COGNITIVE PSYCHOLOGY 25, 524–575 (1993)

# The Roles of Similarity in Transfer: Separating Retrievability from Inferential Soundness

DEDRE GENTNER

*Northwestern University*

MARY JO RATTERMANN

*School of Communications and Cognitive Science, Hampshire College*

AND

KENNETH D. FORBUS

*Department of Computer Science, Northwestern University*

Similarity is universally acknowledged to be central in transfer, but recent research suggests that its role is complex. The present research attempts to isolate and compare the determinants of similarity-based *access* to memory and the determinants of the subjective *soundness* and *similarity* of a match. We predicted, based on structure-mapping theory, that subjective soundness would depend on the degree of shared relational structure, particularly higher-order structure such as causal bindings. In contrast, we predicted that memory retrieval would be highly sensitive to surface similarities such as common object attributes. To assess *retrievability*, in three studies, subjects were asked to read a large set of stories and were later given a set of probe stories that resembled the original stories in systematically different ways; e.g., purely relational analogies, surface-similarity matches, or overall (literal similarity) matches. Subjects were told to write out any of the original stories that came to mind. To assess subjective *soundness*, independent subjects (and also the same reminding subjects) were

This research was supported by the Office of Naval Research under Contract N00014-89-J-1272 and by the National Science Foundation under Grant BNS-9096259. The first experiment was carried out by Russell Landers as a senior honors project in the MIT Psychology Department in 1984. We thank Cathy Clement, Brian Falkenhainer, Rob Goldstone, Keith Holyoak, Laura Kotovsky, Art Markman, Doug Medin, Brian Ross, Bob Schumacher, Colleen Siefert, and Janice Skorstad for discussions of these issues and comments on prior drafts of this paper. We also thank Ed Bowden for assistance on Experiment 4 and Verna Holland and Andy Stevenson for editorial assistance. Correspondence and reprint requests concerning this paper should be directed to Dedre Gentner, Department of Psychology, Northwestern University, 2029 Sheridan Rd., Evanston, IL 60208-2710. The materials can be obtained by writing Mary Jo Rattermann, Hampshire College, CCS Adele Simmons Hall, Amherst, MA 01002.



0010-0285/93 $9.00

asked to rate the inferential soundness of each pair; i.e., how well inferences true of one story would apply to the other. As predicted, subjective soundness was highly related to the degree of common relational structure, while retrievability was chiefly related to the degree of surface similarity. Ratings of the similarity of the pairs did not predict the retrievability ordering, arguing against the possibility that the retrieval ordering simply reflected overall similarity. Further, a fourth study demonstrated that subjects given a forced-choice recognition task could discriminate between possible matches on the basis of relational structure, ruling out the possibility that the poor relational retrieval resulted from forgetting or failing to encode the relational structure. We conclude that there is a dissociation between the similarity that governs access to long-term memory and that which is used in evaluating and reasoning from a present match. We describe a model, called MAC/FAC ("Many are called but few are chosen"), that uses a two-stage similarity retrieval process to model these findings. Finally, we speculate on the implications of this view for learning and transfer. 

Analogy and similarity are central in learning and transfer. People solve problems better if they have solved prior similar problems (e.g., Anderson, Farrell, & Sauers, 1984; Holyoak & Koh, 1987; Novick, 1988; Pirolli, 1985; Reed, 1987; Ross, 1987, 1989) and these benefits extend to children as well as to adults (Brown & Kane, 1988; Chen & Daehler, 1989; Gholsen, Eymard, Long, Morgan & Leeming, 1988). One of the most enduring findings in the field is that similarity promotes reminding and transfer (Ellis, 1965; Osgood, 1949; Thorndike, 1903).

However, although the relationship between similarity and transfer is strong, it is not simple. Recent research has confirmed the venerable finding that transfer increases with similarity (e.g., Anderson, Farrell, & Sauers, 1984; Holyoak & Koh, 1987; Novick, 1988; Pirolli, 1985; Reed, 1987; Ross, 1987, 1989; Simon & Hayes, 1976) but has also brought home the complexity of similarity's role in transfer. Three findings have emerged. First, accuracy of transfer depends critically on the degree of structural match—e.g., match in causal structure (Schumacher & Gentner, 1988a,b; Holyoak & Koh, 1987) or in the principle applied (Novick, 1988; Ross, 1984, 1987). Second, people often fail to access structurally appropriate materials, even when such materials are present in long-term memory. In Gick and Holyoak's (1980, 1983) insightful series of studies, subjects were given Duncker's (1945) radiation problem: how can one cure an inoperable tumor when enough radiation to kill the tumor would also kill the surrounding flesh? The solution, which is to converge on the tumor with several weak beams of radiation, is normally discovered by only about 10% of the subjects. But if given a prior analogous story in which soldiers converged on a fort, three times as many subjects (about 30%) produced the correct answer. This indicates that spontaneous analogical transfer can occur to good effect. But it also reveals a limitation, for the great majority of the subjects failed to retrieve the analogous story.

Yet, when given a general hint that the story might be relevant, about 75% solved the problem, indicating that the failure was not of storage but of access.[1]

Third, similarity-based remindings are often based on superficial commonalities as well as, or instead of, structural commonalities (Holyoak & Koh, 1987; Novick, 1988; Reed, Ernst, & Banerji, 1974). For example, Ross (1984) taught subjects six probability principles in the context of different story lines (e.g., a drunk and his keys). Subjects were then tested on problems whose story line was similar to that of the appropriate principle, similar to that of an inappropriate principle, or unrelated to any of the study problems. Compared to the unrelated baseline, surface similarity in story lines improved performance when linked to an appropriate principle and hurt performance when linked to an inappropriate principle. In later research, Ross (1989) found that similarities in story lines had a large influence on the probability of accessing the prior problem, but little influence on the probability of correctly applying the principle.

Such findings lead inescapably to the conclusion that a finer-grained analysis of similarity is necessary to capture its role in transfer. Our research is directed at three simple questions: (1) what kinds of similarity do people think constitute a good match?; (2) what kinds of similarity promote access to long-term memory? and (3) how do the answers to (1) relate to the answers to (2)? We first lay out the theoretical distinctions necessary to proceed. Then we present four experiments[2] and relate their findings to a simulation of similarity-based access and inference.

To achieve a sufficiently precise vocabulary, we must decompose *similarity* into different classes and we must decompose *transfer* into multiple subprocesses. We begin with similarity. Our decomposition of similarity uses the distinctions of structure-mapping theory (Gentner, 1980, 1983, 1989a). First, *analogy* is defined as similarity in relational structure; more specifically, analogy is a one-to-one mapping from one domain representation (the base) into another (the target) that conveys that a system of relations that holds among the base objects also holds among the target objects independently of any similarities among the objects to which those relations apply.[3] An important assumption is that a further selection

[1] It has been pointed out (e.g., Holyoak & Thagard, 1989a; Keane, 1985; Wolstencroft, 1989) that this analogy is structurally imperfect, since the general's goal is to protect his troops from their surround (the populace) whereas the surgeon's is to protect the surround (the flesh) from the rays. However, Keane's replication, which addressed this issue, produced much the same results.

[2] Experiments 1, 2, and 3 have been presented at conferences (Gentner & Landers, 1985; Rattermann & Gentner, 1987) but have never been fully published. One purpose of the present paper is to present these findings completely.

[3] This represents an ideal competence model of analogy—a description at Marr's (1982)

TABLE 1
Examples of Different Similarity Types

| | |
|---|---|
| **Literal Similarity** | Milk is like water |
| **Analogy** | Heat is like water |
| **Mere Appearance (Surface Similarity)** | The glass tabletop gleamed like water |

is made: among the common relations that could participate in the analogy, people prefer to focus on interconnected systems of relations (the *systematicity principle*). That is, they prefer to map sets of relations that include common higher-order relations that constrain the common lower-order relations. Given these distinctions, we can distinguish further classes of similarity according to the kinds of predicates that are shared. In *analogy*, only relational predicates—low-order and higher-order—are shared. In *literal similarity* (overall similarity), both relational predicates and object-attributes are shared. In *surface matches* (or *mere-appearance matches*), only object-attributes and low-order relations are shared. Table 1 gives examples of these different kinds of similarity. Although these distinctions are continua, not dichotomies, it is nonetheless useful to lay out the dimensions.

The canonical situation for similarity-based transfer is that a person has some current topic in working memory and is reminded of some similar situation stored in long-term memory. There is general consensus that similarity-based transfer involves (1) *accessing* a prior potential analogy, (2) *matching* the prior (base) analog with the target, (3) *mapping* further inferences from the base to the target, (4) (possibly) *adapting* the inferences to fit the target, (5) *evaluating* the soundness of the analogy, and (6) (possibly) *extracting the common structure* for later use (J. Clement, 1986; Gentner, 1988a; Gentner & Landers, 1985; Hall, 1989; Holyoak & Thagard, 1989b; Keane, 1985; Kedar-Cabelli, 1988; Novick, 1988; Thagard, 1988; Winston, 1982). On this account, similarity-based transfer involves from four to six subprocesses: *accessing, matching, evaluating, drawing inferences,* and often *adaptation* of the analogy and *abstraction* from the analogy.

This brings us to the core claims of this paper. We suggest that different subprocesses utilize different kinds of similarity. Specifically, structure-mapping theory predicts that evaluations of the *soundness* of a match should be based on the degree of overlap in relational structure. In contrast, we predict that similarity-based *access* to *long-term memory* will be

"computational level" or Palmer and Kimchi's (1985) "informational-constraints level." For a discussion of performance factors in analogical processing, see Gentner (1989a) or Schumacher and Gentner (1988a,b).

much more influenced by surface similarity than by overlap in relational structure. There will thus be a dissociation between the similarity types used at different subprocesses of transfer. After presenting the empirical evidence, we describe a simulation called MAC/FAC ("Many are called but few are chosen") that uses a two-stage similarity retrieval process to model these theoretical claims.

This prediction contrasts with two prominent positions on memory retrieval. First, the case-based-reasoning view states that items are indexed in memory via shared abstractions such as common goals and causal structures (e.g., Hammond, 1989; Keane, 1988; Schank, 1982). On this view, common higher-order structure should be the best retrieval probe. The second contrasting position, implicit in many psychological models of memory retrieval, is that there is a unitary notion of similarity on which transfer depends (e.g., Gillund & Shiffrin, 1984; Hintzman, 1984). In this case, the best predictor of retrievability should be overall similarity.

What do we know so far concerning the specificity of similarity types with respect to transfer subprocesses? Recent research in problem-solving and transfer has gone well beyond the stage of simply studying overall similarity in transfer and has made progress in tracing the details of the transfer processes. For example, Ross (1989), in a study of transfer in problem-solving, measured not only the proportion of correct solutions, but also the proportion of remindings, as measured by whether subjects wrote out a prior formula, regardless of whether they solved the problem correctly. This allows a contrast between *solution* rate—a measure which presumably includes mapping, adaptation, evaluation, and drawing inferences–and *reminding* rate. Ross found that reminding rate was relatively strongly affected by surface similarity and solution rate by structural similarity. Novick (1988) gave novice and expert mathematicians study problems that included both surface-similarity distractors and remote analogs, followed by a later target problem. Although both experts and novices often initially retrieved surface-similar distractors, experts were quicker to reject initially incorrect retrievals, suggesting stronger effects of domain knowledge in mapping and evaluation than in retrieval. Keane (1988) also found stronger effects of surface similarity on retrieval than on mapping and use. He carried out a series of studies manipulating aspects of the Gick and Holyoak (1980, 1983) convergence problem discussed earlier and demonstrated that a *literally similar* prior story (i.e., a story about a surgeon) was retrieved much more often than the remote (general and army) analog, but that subjects were equally good at mapping from the prior story to the target problem once both were present. Holyoak and Koh (1987) also manipulated the convergence problem: they gave subjects one of four prior versions of the convergence problem,

varying the structural and surface similarity to the target problem (the radiation problem). They found that spontaneous access was influenced about equally by surface similarity and structural similarity, but that subjects' ability to successfully perform the mapping was influenced only by structural similarity.

The evidence thus suggests that surface similarity is more important at access than at later stages. In fact, the difference may be even greater than these results suggest. In all these studies, subjects' remindings were elicited in the course of a problem-solving task, and so may have been filtered by the subjects' assessment of their usefulness. In this case the measured reminding rate reflects a combination of access, mapping and evaluation. Another complication is that, as Keane (1987) points out, findings from the problem-solving literature may underestimate people's abilities to retrieve and process analogies because subjects may lack the domain knowledge required to recognize a correct analogy.

