants to perform a spatialized encoding of the premises that facilitated the resolution of the problem (Mynatt & Smith, 1979). Participants would have then stored the premises in a spatialized linear order in accordance with Potts's (1976) hypothesis.

Potts's (1976) hypothesis, which holds that participants interpret the inclusion relation as a similarity relation, also accounts for the majority of the observed results and, in particular, the Truth × Distance interaction. It may also agree that performance is linked to WM span. Participants with a large WM span would be better able than the others to construct a correctly integrated representation (i.e., the ordered scale A-B-C-D-E) of the set of premises on the basis of the relations that will subsequently be evaluated. However, this hypothesis does not account for the important fact that the correct identification of TA statements and the calculation of inferences appear to constitute two distinct processes. In effect, if Potts's hypothesis is correct, an exact representation of the premises should lead both to the identification of adjacent statements and the extraction of the inferences that are "read" from this representation. Thus, the scores obtained in the four types of problems should all be correlated with one or other of the WM tasks because, in each case, their solution would be based on the same process. However, we have seen that RS, which functions as a predictor of the ability to identify premises and reject FA statements, was not correlated with the transitivity index. In contrast, performance in the AR task, which predicts the ability to produce inferences (correlation between AR and transitivity index: $r = -0.31$), was independent of success in problems that contained only adjacent statements (correlation between the AR task and TA/FA problems: $r = 0.20$).

Chronometric analyses should make it possible to decide. Our hypothesis supposes that the time required for the verification of a statement increases with the number of inferential steps, whereas Potts's (1976) hypothesis predicts the opposite effect, which is habitually observed in the processing of linear sequences. In fact, Potts seems to confirm these predictions. However, Griggs and Osterman (1980) make no response to these results and maintain their hypothesis that participants store statements in an isolated way. They conclude that these differences may be due to the different problem-solving strategies implemented by the participants.

In fact, the presence of inferential distance effects in all of our participant groups, whatever their level of performance, does not in itself mean that they all use the same strategy. It is possible that the participants who achieve greatest success use an ordered linear representation of the terms and that the others store the premises in an isolated way before then attempting to perform the calculations. In both cases, the effects due to inferential distance would be the same. Similarly, one might imagine that the participants who possess the greatest processing capacity are better able than the others to construct and maintain an integrated linear representation of the premises. In such cases the reaction time (RT) patterns would be distinct and a function of the participants' performance levels. The RTs of the participants who achieve the best performances should decrease as the number of inferential steps increases (just as was observed in connection with linear sequences), whereas the RTs of the poorest performers should lengthen as the number of inferential steps increases. An interparticipant diversity of RT patterns has also been identified by Griggs and Osterman (1980). At the same time, WM capacities alone can in no way explain all the variance observed in the reasoning task. The hypothesis that a variety of problem-solving strategies may exist cannot be discarded.

Griggs and Osterman (1980) concluded from their work that the most interesting approach to the study of the hierarchical inclusion task is an analysis of the factors responsible for observed differences between participants. Our study has revealed that one of these factors is the participants' processing capacities. Our approach allows us to account simultaneously for (a) interparticipant differences and (b) the principal phenomena observed in hierarchical inclusion tasks. First, the limited capacity of WM accounts for the effects of inferential distance and the Truth × Distance interaction that has been the object of so many descriptions. The conclusion drawn by Toms, Morris, and Ward (1993) from their study of the role of WM in conditional reasoning applies to the set-inclusion task: "Utilization of working-memory resources is an important factor in determining problem difficulty" (p. 697). Interparticipant differences in WM capacity also account in part for the differences in the performance of different individuals. Thus, transitive inference tasks based on inclusion relations would appear to count among those cognitive domains about which Just and Carpenter (1992) wrote, "One implication of capacity theory is that some of the performance differences among individuals within a task domain will be explained in large part in terms of working memory capacity" (p. 143).

However, the results obtained in the present experiments may not exclude the hypothesis that a variety of problem-solving strategies exist. Some participants may arrange the
terms in an ordered series that they use to evaluate relations. Others may calculate the inferences on the basis of individually stored premises. The same inferential distance effect may then result from two different strategies.

References


Appendix A

Experimental Texts

Text Used in Experiment 1

Paul has built a new board game. He has made different counters in order to be able to play with his friends. You can see that all the wooden counters are square. Not all the counters have the same color and all the red counters are varnished. Of the counters, some are square, some are round and yet others are triangular. All the square counters are red.

Example of Text Used in Experiments 2 and 3

As a present, Jean has been given a building toy made of parts of different shapes, sizes, and colors. Some of the parts are pierced so that they can be assembled and all the square parts are pierced. When he looks at his toy he notices that all the pierced parts are red. He also notices that all the small parts are square.

Appendix B

Criteria Used to Assign Nguyen and Revlin’s (1993) Rule 1, 2, or 3 to the Participants

The Rule 1 proposed by Nguyen and Revlin (1993) stated that participants consider the relation as transitive and asymmetric. Therefore, Rule 1 users would obtain a high score in each type of problem (i.e., TA/FA, TA/FR, TR/FA, TR/FR).

For Rule 2 users, the relation is transitive but symmetrical. These participants would accept both the FA statements (symmetry) and TA statements, and all the remote statements. This led to a random response in each type of problem because all of the possible statements appear to be true. Therefore, Rule 2 users would obtain a medium score in each type of problem.

For Rule 3 users, the relation is intransitive and symmetrical. So, these participants would accept all adjacent statements whether true or false and would reject remote statements. Therefore, Rule 3 users would obtain a high score in TA/FR problems, a medium score in TA/FA and TR/FR problems, and a low score in TR/FA problems.

For each type of problem (TA/FA, TA/FR, TR/FA, TR/FR), a score of 7 or more out of 9 was considered high, whereas a score of 2 or less was considered low. A score from 3 to 6 was considered medium.

These criteria could be summarized as follows:

A strict application of these criteria led to isolating only 6, 5, and 22 participants who used Rules 3, 2, and 1, respectively, in Experiment 3. So, weaker criteria had been used in which one cell of the table could be slightly different from the strong criteria (i.e., a high score could be medium, a medium score could be high or low). The L cell in Rule 3 criterion could not be changed because the M-M-M-H pattern results in a Rule 2 pattern.

Received February 10, 1995
Revision received February 21, 1996
Accepted March 5, 1996