In analogical *mapping*, there is considerable evidence that mapping accuracy is influenced by the consistency and systematicity of an analogy, whether in problem-solving (Bassock & Holyoak, in press; Keane, 1987; Novick, 1988; Reed, 1987; Ross, 1984, 1987), in story transfer (Gentner & Toupin, 1986) or in transfer of mental models from one device to another (Collins & Gentner, 1987; Gentner & Gentner, 1983; Schumacher & Gentner, 1988). Clement and Gentner (1991) found a focus on common systems in processing novel explanatory analogies. Asked to choose which facts were most relevant to the analogies, the subjects attended not just to whether the facts matched, nor to how important the facts were, but to whether the matching facts were *connected to other matching facts*. The same preference for common interconnected systems held for making predictions from an analogy. Given an analogy that allowed two possible predictions—i.e., that had two facts present in the base but not the target—subjects predicted the fact that was connected to a common relational system.

But if common relational structure is important in successful transfer, and if, as we will argue, retrieval processes often produce poor matches as well as good ones, it is important to know whether people can tell good analogies from poor ones. That is, do people weight structural consistency and systematicity strongly when evaluating the *soundness* of an analogy? This is a key prediction of structure-mapping theory (Gentner, 1983, 1989). There is not much direct evidence on this point. Gentner and Clement (1988) found that common relational structure is important in judging the aptness of metaphors.[4] People rate comparisons based on

[4] Although the literature on metaphor is strangely disconnected from that on analogy, there is a close connection between the phenomena. Many of the metaphors and similes used in current research could as well be called analogies. In particular, relational meta-

shared relations (e.g.,"A camera is like a tape recorder.") as more apt than those based on shared attributes (e.g., "The sun is like an orange.") and their aptness ratings are correlated with the degree to which their interpretations include relational (but not attributional) information (Gentner, 1988b; Gentner & Clement, 1988; Gentner, Falkenhainer, & Skorstad, 1988). But these findings address metaphoric *aptness*, not analogical *soundness*, and further, although they demonstrate a preference for shared relations, they do not address the more specific claim of a preference for common *systems* of relations governed by higher-order relations (the *systematicity* claim). Moreover, a contrary result was found by Reed (1987). He found that subjects weighted surface similarities strongly when asked to rate the potential usefulness of one mathematical problem for solving another. In some cases they even rated surface-similar but disanalogous problems as more useful than analogous, surface-dissimilar problems, suggesting that structural similarity was not their criterion for a useful analogy. However, usefulness may not be equivalent to soundness; subjects may have felt that a surface-similar problem would be more *useful* to them in a task because they lacked the domain knowledge necessary to apply a non-transparent match. Further, returning to Keane's (1987) point, the subjects may not have had the requisite domain knowledge to recognize a structurally sound analogy. Novick's (1988) studies show that expert mathematicians are considerably better than novices at discarding inappropriate retrievals, suggesting that with sufficient knowledge people may be able to judge the soundness of a mapping.

In the present research, we attempted to design clear tests of the kinds of similarity involved in similarity-based *access* and in *soundness evaluation*. We also strove to decompose levels of commonality precisely and to do so for relatively large stimulus sets. Much of the existing literature is based on very small sample sizes, often from one to four stories. Aside from the problems of generalizing over small samples, to study the phenomena of memory retrieval properly would seem to require moderately large sample sizes.

The basic method was the same across all three studies. We first created sets of story pairs differing systematically in their similarity class, as described below. To investigate the *access* process, we used a story-memory task. Subjects read a set of stories and later wrote out any remindings they experienced when reading a second set of similar stories. In contrast to the problem-solving context, here there is no additional criterion for the memory task beyond being reminded. This minimizes the

---

phors/similes, such as "Sermons are like sleeping pills." (Ortony, 1979), can readily be analyzed as analogies (Gentner, 1982; Gentner & Clement, 1988; Gentner, Falkenhainer & Skorstad, 1987).

contribution of other subprocesses, such as mapping and evaluation, and comes close to being a pure measure of access. To investigate the evaluation process, we presented subjects with the pairs of stories used in the memory task and asked them to rate the soundness (and in later studies, the similarity) of the pairs. This should require only mapping and evaluation.

The story pairs were designed as combinations of different levels of predicate matches. Matches could occur at three levels of predicates: (1) *object attributes*, such as *countries/nations;* (2) *first-order relations*, such as events and other relations between objects (e.g., *X shooting at Y/ P firing at Q*); and (3) *higher-order relational structure*, such as causal relations or other kinds of plot structure. By the systematicity assumption, the matches should be considered more sound when they share higher-order relational structure than when they do not. Note that it is not enough for the stories simply to *contain* the same higher-order relations: they must also have like *bindings* to lower-order relations: e.g., (1) is analogous to (2) but not to (3) (where A, B, C and x, y, z are events).

(1) A causes B, B causes C.
(2) x causes y, y causes z.
(3) x causes y, z causes x.

These three levels of matches—objects, first-order relations, and higher-order relational structure—were combined to yield different kinds of story matches. In the first experiment,[5] the story matches were *surface matches*, which matched in object descriptions and first-order relations; *analogies*, which matched in first-order and higher-order relations; and FOR *matches*, which matched only in first-order relations. Subjects were told that if the new story reminded them of any of the original stories they should write out that original story in as much detail as possible. After the reminding task, subjects rated the *soundness* of each pair of original and cue stories. Soundness was explained as the degree to which inferences from one story would hold for the other.

According to structure-mapping, the subjective soundness of an analogy should depend on the degree and depth of the common relational structure and not on the amount of overlap in object-attributes. Therefore, analogical matches should be rated as more sound than either surface matches or FOR matches. In contrast, the surface superiority claim for access predicts that access should depend much more heavily on the degree of object matches and event matches than on the degree of match in relational structure. Thus the access and soundness tasks should show

[5] The first study was carried out by Russell Landers in 1984 as an M.I.T. undergraduate honors project and appears in the Proceedings of the IEEE (Gentner and Landers, 1985).

different orderings of similarity types. In contrast, if there is one unitary kind of similarity on which both these processes depend, soundness and access should show the same ordering of similarity types.

## EXPERIMENT 1

### Method

#### *Subjects*

The subjects were 30 students from the MIT Psychology Department.

#### *Design and Materials*

Similarity type was varied within subjects, with each subject receiving ⅓ analogies, ⅓ surface matches, and ⅓ FOR matches. For counterbalancing, subjects were divided into three groups that differed in which story pairs they received in each Similarity type. For this purpose, the 18 stories were divided into three sets of six. This gave a 3 × 3 Group (between) X Similarity type (within) design.

The dependent measures for the reminding task were three measures of subjects' recall of the original stories: judges' ratings of their recalls, proportion of recalls rated above criterion, and proportion of recalls of a predefined keyword. The dependent measure for the soundness task was subjects' ratings of the soundness of the matches.

*Materials.* The materials were 18 sets of stories, with four stories per set, plus 14 filler stories. The stories were two or three paragraphs long. Each of the 18 story sets contained an original story plus three matching cue stories, which differed in the amount and level of similarity they shared with the original (See Table 2 and Appendix for sample sets of stories from Experiments 1 and 2). All cues shared identical or nearly identical first-order relations (e.g., events and actions) with the original story. They differed in the other levels of shared similarity.

In *analogy* (AN) cues, common higher-order relational structure was added to the first-order relation matches. The objects (i.e., the characters, physical objects, and locations) differed.

In *surface-similarity match* (SS) cues, object matches were added to the first-order relation matches. The higher-order relational structure (i.e., the causal structure or plot structure) differed.

In *FOR match* (FOR) cues, only the first-order relations matched. The objects as well as the higher-order structure differed.

Each subject received only one matching cue story for each of the 18 original stories. All subjects received the same memory set; the difference was in the type of story used in the cue set. This ensured that the measure of reminding would not be contaminated by differential availability of the memory material. To ensure comparability, the cue stories in a given set were made as similar as possible, subject to the constraints of the theory. In particular, the AN and FOR cues were constructed to have the same objects, and the SS and FOR cues were constructed to have the same higher-order structure. The objects were people, animals, companies, and countries. To avoid purely lexical remindings, the use of identical words between original and cue was avoided; similarities were expressed by means of synonyms or closely similar words. (The exceptions were function words and certain connectives and prepositions that had no acceptable substitutes.) For each story, one to

TABLE 2
Sample Stimuli from Experiments 1 and 2

Base Story

Karla, an old hawk, lived at the top of a tall oak tree. One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. The hunter took aim and shot at the hawk but missed. Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. The hunter was so grateful that he pledged never to shoot at a hawk again. He went off and shot deer instead.

Literal-similarity match

Once there was an eagle named Zerdia who nested on a rocky cliff. One day she saw a sportsman coming with a crossbow and some bolts that had no feathers. The sportsman attacked but the bolts missed. Zerdia realized that the sportsman wanted her tailfeathers so she flew down and donated a few of her tailfeathers to the sportsman. The sportsman was pleased. He promised never to attack eagles again.

Analogy match

Once there was a small country called Zerdia that learned to make the world's smartest computer. One day Zerdia was attacked by its warlike neighbor, Gagrach. But the missiles were badly aimed and the attack failed. The Zerdian government realized that Gagrach wanted Zerdian computers so it offered to sell some of its computers to the country. The government of Gagrach was very pleased. It promised never to attack Zerdia again.

Surface-similarity match

Once there was an eagle named Zerdia who donated a few of her tailfeathers to a sportsman so he would promise never to attack eagles. One day Zerdia was nesting high on a rocky cliff when she saw the sportsman coming with a crowsbow. Zerdia flew down to meet the man, but he attacked and felled her with a single bolt. As she fluttered to the ground Zerdia realized that the bolt had her own tailfeathers on it.

FOR Match

Once there was a small country called Zerdia that learned to make the world's smartest computer. Zerdia sold one of its supercomputers to its neighbor, Gagrach, so Gagrach would promise never to attack Zerdia. But one day Zerdia was overwhelmed by a surprise attach from Gagrach. As it capitulated the crippled government of Zerdia realized that the attacker's missiles had been guided by Zerdian supercomputers.

three of the characters were given proper names. These always differed between original and cue but were similar or identical for all three cue stories.[6]

Finally, following a technique used by Read (1983, 1984), the final sentence in each original story was not paralleled in any of the cue stories. This sentence served as a pure memory test; it could not be reconstructed by backwards mapping from the cue stories. It

[6] In some story sets, the sexes of the AN and FOR characters were changed from those of the original to heighten the lower-order differences. (The SS characters were, of course, always the same sex as the original.) In these cases, the names of the cue characters were kept as similar as possible: e.g., Mr. Boyce and Ms. Boyce, Christian and Christine, and Sidney and Cindy.

was designed so that it could not be predicted from the rest of the original, did not alter the plot structure, and contained at least one distinct new word or concept (the keyword), whose presence could be easily detected in the recall. For example, in the "Karla the hawk" set shown in Table 2, the final original sentence is "He went off and shot deer instead." with *deer* the unique word.

## Procedure

*Reminding task.* Subjects were tested in groups of three to eight in two separate sessions. In the first session, subjects read a booklet containing the 18 original stories intermixed with 14 filler stories. (The first three and last three stories were always fillers.) All subjects read the same 32 stories, in different semi-random orders. They were told to read the stories carefully, so that they would be able to remember them a week later. Subjects took about 30 min for this task.

The second session took place 6 to 8 days later. Subjects received a workbook containing 18 cue stories that corresponded to the 18 original stories read in the previous session. Each workbook consisted of six SS cues, six AN cues, and six FOR cues. Subjects were told that for each cue story they should write down any original story of which they were reminded. If they were reminded of more than one story, they were to write the one that best matched the current story. They were told to include as many details as they could remember—if possible, the names of characters and their motives as well as the events.

*Soundness-rating task.* Following completion of the reminding task in the second session, the subjects were given a soundness-rating task. In this task they were given the same pairs of stories they had received in the reminding task (regardless of whether the subject had succeeded in the recall task). Thus each subject rated six SM, six AN, and six FOR matches in the same counterbalancing groups as in the reminding task. They were asked to rate each pair for the inferential soundness of the match. In explaining the task, we avoided using terms like "analogy" or "analogous." Instead, *soundness* was explained as follows: "This part of the experiment is about what makes a good match between two stories or situations. We all have intuitions about these things. Some kinds of resemblances seem important, while others seem weak or irrelevant. . . ." We then described a *sound match* as one in which "two situations match well enough to make a strong argument," and in which "the essential aspects of the stories match . . ." so that "you can draw conclusions about the second story from the first." Subjects were told that the opposite of a sound match is a *spurious match*, in which "the resemblance between the stories is superficial." They were told that if they could "infer or predict much of the second story from the first . . . then this is a *sound match*. If you could not, then this is a weak, or *spurious match*."

## Scoring

The subjects' written recalls were scored by two judges. The judges were provided with a list of the 18 original stories, and for each of the subjects' recalls, the judges were told which original story was the intended match. However, they were *not* told which cue story the subject had seen; that is, they were blind to the Similarity condition. Three kinds of reminding scores were obtained:

(1) *Overall score.* The judges rated how close the subject's recalled version was to the actual original story, using a 0–5 scale, as follows:

5 = All important elements of the original and many details.
4 = All important elements of the original and some details.
3 = All important elements of the original but very few or no details.
2 = Some important elements of the original; others missing or wrong.

1 = Some elements from the original but not enough to be certain that the subject genuinely recalled the original.

0 = No recall or different story.

(2) *Proportion remindings. (i.e., number above criterion).* The 0–5 recall score reflects the *quality* of the recalls. This could lead to confoundings if some kinds of similarity matches permit more accurate recalls (through reconstruction). Therefore, as a second dependent measure, we dichotomized the judges' scores. All responses that were judged to be clearly identifiable recalls of the original (i.e., recalls that received an overall score of 2 or better) were classified as *remindings;* all descriptions with scores of one or zero were classified as non-remindings. This allowed a measure of the number of remindings each subject produced for each of the three similarity types.

(3) *Keyword score.* As discussed above, each original story contained a final key sentence that was not paralleled in the cue stories. To facilitate scoring, these sentences each contained a distinctive word or concept. Subjects' descriptions were scored as 2 if they contained the keyword or a close synonym, 1 if they contained a dubious synonym and 0 otherwise.

The independent scores assigned by the two judges agreed 75.5% of the time and were within one point of each other 97.5% of the time. Disagreements were resolved by discussion.

## Results and Discussion

### *Soundness*

As predicted, subjects judged the pairs that shared higher-order relational structure (the analogy matches) to be significantly more sound than the pairs that did not (the FOR matches and surface matches). Figure 1a shows the mean soundness ratings for the three types of similarity matches.

A one-way analysis of variance confirmed a main effect of Similarity type, $F(2,58) = 70.51$, $p < .0001$. Planned comparisons using the Bonferroni adjustment ($\alpha = .05$)[7] showed a significant advantage for AN matches ($M = 4.40$) over SS ($M = 2.80$) and FOR matches ($M = 2.74$) and no significant difference between SS and FOR matches. A one-way ANOVA using items as the random variable also revealed a significant effect of Similarity types, $F(2,34) = 46.31$, $p < .0001$.

### *Reminding*

The results of the reminding task show a different pattern. As predicted by the surface superiority hypothesis, the SS matches were the most effective reminding cues. On all three measures of reminding, the AN matches were much less effective than the SS matches. Figure 1b shows the proportion of stories recalled (i.e., the proportion of recalls assigned a score of 2 or better) for each type of cue. A one-way analysis of variance

[7] All planned comparisons in this paper were performed using the Bonferroni adjustment at $\alpha = .05$.

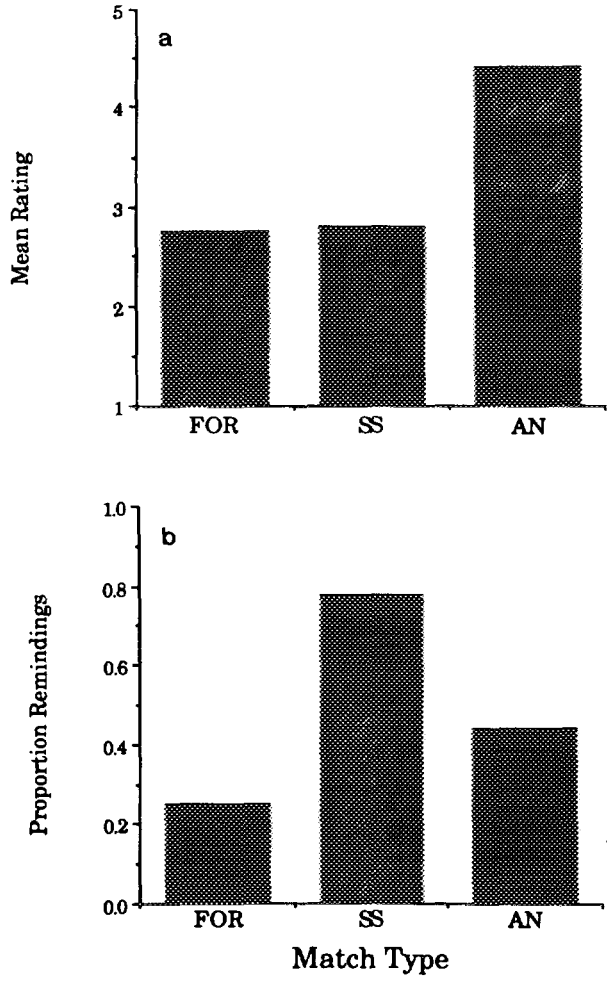


FIG. 1. (a) Mean soundness ratings for the three similarity types in Experiment 1. (b) Proportion recalled for the three similarity types in Experiment 1.

confirmed a main effect of Similarity type, $F(2,58) = 75.45$, $p < .0001$. Planned comparisons showed a significant advantage in proportion of remindings for SS matches ($M = .78$) over AN matches ($M = .44$) and for AN matches over FOR matches ($M = .25$). A one-way analysis of Similarity type over items again confirmed the main effect of Similarity type, $F(2,34) = 31.80$, $p < .0001$.

The other two measures of reminding showed a similar pattern to the proportion recalled. Table 3 shows the mean ratings of quality of recall

TABLE 3
Results of Experiment 1: Mean Rating of Soundness Compared to Three Measures of Reminding across the Three Similarity Types

| | Match type | | |
|---|---|---|---|
| | Analogy (R1 + R2)[a] | SS matches (OA + R1) | FOR matches (R1) measure |
| Soundness[b] | 4.40 | 2.80 | 2.74 |
| Proportion recalled | .44 | .78 | .25 |
| Proportion of keywords recalled | .21 | .26 | .13 |
| Quality of reminding | 1.48 | 2.36 | .86 |

[a] Here OA refers to object commonalities, R1 to first-order relational commonalities, and R2 to higher-order relational commonalities.
[b] Soundness and quality of reminding are mean ratings on 1 (low) to 5 (high) scales.

and the mean keyword scores across the three match types. For comparison, the proportion recalled and the mean soundness ratings are also given.

Similar analyses of the judges' ratings of quality of recall revealed the same pattern as for proportion recalled. There is a main effect of Similarity type, $F(2,58) = 72.81$, $p < .0001$. The mean rating for SS ($M = 2.36$) exceeded that for AN ($M = 1.48$), which exceeded that for FOR ($M = .86$). Again, the items analysis also showed a significant effect of Similarity type, $F(2,34) = 24.79$, $p < .0001$. The keyword analyses revealed weaker effects in the same direction. There was a marginal main effect of Similarity type in the subjects analysis, $F(2,58) = 2.80$, $p < .06$, but not in the items analysis.

Overall, the reminding results present a uniform picture. Not surprisingly, there is an effect of sheer degree of match. Starting with a first-order match, accessibility was increased either by adding common object-attributes (since SS was superior to FOR) or by adding higher-order relational structure (since AN was superior to FOR). However, the advantage was greater for adding object descriptions. Surface matches produced the greatest proportion of remindings, followed by AN and then by FOR matches, supporting surface superiority in retrieval.

This pattern differs markedly from the pattern found for subjective soundness. Indeed, the two measures were uncorrelated $r(52) = .15$, NS. For subjective soundness, as predicted by structure-mapping theory, common higher-order relational structure is a crucial determinant of the subjective "goodness" of an analogy. Subjects rated AN matches (R2 + R1) as more sound than FOR matches (R1), but SS matches (R1 + OA) were rated as no better than FOR matches. These results make it unlikely that soundness is determined by the mere number of features shared.

Rather, they suggest that common higher-order relational structure is particularly valued.

The dissociation between access and inferential soundness is striking. Analogical matches were rated as sound but were not well retrieved, and the opposite was true for surface matches. It appears that subjects' memory access mechanisms did not provide them with the matches that they themselves considered most valuable.

## EXPERIMENT 2

Experiment 2 was conducted to replicate and extend Experiment 1. First, we wished to investigate the possibility that the retrievability ordering found in Experiment 1 was simply a function of overall similarity, as suggested by Holyoak and Koh (1987). It could have been the case that the surface-similar cues were simply more similar to their originals than were the analogy cues, and the analogies more similar than the FOR matches. In this case, there would be no reason to invoke a special role for surface attributes in similarity-based access and it might be possible to preserve a unitary similarity view of retrieval. (Soundness evaluations would be dealt with separately in such a theory of transfer.) To address this concern, we added a similarity-rating task, again performed by independent subjects.

A second concern was the soundness-rating task. We wanted these ratings to reflect subjects' assessment of the inferential utility of the match. However, we became concerned that the instructions might have biased subjects against surface matches.[8] Although we described sound matches as ones from which predictions could be made, we also described spurious matches as "superficial" and cautioned the subjects against using shallow similarities as a basis for soundness judgments. Subjects could well have interpreted these instructions as warning against using object-level commonalities, in which case their seeming preference for relational matches is uninformative. To be sure we were gauging subjects' intuitions about inferential usefulness, we modified the instructions for the soundness-rating task, removing any caution against using surface or superficial features and stressing instead that subjects should judge whether they could make inferences or predictions from one story to the other.

A third concern was to avoid contamination from the reminding task to the ratings tasks, since the ratings task must of necessity follow the reminding task. We therefore conducted each of the ratings tasks with independent subjects (as well as with the recall subjects).

[8] We thank Keith Holyoak and an anonymous reviewer for pointing out this potential problem.

In addition to these concerns, an important addition was made to the design. In Experiment 1, we began with FOR matches (R1) and added either object-attribute commonalities to create surface matches (OA + R1) or higher-order relational commonalities to create analogy matches (R1 + R2), with the former leading to an access advantage and the latter leading to a soundness advantage. In Experiment 2, we completed this design, adding a *literal similarity* condition in which all three kinds of commonalities were present—OA, R1 and R2. With the addition of literal similarity (LS) matches, the set of match types forms a 2 × 2 design, as shown in Fig. 2.

This design enables a further test of the predictions for soundness. According to structure-mapping, the subjective soundness of a similarity match depends only on the degree to which relational structure is shared. Therefore, literal similarity matches and analogies should be rated as highly and about equally sound, since both kinds of matches include substantial and roughly equivalent common relational structure. In particular, LS soundness ratings should be no higher than AN soundness, despite greater numbers of shared predicates.

For access, the surface superiority hypothesis predicts that both LS matches and SS matches will be better recalled than AN matches and FOR matches. Thus the advantage of LS over AN should be greater than the advantage of LS over SS. In contrast, the goal-oriented memory view predicts that LS and AN will be better recalled then SS and FOR matches. Finally, the unitary similarity account predicts that the acces-

| | Common Higher-Order Relations | No Common Higher-Order Relations |
|---|---|---|
| Common Object-Attributes | LS | FOR |
| No Common Object-Attributes | AN | SS |

FIG. 2. Design of materials for Experiment 2. All pairs shared first-order relations (e.g., events).

sibility pattern will match the pattern of subjective similarity. The design allows two parallel tests of the relative contributions, since LS–AN and SS–FOR both reflect the additional contribution of object commonalities, while LS–SS and AN–FOR reflect the additional contribution of higher-order relational commonalities.

## Method

### *Subjects*

The subjects for the reminding task were 36 undergraduates from the University of Illinois who received class credit for participation. The subjects for the soundness and similarity rating tasks were 20 volunteer subjects from Hampshire College, who performed the soundness task with new instructions, and 40 undergraduates at the University of Illinois, 20 of whom received psychology class credit for participation and 20 of whom were paid for their participation. The paid and unpaid subjects from University of Illinois were evenly distributed across the similarity rating task and the soundness rating task with old instructions.

### *Materials and Design*

The materials were the same as those used in Experiment 1 with two additions. First, two new story sets were added for a total of 20. Second, the story sets were expanded to include, along with the original story, literal similarity (LS) cues as well as AN, FOR, and SS cues. This addition required a few minor changes in the prior stories to preserve the parallel design. The LS cues had in common with the original stories all three levels: object attributes (OA), first-order relations (R1), and higher-order relations (R2); SS had two levels (OA and R1) as did AN (R1 and R2); and FOR had one level of commonality (R1). (See Table 2 and Appendix for examples of LS cues.)

As in Experiment 1, Similarity type was varied within subjects. Subjects were divided into four counterbalancing groups which each received ¼ of the stories in each of the four similarity conditions. Each story set was rated equally often in each similarity condition and no subject received more than one pair from any story set. The presence of matching higher-order structure and object attributes was varied systematically among the four match types, making a 2 × 2 design: Relational similarity (within) × Object similarity (within).

### *Procedure*

*Soundness.* The soundness-rating task was run both with the original instructions and with new intructions that avoided biasing against surface similarities, as follows:

"This part of the experiment is about what makes a good match between two stories or situations. We all have intuitions about these things. Some kinds of resemblances seem important, while others seem weak or irrelevant . . . In this part of the experiment, we want you to use your intuitions about soundness—that is, about when two situations match well enough to make a strong argument. . . .

A *sound match* between two stories is one in which the essential aspects of the stories match. To put it another way, a *sound match* is strong enough that you can infer or predict things about the second story from the first. For example, suppose your read the first story and just the first part of the second story. Could you infer or predict, with fair accuracy, what happens in the rest of the second story? If you could predict it with reasonable accuracy, then this is a *sound match.* If you could not, then this is a weak, or *spurious match.*"

*Similarity.* The method for the similarity-rating task was similar to that used for the soundness ratings. Subjects saw pairs of stories and rated them on a 1–5 point rating scale

with 5 = extremely similar and 1 = extremely dissimilar. In the instructions, "similarity" was explained as "overall, how the characters and actions in the two stories resemble one another; or how much there is a general resemblance between the two stories." The vagueness of the description was intentional; we wanted this rating to reflect the subjects' own opinion of what makes two items similar.

*Reminding.* The procedure for the reminding task was as in Experiment 1. In the first session, subjects were told to read and remember 20 original stories and 12 filler stories. One week later, they returned for the reminding session and were given booklets of 20 cue stories, each of which matched one and only one original story. The instructions for the reminding task were identical to those of Experiment 1.[9]

The data were scored as in Experiment 1, with two blind judges rating each recall on a 0–5 scale for accuracy to the original story. The judges agreed 79% of the time and were within 1 rating point of each other 97% of the time. The proportion of remindings above criterion and the score for keyword recall were computed as in Experiment 1.

## Results and Discussion

### *Soundness*

As predicted, subjects based their soundness judgments primarily on the degree to which the two stories shared a relational system. (See Fig. 3a.) Planned comparisons revealed an advantage for LS matches ($M$ = 4.41) and AN matches ($M$ = 4.16) over SS matches ($M$ = 2.70) and FOR matches ($M$ = 2.58). As predicted, there was no significant difference between LS and AN matches nor between SS and FOR matches. A 2 × 2 Relational similarity (within) × Object similarity (within) analysis of variance confirmed a main effect of Relational similarity, $F(1,19)$ = 185.29, $p < .0001$. Neither the main effect of Object similarity nor the interaction of Relational similarity and Object similarity was significant ($F(1,19) = 1.99$, $p < .17$ and $F(1,19) = .638$, $p < .6$). The item analysis also revealed a main effect of Relational similarity, $F(1,19) = 86.19$, $p < .0001$.

We also ran the soundness-rating task on 20 other subjects using the instructions from Experiment 1, which emphasized avoiding superficial commonalities. The same pattern of results was obtained as above. Subjects rated LS matches ($M$ = 4.26) and AN matches ($M$ = 4.01) as more sound than SS matches ($M$ = 3.10) and FOR matches ($M$ = 2.87) (Bonferroni comparisons). The analysis of variance produced the same pattern of significance.

### *Similarity*

As predicted, subjects' similarity ratings were sensitive to both relational commonalities and object-attribute commonalities. (See Fig. 3b.)

[9] These subjects then performed the soundness-rating and similarity-rating tasks. The patterns of results were substantially the same as those of the independent subjects and are omitted for brevity.

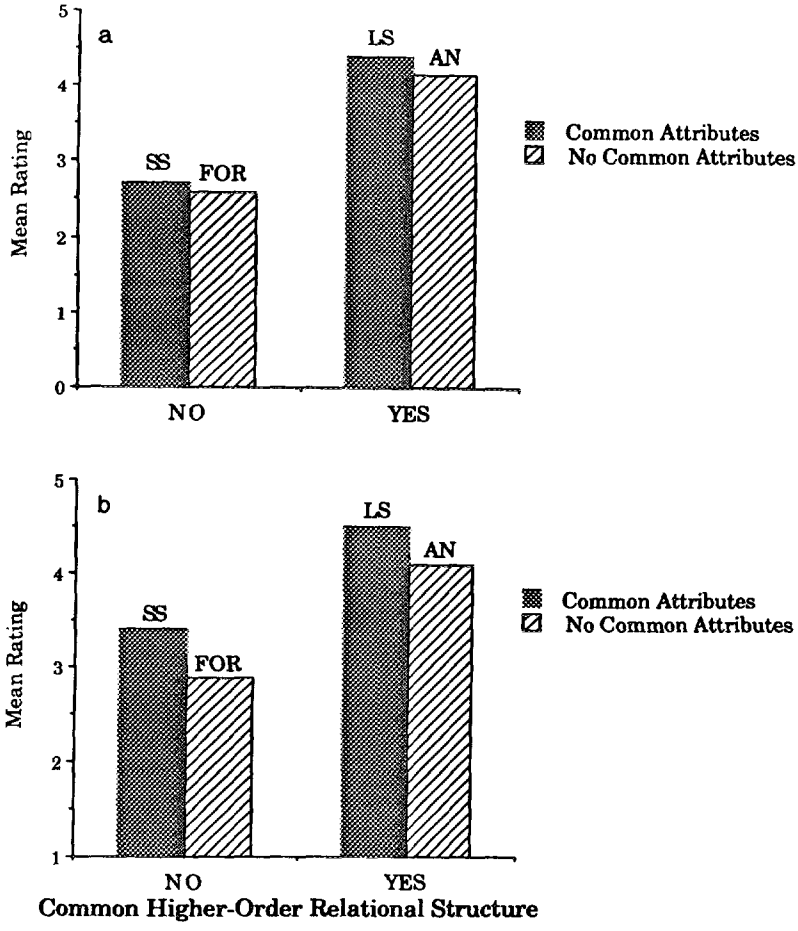


FIG. 3. (a) Mean soundness rating for the four similarity types in Experiment 2. (b) Mean similarity rating for the four similarity types in Experiment 2.

Bonferroni comparisons revealed an advantage of LS matches ($M$ = 4.50) over AN matches ($M$ = 4.09), of AN matches over SS matches ($M$ = 3.40), and of SS matches over FOR matches ($M$ = 2.88). The analysis of variance revealed main effects of Relational similarity, $F(1,19) = 85.78$, $p < .0001$, and Object similarity, $F(1,19) = 7.01$, $p < .02$. The interaction between Relational similarity and Object similarity was non-significant, $F(1,19) = .24$. The item analysis also revealed main effects of Relational similarity, $F(1,19) = 65.25$, $p < .0001$, and Object similarity, $F(1,19) = 5.98$, $p < .05$.

The comparison between similarity and soundness is revealing. Simi-

larity, like soundness, is sensitive to the presence of common higher-order relational structure. In fact, across the 80 story matches, similarity is highly correlated with soundness, $r(78) = .69$, $p < .001$. This is consistent with accounts of similarity that emphasize alignment of structure, rather than simple feature-matching (Falkenhainer, Forbus, & Gentner, 1989; Gentner, 1983, 1989; Goldstone & Medin, in press; Markman & Gentner, 1990, in press; Medin, Goldstone & Gentner, in press). However, whereas similarity is sensitive to both object commonalities and relational commonalities, soundness appears more specifically focused on relational commonalities.

*Reminding*

The pattern for reminding was very different from that for soundness and similarity. Object commonalities contributed strongly to memory access and higher-order relational commonalities had little effect. Subjects recalled more LS ($M = .56$) and SS ($M = .53$) matches than AN ($M = .12$) and FOR ($M = .09$) matches (Bonferroni comparisons). The difference between LS and SS and that between AN and FOR were nonsignificant. Figure 4 shows the proportion of remindings (i.e., the proportion of recalls receiving scores above criterion) for the four kinds of similarity matches.

A 2 × 2 Relational similarity (within) × Object similarity (within) analysis of variance performed on the number of remindings revealed a main effect of Object similarity, $F(1,35) = 108.73$, $p < .0001$. There was no

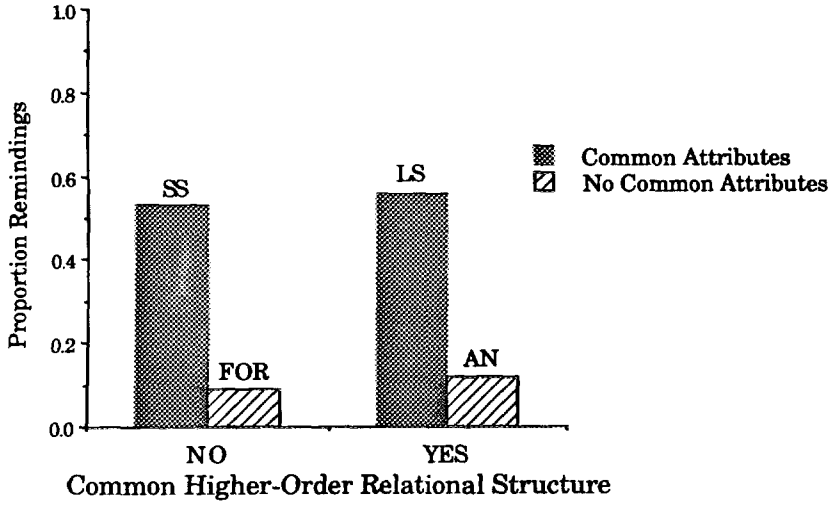


FIG. 4. Proportion recalled for the four similarity types in Experiment 2.

effect of Relational similarity, $F(1,35) = .89$. Also, as throughout the reminding results, there was no interaction between Object similarity and Relational similarity. An item analysis also revealed a main effect of Object similarity, $F(1,19) = 78.45$, $p < .0001$.

Similar analyses were performed on the overall quality-of-recall scores and the keyword data. Table 4 shows all three measures of reminding—proportions of remindings, mean ratings of quality of recall, and mean keyword scores across the three similarity match types. For comparison, the mean ratings of soundness and similarity are also included.

The quality-of-reminding results showed the same ordering of means: there was a significant advantage for LS matches ($M = 1.92$) and SS matches ($M = 1.64$) over AN matches ($M = .44$) and FOR matches ($M = .27$). The analysis of the quality-of-reminding data revealed a main effect of Object similarity, $F(1,35) = 124.18$, $p < .0001$, and also a main effect of Relational similarity, $F(1,35) = 4.98$, $p < .05$, presumably reflecting the advantage of relational matches in construction during recall. The item analysis also revealed a main effect of Object similarity, $F(1,19) = 77.82$, $p < .0001$, and a main effect of Relational similarity, $F(1,19) = 5.29$, $p < .01$. The keyword results showed the same ordering of means (See Table 4), although there were no significant pairwise comparisons. The keyword analysis revealed main effects of Object similarity, $F(1,35) = 11.86$, $p < .01$, but not Relational similarity, $F(1,35) = .5$. An item analysis also confirmed a main effect of Object similarity, $F(1,19) = 7.45$, $p < .05$, but not of Relational similarity, $F(1,19) = .65$.

TABLE 4
Results of Experiment 2: Mean Ratings of Soundness and Similarity Compared to Three Measures of Reminding across the Four Similarity Types

| | Match type | | | |
|---|---|---|---|---|
| | Literal similarity (OA + R1 + R2) | Analogy (R1 + R2) | SS matches (R0 + R1) | FOR matches (R1)[a] measure |
| Soundness[b,c] | 4.41 | 4.16 | 2.70 | 2.58 |
| Similarity | 4.50 | 4.09 | 3.40 | 2.88 |
| Proportion recalled | .56 | .12 | .53 | .09 |
| Proportion of keywords recalled | 1.11 | .17 | .81 | .25 |
| Quality of reminding | 1.92 | .44 | 1.64 | .27 |

[a] Here OA refers to object commonalities, R1 to first-order relation commonalities and R2 to higher-order relational commonalities.

[b] Soundness and similarity were rated by two independent groups of subjects.

[c] Soundness, similarity and quality of reminding are mean ratings on 1 (low) to 5 (high) scales.

The results of the reminding task replicate and extend the pattern of object-dominance in remindings found in Experiment 1. Matches that shared object-level commonalities—LS and SS matches—produced many more remindings than AN and FOR matches. In contrast, higher-order relational similarity did not lead to a significant recall advantage in any of the three measures except quality of reminding, and this measure seems likely to overestimate the importance of relational similarity in memory. (This is because, as discussed above, any tendency subjects may have to use the cue story as a template for their written recall will tend to lead to improved quality-of-reminding scores for LS and AN (the two relationally similar match types) but not for SS and FOR.) Moreover, even under the quality of reminding measure, AN matches were not significantly better than FOR matches in access. The other two measures of recall, keyword recall and proportion recalled, show effects of object similarity but not of higher-order relational similarity.

*Comparisons across tasks.* As in Experiment 1, the results of the reminding task contrast sharply with those of the soundness-rating task. In two different versions of the soundness task, subjects consistently relied on relational commonalities. Thus, as predicted by structure-mapping, common systematic relational structure appears paramount in subjects' judgments of inferential soundness.

This attention to relational structure was not reflected in subjects' recall performance. There was no correlation between the soundness ratings and proportion reminding, $r(78) = .008$, NS. Instead, similarity-based access to memory appears strongly influenced by surface similarity, with common relational structure playing a smaller role. In Experiment 2, common higher-order relational structure had no significant effect on the proportion of remindings; in Experiment 1 there was an effect, in that analogies were better recalled than FOR matches, but the gain was small compared to the gain for SS matches.

We now turn to whether memorial access reflects overall similarity. In this case, the reminding results should parallel the similarity ratings. The results do not bear out this possibility. There was no correlation between similarity and proportion recalled, $r(78) = .17$, NS. Further, as can be seen in Figs. 3b and 4, the results of the reminding task show a very different pattern from those of the similarity-rating task.

But although the pattern for similarity does not resemble that for recall, it does rather resemble that for soundness, and as noted above, similarity is highly correlated with soundness. Perhaps this degree of attention to structure in similarity judgments is a little surprising. An advocate of the "memory-reflects-similarity" view might even suspect that our subjects were not reporting their true feelings of similarity. But the finding that subjects are sensitive to common relational structure in judging similarity

is consistent with several other recent findings (e.g., Markman & Gentner, 1990, in press; Goldstone, Gentner, & Medin, 1991). We return to this point in the discussion.

Subjects considered relational matches more similar and more sound than object matches yet recalled more object matches than relational matches. These results support the surface superiority account of retrieval over the clever-indexing view. Further, these results suggest that different kinds of similarity govern the process of *access to long-term memory* and the processes of *mapping and evaluating* similarity comparisons when both members are present, arguing against a unitary model of similarity in transfer.

## EXPERIMENT 3

In the first two studies, surface matches were highly accessible in memory. Adding object commonalities to stories that shared first-order relations was much more effective in promoting recall than adding higher-order relations. This seems to support the claim of surface superiority in access. But there is another interpretation, namely, that the SS stories actually invited subjects to build appropriate higher-order relations, which then contributed to the reminding (suggested by Hammond, personal communication). The SS stories shared both objects and events with the base stories. If we suppose that particular combinations of objects and events are associated with typical higher-order causal patterns, then subjects receiving the SS matches might have invoked higher-order relational commonalities as well. For example, presented with an SS match that includes one country invading another, subjects might think of typical causal interactions among warring countries, and these higher-order patterns might contribute to their being reminded of the original story. In this case the SS matches would effectively have been (OA + R1 + R2) matches, possessing both higher-order commonalities and surface commonalities. Their retrievability advantage over AN (R1 + R2) could then be due to their effectively possessing more matching levels. Such an explanation would preserve the possibility that common higher-order structure is the crucial element in retrievability.

In Experiment 3, we tested this possibility. We created *object-only matches*—pairs of stories that had common object-descriptions but virtually no relational commonalities either in first-order events or in higher-order causal structure, as shown in Table 5. If in the first two experiments the seeming object-superiority in memory access was an artifact of higher-order relations being invoked, then object matches by themselves should be ineffective at promoting recall. In particular, objects-only matches (OA) should lead to considerably less retrieval than analogies

TABLE 5
Sample Materials from Experiment 2, Showing a Base Story and its Objects-Only Match

Base story

Two small countries, Bolon and Salam, were adjacent to a large, warlike country called Mayonia. Bolon decided to make the best of the situation by taking over Salam. Salam started looking for aid from other strong countries but soon Bolon succeeded in taking it over. Then victorious Bolon proposed to make a treaty with its warlike neighbor Mayonia. Bolon proposed to give Mayonia control over Salam in exchange for a guarantee that Bolon would remain independent. Mayonia responded by overrunning both Bolon and Salam. Bolon was so busy maintaining control of Salam, it could do nothing to stop Mayonia. Thereupon Mayonia installed puppet governments in both Bolon and Salam and took over the newspapers and ratio stations.

Objects-only match

Two weak nations, Lincoln and Moreland, bordered each other. Both countries relied heavily on the tourist trade to keep their economies afloat. They competed with each other over which one of them would get the most tourists. Meanwhile, another nearby nation, Chad, had a very strong economy with a thriving tourist trade. Tourist cruises flocked into its harbors and planes full of visitors were constantly landing in its airport. Because of this, Moreland tried to join forces with Chad in its new advertising campaign to entice still more tourists. Unfortunately a hurricane hit the coast and bankrupted all three nations.

(R1 + R2). Apart from this prediction, it is of some interest to know to what degree common object-descriptions by themselves can promote access.

This design also allows a further test of the predictions for soundness and similarity. According to structure-mapping, soundness should be based on the degree of common relational structure: highest for AN matches (R1 + R2), then SS matches (OA + R1), and finally OO (objects-only) matches (OA). Similarity should increase with the number of commonalities, so that OO matches should be rated lower in similarity than SS and AN matches. If we further find that AN is rated as more similar than SS, then this will strengthen the conclusion from Experiment 2 that structural commonalities are important in subjective similarity.

The same cued-recall method was used as in the prior studies, with one change. The keyword measure, which yielded consistent but weak results, was dropped in order to simplify the scoring procedure.

## Method

### *Subjects*

The subjects were 72 undergraduates from the University of Illinois, who were fulfilling a course requirement. Thirty-two subjects participated in the reminding task, 20 in the similarity-rating task and 20 in the soundness-rating task.

### *Materials and Design*

The materials were similar to those of the previous study, except that only AN matches and the two kinds of surface matches were used. Each subject received seven analogy matches (AN) and seven superficial matches. For half the subjects, the superficial matches were OO matches, which shared only object descriptions, and not first-order events nor causal structure. Table 5 shows sample materials. The other half received SS matches of the same kind used in the prior experiments, which shared first-order relations as well as objects. In addition, each subject received six LS matches (the same six across all subjects). This was done to give subjects some literally similar remindings to anchor their responses. Within each group, subjects were further divided into two counterbalancing groups that differed as to which stories were presented in each similarity type. Subjects were divided into two groups, an AN–SS group and an AN–OO group.

### *Procedure*

*Reminding.* The procedure was the same as that used in Experiments 1 and 2, except that keyword scoring was dropped. In the first session, the subjects read the 20 original stories and 12 filler stories. One week later, they were given 20 matching stories in the reminding task. Two judges rated the subjects' recalls. The inter-judge reliability was 75.8%.

*Soundness and similarity.* The procedure was as in Experiment 2, including the improved soundness-rating instructions.

## Results and Discussion

### *Soundness*

Table 6 shows the results for soundness, similarity, and reminding. As predicted, the soundness ratings reflected degree of relational overlap. The AN matches ($M$ = 3.85) were rated as significantly more sound than both the SS matches ($M$ = 2.38) and the OO matches ($M$ = 1.76) (within-group $t(9) = 3.03$, $p < .01$ and $t(9) = 9.73$, $p < .0001$, respectively). The

TABLE 6
Results of Experiment 3: Mean Ratings of Soundness and Similarity Compared to Three Measures of Reminding across the Three Similarity Types

| | Match type | | | |
|---|---|---|---|---|
| | Literal[c] similarity (OA + R1 + R2) | Analogy (R1 + R2) | SS matches (OA + R1) | OO matches (OA)[a] |
| Soundness[b] | 4.16 | 3.85 | 2.38 | 1.76 |
| Similarity | 4.59 | 4.01 | 3.21 | 2.30 |
| Proportion recalled | .62 | .07 | .52 | .17 |
| Quality of reminding | 1.79 | .22 | 1.39 | .47 |

[a] OA refers to object attribute commonalities, R1 to first-order relational commonalities, and R2 to higher-order relational commonalities.

[b] Soundness, similarity, and quality of reminding are mean ratings on 1 (low) to 5 (high) scales.

[c] The literal similarity results are included for comparison, although these matches served merely as anchoring stimuli.

SS matches were rated as significantly more sound than the OO matches (between-groups $t(18) = 2.49, p < .01$). Although LS matches were omitted from the analysis (because they served as anchors and did not vary between groups), their soundness rating ($M = 4.16$) was high, as expected. Finally, the soundness advantage of analogy over superficial similarity was greater for objects-only matches than for SS matches: that is, the AN–OO difference was greater than the AN–SS difference, between-group, $t(18) = 2.82, p < .05$.

*Similarity*

AN matches ($M = 4.01$) were rated as more similar than SS matches ($M = 3.21$) and OO matches ($M = 2.30$) (within-group $t(9) = 6.49, p < .0001$ and $t(9) = 2.96, p < .01$, respectively). SS matches were rated more similar than OO matches, $t(18) = 3.49, p < .01$. Also as for soundness, the difference between AN and OO matches was greater than the difference between AN and SS matches, $t(18) = 2.76, p < .05$.

*Reminding*

The results are shown in Table 6. As in the previous studies, the reminding pattern differs markedly from the patterns for soundness and similarity. Both SS matches ($M = .52$) and OO matches ($M = .17$) were better retrieved than AN matches ($M = .07$), within-subject $t(25) = 6.67$, $p < .0001$ and $t(25) = 2.78, p < .01$. SS matches were better retrieved than OO matches, $t(50) = 4.56, p < .0001$. As expected, the anchoring LS matches were highly retrievable ($M = .62$). Finally, the advantage of SS over AN was greater than the advantage of OO over AN, $t(50) = 4.56, p < .0001$.

The pattern for quality-of-recall was similar. The mean quality of recall was significantly higher for SS matches ($M = 1.39$) than for OO matches ($M = .47$), $t(50) = 5.06, p < .0001$. Quality of recall was higher for both SS matches and OO matches than for AN matches ($M = .22$), $t(25) = 7.22, p < .0001$ and $t(25) = 4.25, p < .001$. The significant advantage for OO matches over AN matches is especially telling, since the quality-of-recall measure should tend to favor AN matches over OO matches (because the AN matches can be used as templates to guide reconstruction of the original story). As expected, LS matches received high ratings ($M = 1.79$). The difference between the AN and the SS matches was greater than the difference between the AN and OO matches. ($t(50) = 4.64, p < .0001$).

*Summary*

Experiment 3 was carried out to find out which aspects of the SS matches led to the high accessibility observed in the previous experi-

ments. One way to save the clever-indexing position would be to argue that the apparent surface superiority in retrieval was merely an artifact of the fact that the first-order relational commonalities in the SS matches gave rise to further higher-order structural commonalities. This possibility is ruled out by the fact that not just the SS matches but also the OO matches are superior to the AN matches. The fact that OO matches (with object-attributes [OA] only) were recalled better (17% retrieval) than the purely relational AN matches (R1 + R2, at 7% retrieval) indicates that object-level commonalities, even by themselves, can compete with relational structure matches in accessibility.

On the other hand, the results of Experiment 3 also tend to rule out the possibility that the high recall of SS matches was due simply to the presence of object commonalities. Had this been the case, the OO and SS matches would have been equally well recalled in Experiment 3. In fact, SS matches (OA + R1 at 52%) were recalled significantly more often than OO matches.

As predicted, and as in the prior studies, the soundness ordering reflects the degree of relational overlap: AN, then SS, then OO. As before, this ordering contrasts sharply with the ordering of recallability: for example, analogy is rated as most sound but is least well recalled. Also as in Experiment 2, the similarity ordering is much like the soundness ordering, supporting the claim that structural commonalities are important in similarity, and quite unlike the accessibility ordering, undermining the notion of a unitary similarity governing the various stages of transfer.

## EXPERIMENT 4

We have interpreted these results as indicating that similarity-based retrieval is more sensitive to object similarity than to relational similarity. But an alternate possibility is that the relational disadvantage arises from encoding and storage rather than from retrieval. Perhaps subjects simply failed to process the original stories with sufficient depth to be able to tell the difference between surface and deep matches, or perhaps they forgot the higher-order structure. This possibility was raised by Hammond, Seifert, and Gray (1991), who carried out a study using the same materials as in our Experiment 2, but varying whether subjects received intact or scrambled stories as retrieval probes. Not surprisingly, they found that retrieval was worse with scrambled than intact probes. However, the overall pattern was the same in both conditions as it was in our study: stories with surface matches (LS and SS) were retrieved better than stories with low surface similarity (AN and FOR). Hammond, Seifert and Gray pointed out that the similarity in retrieval patterns between scrambled and intact probes is consistent with the possibility that the initial encodings were superficial. Although they concede that this pattern is

also consistent with our claims that the *encodings* were adequate and the *retrieval* was surface-oriented, their point is important and must be dealt with.

To determine the locus of the surface superiority effect, we need to know whether subjects possessed the requisite relational knowledge at test time. Therefore we carried out a further study. We gave subjects the same study task as in Experiment 2—20 base stories plus 12 fillers—and tested them one week later with a forced-choice recognition task in which they received the LS match (OA + R1 + R2) and the SS match (OA + R1). Since these two matches are equated for degree of surface match, if subjects prefer the LS match, we can infer that their encodings of the base stories included sufficient R2 information to permit the distinction.

A second group of subjects was given the same task, except that their recognition task consisted of AN matches (R1 + R2) versus FOR matches (R1). The reasoning here is parallel to the previous case, but with an added refinement. If subjects choose the AN match over the FOR match, this will indicate that their stored representations possessed sufficient higher-order relational information to make the distinction, *and* that the AN and/or FOR match is a good enough retrieval cue to provide access to the base representation.

## Method

### *Subjects*

The subjects were 27 undergraduates from Northwestern University, 14 in Group 1 and 13 in Group 2, who received course credit for their participation.

### *Materials and Design*

The materials were as in Experiment 2: 20 base stories plus 12 fillers, with the associated LS and SS matches used at test. All subjects received the same encoding task. Each of the two groups was further divided into four counterbalancing groups to control left-right position of the recognition test items and order of presentation.

### *Procedure*

In the first session, subjects read the 32 initial stories with the same instructions as in the previous studies. One week later, they returned and were given a forced-choice recognition task. Group 1 received the 20 pairs of LS and SS matches for the 20 base stories. They were told to choose the one that best matched the story they had read. The procedure was the same for Group 2, except that the AN and FOR matches were given in the recognition test.

## Results and Discussion

Group 1 subjects chose the LS matches 82% of the time ($M = 16.46$ out of 20), $\chi^2 = 61$ ($df = 12$), $p < .0001$. From this we can infer that the subjects' stored representations contained higher-order relational information. This suggests that the locus of the surface bias is at retrieval, and

this conclusion is further buttressed by the results for Group 2. Group 2 subjects were at chance: they chose the AN response only 55% of the time ($M$ = 11 out of 20). They presumably possessed the same sort of relational representation as the Group 1 subjects, since they had received precisely the same treatment. But whereas Group 1 subjects with their surface-similar probes were able to successfully access their base representations, Group 2 subjects, with their relationally similar probes, were not.

From these results we infer first, that subjects encoded and stored the relational structure of the base stories. Second, they were much better able to retrieve these descriptions when they later encountered surface-similar stories than when they later encountered relationally similar stories. We conclude that the locus of the surface superiority effect is at retrieval.

## COMPUTATIONAL SIMULATION

The results of these four experiments present a complex picture of similarity in transfer. These results, along with the related results discussed earlier, place constraints on a computational model of transfer: (1) the model must be able to store structured representations (the *structural representation* criterion); (2) it must capture processes of structural alignment and mapping over these representations (the *structural mapping* criterion). At the same time, the retrieval process must be such that (3) occasional relational remindings occur (the *rare insights* criterion); but (4) the majority of its retrievals are LS matches (the *primacy of the mundane* criterion); and (5) retrievals based on surface similarity are frequent (the *surface superiority* criterion). Finally, criterion (6), the *scalability* criterion, is that the system must be capable of being extended to large memory sizes.

Current models can be divided into those that capture the first two criteria and those that capture the last three. Models of similarity that assume smart processes operating over richly articulated representations can capture the first two criteria. Most case-based reasoning models have this character (Hammond, 1989; Kolodner, 1984; Schank, 1982). These models are sufficiently rich to capture processes like case-alignment, and even more clever processes, such as pragmatic matching and adaptation of cases (Kass, Leake, & Owens, 1987; Reisbeck & Schank, 1981). However, these models typically assume use of intelligent indices that capture significant higher-order abstractions, so that people should typically access the best structural match. Such models typically fail to predict the surface superiority effect in retrieval. It is also not clear how abstract indices would fare with very large memories.

The reverse set of advantages and disadvantages holds for approaches

that model similarity as the result of a dot product (or some other operation) over feature vectors, as is commonly done in mathematical models of human memory (e.g., Gillund & Shiffrin, 1984; Hintzman, 1984; Medin & Schaffer, 1978; but see Murphy & Medin, 1985; see Humphreys, Bain & Pike, 1989 for a review) and in many connectionist models of learning (e.g., Smolensky, 1988). These models, with their nonstructured knowledge representations and their relatively simple match processes, do not allow for the structural precision of people's similarity judgments and inferences. However, the use of feature-vectors has some advantages for modeling access to long-term memory. The computations are simple enough to make it feasible to compute many such matches and choose the best, thus satisfying criterion (6), *scalability*. It should also be straightforward for feature-vector representations to satisfy the *surface superiority* criterion (5), and (provided surface and structural features are correlated), the *primacy of the mundane* criterion (4).

We seek to combine the advantages of both these approaches by using a two-stage model of retrieval and mapping. The first stage carries out a cheap but error-prone search and the second stage carries out a full structural matching and mapping process. We begin with the second stage, SME, a simulation of mapping and evaluating comparisons. Then we summarize the large simulation, MAC/FAC, a simulation of similarity-based retrieval and mapping which uses SME as its second stage. We compare their results with the data presented above.

## SME and Subjective Soundness

The Structure-mapping engine (SME) is a simulation of the process of interpreting and evaluating comparisons. Given a pair of descriptions it carries out a structural alignment, projects candidate inferences from base to target and produces a structural evaluation of the interpretation. We briefly summarize the process here. (For details see Falkenhainer, Forbus & Gentner, 1986, 1989; Forbus & Oblinger, 1990; Gentner, 1989b; Skorstad, Falkenhainer, & Gentner, 1987). SME uses a local-to-global matching process to arrive at the maximal structurally consistent interpretation(s) of an analogy or similarity comparison. Given propositional representations of two potential analogs, SME begins by finding all possible local matches between predicates in the base and predicates in the target. For each predicate in the base, it finds all possible matches in the target. Two items can match if either (a) they are intrinsically alike[10] or (b) they

[10] SME's constraint of matching identical predicates assumes canonical *conceptual* representations, not lexical strings. Two concepts that are similar but not identical (such as "bestow" and "bequeath" are assumed to be decomposed into a canonical representation language so that their similarity is expressed as a partial identity (here, roughly "give").

are corresponding arguments of matching predicates and are objects or functions.

These local matches are often mutually inconsistent. In the next stage, SME imposes structural consistency: it combines these local hypotheses into subsystems that are mutually consistent (e.g., that have one-to-one object bindings and parallel relational connectivity). It then tries to combine these subsystems into larger mutually consistent mappings. In this way it generates one or a few global interpretations of the comparison.[11] SME also draws any further *candidate inferences* that would follow from the interpretation. Predicates that belong to the base system but were not initially present in the corresponding target system are hypothesized to be true in the target system.

Finally, each interpretation is given a structural evaluation. To ensure a preference for systematic relational structure, the depth of the common structure is taken into account in determining the evaluation. SME uses a cascade-like process in which predicates pass some multiple of their match score down to their arguments (Falkenhainer, Forbus & Gentner, 1986, 1989; Forbus & Gentner, 1989). This means that a deeply nested interpretation will be preferred over a flat interpretation containing the same number of matching predicates.

*Comparing SME's Evaluations with Human Soundness Judgments*

According to theory, the structural evaluations assigned by SME in analogy mode should ordinally match the soundness ratings assigned by human subjects across the story pairs. In *analogy* mode, SME initially ignores object attributes and matches identical relations; all other predicates are put into correspondence based on their structural roles. SME's *literal similarity* mode is similar except that object attributes are also matched initially. The theoretical commitment here is that *soundness* depends only on common relational structure, and hence should be captured by analogy mode, while *similarity* depends on an overall match (so object similarity as well as structural alignment is important). To test the soundness claims, we constructed propositional representations for 9 of the 20 story sets used in Experiment 2. The 9 base-AN pairs and 9 base-SS pairs were given to SME in analogy mode. The results are shown in Table 7, along with the human soundness results from Experiment 2.

Consistent with the theory, SME's structural evaluation scores are higher throughout for the base-AN pair than for the base-SS pair. To further test the apparent parallel between SME's pattern of preference

[11] SME can be run exhaustively to produce all possible interpretations of a given comparison, but we think it more plausible that humans produce only one or two interpretations (Forbus & Oblinger, 1990).

TABLE 7
Comparison of SME's Structural Evaluation Scores with Human Soundness Ratings

| | SME's structural evaluation score | | | Human subjects' soundness ratings | | |
|---|---|---|---|---|---|---|
| | AN | SS | AN > SS | AN | SS | AN > SS story # |
| 5: Karla, hawk | 23.5 | 17.0 | + | 4.4 | 2.4 | + |
| 7: Julius, mule | 26.0 | 21.0 | + | 4.6 | 2.0 | + |
| 8: Percy, squirrel | 19.5 | 17.5 | + | 4.2 | 3.2 | + |
| 9: Steak, dog | 38.5 | 33.0 | + | 4.0 | 1.8 | + |
| 10: Boris, business | 39.0 | 34.5 | + | 3.4 | 1.4 | + |
| 13: Morris, prisoner | 45.5 | 28.5 | + | 5.0 | 1.4 | + |
| 15: Fred, shepard | 55.0 | 45.5 | + | 4.8 | 1.8 | + |
| 17: Pioneers, divide | 21.5 | 19.5 | + | 4.2 | 2.2 | + |
| 19: Cobra, Pierre | 51.0 | 47.5 | + | 4.6 | 4.2 | + |

[a] These figures are for SME in analogy mode.
[b] + Indicates that the score was greater for the analogy (AN) pair than for the surface-similarity (SS) pair.

and that of human subjects, we computed for each story set the difference between SME's base/AN score and its base/SS score and correlated this with the corresponding difference in the human soundness ratings (i.e., base/AN–base/SS). As predicted, we found a significant positive correlation, $r(7) = .73$, $p < .05$.

*Summary*

SME was designed to embody the structural consistency and systematicity constraints. While the fit between SME's evaluation order and that of humans is not proof of the psychological reality of these constraints, it shows that this account is consistent with human performance. SME's process of beginning with local matches and coalescing these into global structural matches seems apt for modeling human processing. For example, Goldstone and Medin (in press) have found that local similarities have their effects on mapping earlier than global relational similarities in a timed mapping task. Ratcliff and McKoon (1989) found that in a sentence-matching task subjects could discriminate new from old sentences early in processing if the new sentences contained all new words (e.g.,"Helen attracted Jeff." vs. "Andrew accosted Mary."); but only after about 700 msec. could subjects reject sentences based on differences in relational structure (e.g., "Helen attracted Jeff." vs. "Jeff attracted Helen."). In pilot experiments using perceptual stimuli, in which subjects were timed under different kinds of mapping instructions, Arthur Markman and I have found that subjects are faster to choose on the basis of similar

objects than on the basis of similar relations, *even when the two rules dictate the same response.*

Aside from this timing evidence, SME's local-to-global process of computing alignment has the attractive feature that it begins blindly. Since the emergence of clever or deep interpretations results from its ability to utilize interconnectivity between its initial local matches and its preference for systematicity and structural consistency, it does not need advance knowledge of the content of the analogy as in some prior accounts (see Carbonell, 1986; Holyoak, 1985). The local-to-global algorithm is adaptable to a connectionist framework: similar processing principles have been incorporated into Holyoak and Thagard's (1989a) ACME model of analogical mapping and Goldstone and Medin's (in press) SIAM model of similarity processing. ACME has the attractive feature of combining structural, pragmatic and similarity-based constraints, though at the cost of failing to ensure structural consistency. SME's use of structural consistency permits spontaneous inferences from base to target. Hofstadter & Mitchell's (in press) Copycat also develops an overall match from local pressures, using a parallel terraced scan to allocate attention among multiple local clusters (Hofstadter, 1984).

## Simulating Memory Retrieval: The MAC/FAC Model

Of the six constraints on a model of similarity-based transfer, SME satisfies the first two: it accepts *structural representations* and produces *structural mappings*. Now we turn to the other four criteria—the *rare insights* criterion, the *primacy of the mundane* criterion, the *surface superiority* criterion, and the *scalability* criterion. To simulate similarity-based retrieval we require a process that is strongly but not solely influenced by surface similarity and that is somewhat but not wholly insensitive to structural consistency. It should typically retrieve literally similar matches, often retrieve surface-similar matches, and occasionally retrieve purely analogous matches (Wharton, Holyoak, Downing, Lange and Wickens, 1991, 1992). Finally, it must be able to deal with the fact that although *access* to memory may be structurally insensitive, once the knowledge is retrieved its structure can be used.

The model we propose, called MAC/FAC (for "Many are called but few are chosen"), uses a two-stage retrieval process (Gentner, 1989b; Gentner & Forbus, 1991). For a full description of the model see Gentner & Forbus (in preparation); here we briefly summarize it. The first stage (MAC) is a "wide-net" stage in which a crude, computationally cheap match process is used to pare down the vast set of memory items into a small set of candidates for more expensive processing (c.f. Bareiss & King, 1989). The second stage—the FAC stage—is a full structural similarity matcher, namely, SME in literal similarity mode, as discussed

above. The dissociation noted previously, we claim, can be understood in terms of the interactions of its two stages. MAC/FAC's inputs are a pool of memory items and a *probe:* a description for which a match is to be found. Its output is a memory description and a comparison of this description with the probe. We make minimal assumptions concerning the global structure of long-term memory: e.g., whether the pool is the whole of long-term memory or just a subset.

*The MAC stage*

The "Many are called" initial process is meant to be a cheap, fast process whose job is to select a manageable number of possible analogs to pass along to the "Few are chosen" stage. The MAC stage aims at a quick estimate of the overlap between the probe and each item in the memory pool. One way to do this would be to carry out the first part of a full analogy process for each possible pair, and then simply count the match hypotheses for each pair rather than completing the structural alignment. This technique was used in our original version of MAC/FAC (Gentner, 1989b) as well as in ARCS (Thagard, Holyoak, Nelson, & Gochfeld, 1990). The problem with this technique is that it is very costly. Even with parallel and/or neural hardware, it is unlikely that generating match hypothesis networks between a probe and everything in the pool of memory can be accomplished quickly enough to provide realistic response times for a large memory pool.

The method we adopted is to associate with each structured representation a *content vector.* Content vectors are flat summaries of the knowledge encoded in complex relational structures. The content vector for a given description specifies which predicates were used in that description and the number of times they occurred. (Here "predicate" is used in the inclusive sense to include functions, connectives, object attributes, relations, and so on.) Thus if there were four occurrences of IMPLIES in a story, the value for the IMPLIES component of its content vector would be four. (Other values are possible depending on the normalization used (see Gentner & Forbus, 1991, in preparation.)) Content vectors are assumed to arise automatically from structured representations and to remain associated with them.

On this account, initial access occurs at the level of content vectors and later stages of access and reasoning occur over structural descriptions. In the initial (MAC) stage, a dot product is taken between the content vector of the probe item and the content vector of each item in the memory pool. The output of MAC is the best match and everything within 10% of it. These pairs are passed to the FAC stage.

*The FAC Stage*

The FAC stage is responsible for structural alignment, interpretation

and evaluation of the best matches from the MAC stage. We use SME in literal similarity mode as the FAC matcher. We use literal similarity rather than analogy on the assumption that the normal criterion for successful retrieval is a good overall match. (Recall that literal similarity is structurally sensitive like analogy, but more inclusive.) FAC operates on the structural representations, not the content vectors. It carries out a structural alignment of each of the MAC output items with the probe, resulting in an interpretation, a set of new candidate inferences, and a structural evaluation. Thus, at the FAC stage issues of parallel structural binding become important. FAC will reject many of the matches given to it by MAC.

FAC thus acts as a structural filter on the output of MAC. This allows a divide-and-conquer strategy. The MAC stage is cheap and could plausibly be carried out in parallel for every item in the memory pool, perhaps in a connectionist architecture. To be sure, MAC's estimate of similarity as the dot product of content vectors has critical limitations. Like feature-vector schemes, it does not take the actual relational structure into account. It only produces a numerical score, and hence doesn't produce the correspondences and candidate inferences which provide the power of analogical reasoning and learning. But these limitations, in combination with FAC's structural filter, may be what is needed to model our highly efficient but somewhat fallible memory.

How successful is this two-stage model? We briefly summarize some comparisons of MAC/FAC's performance with the psychological data. (For the full set of studies, see Gentner & Forbus, 1991; in preparation.) Across the three experiments, the human subjects showed a stable retrievability order: LS > = SS > AN > = FOR (where "> =" means "greater than or equal to"). For example, in Experiment 2 the proportions of remindings were .56 for LS, .53 for SS, .12 for AN and .09 for FOR matches. The question is whether MAC/FAC can duplicate the human pattern: in particular, the propensity for retrieving SS and LS matches rather than AN and FOR matches.

*Comparing MAC/FAC with Human Performance*

For these simulations, we encoded predicate calculus representations for 9 of the 20 story sets (45 stories).[12] Then we gave MAC/FAC different memory sets and different probe types and recorded retrievals. To count as a retrieval, a story had to pass both MAC and FAC.

For example, in one study, MAC/FAC's memory set consisted of all four variants of each of the nine base stories (a memory set of 36 stories).

[12] To have encoded all 20 story sets plus the 12 distractors would have required a prohibitive amount of encoder time.

Each base story in turn was used as a probe. This is a more difficult task than the subjects', for whom there was one and only one memory match for each probe story; but in partial compensation, MAC/FAC had nothing corresponding to the subjects' delay between study and test. In any case it is not absolute difficulty but order of difficulty that is at stake here.

Table 8 shows the mean number of matches of different similarity types that pass both MAC and FAC. There are several points to note. First, the retrieval results (i.e., the number that make it through both stages) ordinally match the results for human subjects: LS > SS > AN > FOR. This degree of fit is reasonably encouraging. Second, as expected, MAC produces some matches that are rejected by FAC. The mean number of memory items produced by MAC is 3.3 and the mean number accepted by FAC is 1.5. Third, as expected, FAC succeeds in its job as a structural filter on the MAC matches. It accepts all of the LS matches proposed by MAC and some of the partial matches (the SS, AN and FOR matches), and rejects most of the inappropriate matches (the "other" matches from different story sets). It might seem puzzling that FAC accepts more SS matches than AN matches, when it normally would prefer AN over SS. The reason is that it is not offered this choice; it must take the best of the matches passed on by MAC for a given probe.

We have tried several other variants with similarly encouraging results. For example, in another study, MAC/FAC was given the 9 base stories in memory along with the 9 FOR stories, which served as distractors. We then used each of the variants—LS, SS, and AN—as probes. The proportion of times the base story made it through both MAC and FAC was 1.0 for LS, .89 for SS, and .67 for AN. Again the results ordinally match those of human subjects.

We can compare MAC/FAC's performance with that of the closest comparison model, Thagard, Holyoak, Nelson, and Gochfeld's (1990) ARCS model of similarity-based retrieval. Thagard *et al.* gave ARCS a

TABLE 8
Results of MAC FAC Simulation: Mean Numbers of Different Match Types Retrieved When Base Stories Are Used as Probes

| Retrievals | MAC | FAC |
|---|---|---|
| LS | .78 | .78 |
| SS | .67 | .44 |
| AN | .33 | .11 |
| FOR | .22 | 0 |
| Other | 1.33 | .22 |

*Notes.* Memory contains 36 stories (LS, SS, AN, and FOR for 9 story sets); the probes were the 9 base stories. *Other* = any retrieval from a story set different from the one to which the base belongs.

memory set consisting of the Karla the hawk story in memory along with 100 Aesop's fables as distractors. When given the four similarity variants as probes, ARCS produced asymptotic activations as follows: LS (.67), FOR (−.11), SS (−.17), AN (−.27). These results differ from the LS = SS > AN > = FOR order found for human remindings. In particular, SS remindings, which should be about as likely as LS remindings, are quite infrequent. In contrast, MAC/FAC produces the same ordinal sequence, explaining the data better than ARCS. This is especially interesting because Thagard et al. argue that a complex localist connectionist network which integrates semantic, structural, and pragmatic constraints is required to model similarity-based reminding. While such highly integrated models are intriguing, MAC/FAC shows that a simpler model can provide a better account of the data.

There are open questions with MAC/FAC. For example, at present it always produces at least one match for every probe. We are experimenting with the use of thresholds to capture cases where nothing is recalled. Also, MAC/FAC does not capture effects of multiple simultaneous interactions of probes with memory items (e.g., Hintzmann, 1984, 1986; Medin & Schaffer, 1978), and competition among retrieval items is less integral than in ARCS (Thagard, Holyoak, Nelson, & Gochfeld, 1990; Wharton, Holyoak, Downing, Lange, & Wickens, 1991, 1992).

However, MAC/FAC's overall pattern of behavior captures the motivating phenomena. It allows for structured representations and for processes of structural alignment and mapping over these representations (criteria 1 and 2). It produces a small number of analogical matches, thus satisfying criterion 3, the existence of rare insights criterion. The majority of its retrievals are LS matches, thus satisfying criterion 4, the primacy of the mundane. It also produces a fairly large number of SS matches, thus satisfying criterion 5, surface superiority. Finally, its algorithms are simple enough to apply over large-scale memories, thus satisfying criterion 6, scalability.

## GENERAL DISCUSSION

We began with the venerable claim "Similarity is central in transfer." Our results suggest that the proper terms are "Similarities" and "transfer processes." An account of similarity's effects on transfer requires making fine distinctions both about similarity and about transfer. Specifically, we found a dissociation between the matches people *get* from memory and the matches they *want*. The accessibility of matches from memory was strongly influenced by surface commonalities and weakly influenced by structural commonalities. In contrast, rated inferential soundness of comparisons was strongly influenced by structural commonalities and not at

all influenced by surface commonalities. Subjective similarity itself was influenced by *both* surface and structural commonalities.

## Soundness

The results bear out the structure-mapping prediction that the subjective soundness of a similarity match is determined by the degree to which the analogs share relational structure. Three aspects of the results lends support to this prediction. First, adding common relations increases the perceived soundness of a match. In every case where a precise comparison is possible, soundness increases with the addition of relational commonalities. When higher-order relations were added in Experiments 1 and 2, analogies (R1 + R2) were judged more sound than first-order matches (R1); and similarly in Experiment 2, literal similarity matches (OA + R1 + R2) were judged more sound than surface matches (OA + R1). When first-order relations were added in Experiment 3, surface matches (OA + R1) were judged more sound than objects-only matches (OA).

The second finding is that adding higher-order commonalities to a first-order relational match increases its soundness more than adding object commonalities. In all three experiments the soundness of analogical matches (R1 + R2) is substantially higher than that of surface matches (OA + R1). By itself this result might simply mean that the number or salience of features was greater for higher-order relational features than for object features. But the finding of a surface advantage in memory access renders this explanation implausible, or at least forces us to assume a different salience weighting for retrieval than for subjective soundness; and this is our point.

The third piece of evidence for structural specificity in soundness judgments is that the addition of object-attribute commonalities *fails* to increase soundness. The conditions for this test are met once in Experiment 1 and twice in Experiment 2. In Experiment 1, surface matches (OA + R1) were considered no more sound than first-order event matches (R1) ($M$ = 2.80 and $M$ = 2.74, respectively). In Experiment 2, surface matches (OA + R1) ($M$ = 2.70) were not significantly more sound than FOR matches (R1) ($M$ = 2.58). Literal similarity matches (OA + R1 + R2) ($M$ = 4.41) were not significantly more sound than analogies (R1 + R2) ($M$ = 4.16). Without embracing the null hypothesis, we can safely say that these studies produced no evidence that adding object commonalities increases soundness. Overall, the results indicate that subjective soundness reflects the degree of common relational structure.

These results are compatible with prior findings concerning aptness of metaphor. Subjects' ratings of the aptness of metaphors were positively correlated with the degree of relationality of their interpretations (as judged by independent judges) and either uncorrelated or negatively cor-

related with the attributionality of their interpretations (Gentner, 1988b; Gentner & Clement, 1988). Finally, the correlation between our subjects' soundness ratings and SME's structural evaluation scores is consistent with the claim that the human perception of inferential soundness is based on the degree of structural match.

## Similarity-Based Access

The results for access were almost the reverse of the results for soundness. Subjects tended not to retrieve the matches they considered most sound. Instead, they were most likely to access surface matches. As above, there are three lines of evidence. First, adding common object attributes to a match consistently increased the proportion retrieved. In Experiments 1 and 2, recall of surface matches (OA + R1) was greater than recall of FOR matches (R1) by at least a factor of three. In Experiment 2, recall of literal similarity matches (OA + R1 + R2) was greater than recall of analogical matches (R1 + R2) by about a factor of five. Second, adding common object attributes contributes more to retrievability than adding common higher-order relations (in contrast to the soundness results). In all three studies, recall of surface matches (OA + R1) exceeded recall of analogical matches (R1 + R2). The proportions of surface matches retrieved across the three studies were .78, .53, and .52, respectively, substantially higher than the retrieval rate for analogies (.44, .12, and.07, respectively).

Finally, we can ask whether this surface advantage is absolute. Does adding higher-order relational commonalities contribute anything to accessibility? The evidence here is mixed. In Experiment 2, analogy (R1 + R2) ($M$ = .12) was no better recalled than first-order matches (R1) ($M$ = .10) and literal similarity (OA + R1 + R2) ($M$ = .56) no better than SS (OA + R1) ($M$ = .53). However, in Experiment 1, analogies (R1 + R2) ($M$ = .44) were retrieved more often than first-order matches (R1) ($M$ = .25). Taken in combination with other studies, these findings suggest that higher-order relational similarities can sometimes promote access, but that the effects are less robust than the effects for surface matches. We will return below to the issue of when relational access is most likely.

## Similarity

Accessibility and subjective soundness are both aspects of similarity in transfer. Their divergent patterns leave us in something of a quandary: which is the real similarity? We therefore asked subjects to rate similarity directly. If subjective similarity had accounted for accessibility, then it might have been possible to argue a unary notion of similarity in that accessibility is governed by similarity and to deal with the divergent soundness patterns in some other way. However, subjective similarity

resembled soundness in its sensitivity to common relational structure. Adding relational information increased subjective similarity in every case where the comparison can be made. Literal similarity matches (OA + R1 + R2) ($M$ = 4.5) were considered more similar than surface matches (OA + R1) ($M$ = 3.4) (Experiment 2); analogical matches (R1 + R2) ($M$ = 4.09) were considered more similar than FOR-matches (R1) ($M$ = 2.88) (Experiment 2); and surface matches (OA + R1) ($M$ = 3.21) were considered more similar than object-only matches (OA) ($M$ = 2.3) (Experiment 3). The second line of evidence is the comparative addition argument. If, starting with a first-order match, we compare the effects of adding higher-order commonalities versus adding object commonalities, we find that analogies (R1 + R2) are rated as more similar than surface matches (OA + R1) in both Experiment 2 and Experiment 3 ($M$ = 4.01 and $M$ = 3.21, respectively). But unlike soundness, similarity is also increased by adding object commonalities. Literal similarity matches (OA + R1 + R2) are considered more similar than analogy matches (R1 + R2); and surface matches (OA + R1) are considered more similar than FOR-matches (R1) (Experiment 2). It appears that both structural and surface similarities contribute to subjective similarity.

*Similarity and Structural Alignment*

The finding that common relational structure contributes to similarity is consistent with other recent findings. For example, Goldstone, Medin and Gentner (1991) and Medin, Goldstone, and Gentner (1990) found that when subjects were asked to choose between a relational match and an attributional match to a standard, they generally preferred the relational match: e.g., *XX* would be considered more similar to TT than to XT. Further, this relational bias is affected by whether relational or attributional commonalities already predominate (Goldstone, *et al.*, 1991). Another source of evidence for the role of relational structure in similarity is configural effects in perceptual similarity. Pomerantz, Sager, and Stoever (1977) found that subjects could more easily discriminate between the stimuli ) and ( when they were presented in the context of a third identical element, ), yielding the discrimination )) vs. ( ). Pomerantz et al. suggested that the effect of adding this contextual component was to promote emergent features, such as symmetry and intersection. The added component introduced different relational structures in the two stimuli, making them more dissimilar and hence more discriminable. (See also Lockhead & King, 1977; Palmer, 1977, 1978, 1989).

The relationship between similarity and structural alignment is underscored by a finding of Markman and Gentner's (1990, in press). Subjects were asked to map between two pictures that were constructed to contain cross-mappings: e.g., XYZYX, voxov. Subjects who judged the simi-

larity of the two scenes before doing the mapping task were more likely to base their mapping on an alignment of relational structure rather than using local object matches: e.g., they would map X onto v in the above pair, rather than X onto x. These findings suggest that aligning relational structure is integral to the perception of similarity and support the claim that the computation of similarity utilizes common structural connectivity among features, rather than simply counting numbers of matching features (Falkenhainer, Forbus & Gentner, 1989; Goldstone & Medin, in press; Holyoak & Thagard, 1989; Hofstadter & Mitchell, in press).

### *The Plurality of Similarity*

The dissociation between surface similarity and structural similarity across different processes is related to several recent discussions. Medin, Goldstone, and Gentner (in press) and Gentner (1989) have argued that similarity is pluralistic, in the sense that there are multiple subclasses of similarity and multiple influences on how it is computed. Rips (1989) demonstrated a dissociation between similarity, typicality, and categorization. Murphy and Medin (1985) and Keil (1989) have commented on the limited usefulness of simple similarity and pointed out that physical resemblance does not provide a sufficient basis for conceptual structure. Barsalou's 1982 ad hoc categories, such as "things to take on a picnic" and Glucksberg and Keysar's (1991) metaphorically based categories, such as "jail" as a prototypical confining institution, are examples of abstract or relational commonalities. The present results argue specifically against the notion of a unitary similarity that governs retrieval, evaluation and inference.

## Similarity-Based Retrieval

### *MAC/FAC*

The fact that access to memory is surface-driven cannot be taken to mean that the memories themselves contain only surface features. Gick and Holyoak's (1980, 1983) studies and our Experiment 4 demonstrate that even when subjects are not able to use structural matches to access memory, they nonetheless may possess that information. The MAC/FAC simulation is aimed at capturing both these facts: that humans successfully store and retrieve intricate relational structures—and that *access* to these stored structures is heavily—although not entirely—surface-driven. MAC/FAC is a two-stage retrieval model whose first stage is attentive to content and blind to structure and whose second stage is structurally sensitive. MAC/FAC's first stage simply computes the dot product of the probe's content vector—a simple list of all the predicates contained in the structural description—with that of each item in the memory pool. Its

second stage computes a structural similarity alignment and evaluation of the match between the probe and each of the top items from the first stage. There are several appealing features of this model. First, since the initial surface match does not require computing structural consistency, it is computationally cheap; yet, if surface and structural information is correlated (as in ordinary experience), there will be many literal similarity matches. This economy means that the first stage sometimes fails to produce a legitimate structural match, capturing the cases when our subjects considered their own retrievals to be structurally unsound. Second, unlike most case-based retrieval systems, the representations do not need to be indexed by an intelligent agent. The content vector acts as a kind of automatic index—a dumb index, to be sure, since it simply lists all the predicates in the representation, but one that may fit human performance. Third, the system fits the data so far fairly well. It displays the appropriate primacy of mundane and surface matches, and it occasionally retrieves purely relational matches.

*Comparison with current approaches.* MAC/FAC's nearest neighbor among models of memory retrieval is ARCS (Thagard, Holyoak, Nelson, & Gochfeld, 1990). Like MAC/FAC, ARCS assumes that reminding is based on a combination of similarity of concepts and structural isomorphism, although ARCS also assigns a key role in reminding to pragmatic centrality. ARCS sets up a competitive network between the probe and all semantically related items, using excitatory and inhibitory connections to model the above pressures. The item that the network settles on is considered to be the item retrieved. ARCS and MAC/FAC have many similarities, but they differ in a few important ways. MAC/FAC makes a distinction between dumb initial processes in memory retrieval and smart later processes where ARCS does not. MAC/FAC often produces more than one retrieval; ARCS produces only one. ARCS utilizes competition among memory items in a manner integral to the computation to a greater extent than does MAC/FAC. On the other hand ARCS has the disadvantage that a potential for a combinatorial explosion is inherent in setting up full networks between the probe and each of a large pool of memory items. Although MAC/FAC accounts for the present results better than ARCS, more research will be required to sort out their respective advantages.

MAC/FAC can also be compared with two important classes of memory models: case-based reasoning models utilizing abstract structural indices and mathematical memory models utilizing feature vectors. These two approaches have different advantages and disadvantages. Case-based reasoning approaches are good at analogical mapping and inference but fail to show the surface superiority effect in access. Feature-vector models are tractable for large data bases and can capture the lack of structural

effects during access, but have no good way of capturing structural alignment and inference once the material is in working memory. Conceivably a model such as Hintzman's (1986) multiple-trace model might be able to capture both effects by storing surface features in one vector and the structural schema in an associated second vector; but some means of binding the surface arguments to the structural schema would have to be provided. MAC/FAC's hybrid approach, with its cheap initial content-vector stage and its structurally sensitive second stage, combines aspects of both kinds of models.

*Expertise and Relational Access*

Despite the gloomy picture painted in the present research and in most of the problem-solving research, there is evidence of considerable relational access under some circumstances. As Keane (1988) argues persuasively, relational access is most prevalent (a) for experts in a domain and (b) when initial encoding of the study set is relatively intensive. For example, Novick (1988) studied remindings for mathematics problems using novice and expert mathematics students. She found that experts were more likely than novices to retrieve a structurally similar prior problem, and when they did retrieve a surface-similar problem, they were quicker to reject it than were novices. Similarly, J. Clement (1982, 1986) reports that expert physicists often retrieve structural analogies when solving challenging problems. The second contributor to relational retrieval, almost certainly related to the first, is intensive encoding. Faries and Reiser (1988) taught subjects LISP in a series of intensive training sessions and then gave them target problems that were superficially similar to one prior problem and structurally similar to another. Given this intensive training, Faries and Reiser's subjects were able to access structurally similar problems despite the competing superficial similarities. Gick and Holyoak (1988) and Catrambone and Holyoak (1989) found that subjects exhibited increased relational retrieval when they were required to compare two prior analogs, but not when they simply read the two prior analogs. Schumacher & Gentner (1987) found increased relational retrieval of proverbs when subjects wrote out the meaning of each proverb on the subject list, as opposed to simply reading it or rating its cleverness. Seifert, McKoon, Abelson, and Ratcliff (1986) investigated priming effects in a sentence verification task between thematically similar (analogical) stories. They obtained priming when subjects first studied a list of themes and then judged the thematic similarity of pairs of stories, but not when subjects simply read the stories.

The increase of relational reminding with expertise and with intensive encoding can be accommodated in the MAC/FAC model. Assuming that experts' representations are richer and better structured than those of

novices (Carey, 1985; Chi, 1978), then their encodings will contain more higher-order relations and this could promote relational retrieval. For example, there is evidence that teaching children to notice and encode higher-order dimensional relations such as *symmetry* and *monotonicity* increases their ability to appreciate abstract cross-dimensional matches (Kotovsky & Gentner, 1990; in preparation). Equally important for access is the *uniformity of the internal relational encoding* (Gentner & Rattermann, 1991). Whereas novices' knowledge may be only locally coherent, we conjecture that experts may use the same set of theoretical notions across the domain and that this promotes uniform relational encodings in the domain. When a given higher-order relational pattern is used to encode situations, it will of course be automatically incorporated into MAC/FAC's content vectors. This means that any higher-order relational concept that is widely used in a domain will tend to increase the uniformity of the vectors' entries (the "indices" in CBR terms) and therefore the mutual accessibility of situations within the domain. Thus as experts come to encode a domain according to a uniform set of principles, the likelihood of appropriate relational remindings increases. A corollary of this view is that the use of uniform relational labels may be an important contributor to achieving analogical access (Catrambone, in preparation; Clement, Mawby, & Giles, 1991).

As domain expertise increases, MAC/FAC's behavior may come to resemble that of a case-based reasoning model with some surface indexing. We can think of its content vectors as indices with the property that they change automatically with any change in the representation of domain exemplars. Thus as domain knowledge increases, MAC/FAC may have sufficiently elaborated indices to permit a fairly high proportion of relational remindings. The case-based reasoning emphasis on retrieving prior examples and generalizations that are inferentially useful may be a reasonable approximation to the way experts retrieve knowledge.

*Why Surface Access?*

These findings may leave us feeling schizophrenic. How can the human mind, at times so elegant and rigorous, be limited to this primitive retrieval mechanism? An intriguing possibility is that in the evolution of cognition, retrieval from memory is an older process than inferential reasoning over symbolic structures. We could thus think of our surface bias in retrieval as a vestige of our evolutionary past, perhaps even a mistake in design that we have never lived down. But regardless of whether the evolutionary account is correct, there are reasons that a surface bias might be a good design choice for humans. First, there is the *kind world* hypothesis: that in much of our experience surface information is strongly correlated with structural information. If something looks like a tiger, it

probably is one. Perceptual information being plentiful and easy for us to process, it might be a good strategy for a being with a large store of knowledge to take advantage of this correlation (Gentner, 1989; Medin & Ortony, 1989). The second point is the argument from scale. Human knowledge bases are vastly larger than any artificial base yet constructed. It is not clear that a structural indexing scheme would scale up properly to human memory size; we might find ourselves swamped with relational remindings.

A third argument that surface access might sometimes be a good idea rests on considerations of learning, reminding-based generalization and the novice-expert shift. Let's agree that a difference between novices and experts in a domain is that experts know the domain theory: in particular, they know the relational constructs for that domain (Chi, 1978; Chi, Feltovich, & Glaser, 1981; Gentner & Rattermann, 1991). Let us further conjecture that the optimal relational constructs are typically domain-specific and may differ considerably between domains. In such a case, for a learner to move quickly to an abstract structural description could lead to nonrecoverable errors. It would be adaptive for notices to utilize highly conservative encodings with rich information about objects and contexts, and to add relational information gradually. If we further assume that reminding-based generalization can lead gradually to relational abstractions (Elio & Anderson, 1981; Forbus & Gentner, 1986; Gick & Holyoak, 1983; Hayes-Roth & McDermott, 1978; Medin & Ross, 1989; Ross, 1989, in press; Skorstad, Gentner & Medin, 1988) then an initial conservative bias would best allow the relational vocabulary to develop according to the demands of the domain. This brings us to our last conjecture as to why an object-oriented retrieval strategy might be useful: that of incommensurability of representations. Accepting Carey's (1991) arguments that fundamental changes in beliefs can occur with learning and development, then it would be adaptive for human retrieval mechanisms to maintain some reliance on those aspects of internal representations that change least. Assuming that object concepts are relatively stable across theory change, then maintaining some reliance on objects in retrieval preserves access to early learning.

## APPENDIX

### Sample Stimuli from Experiments 1, 2, and 3

*Base Story*

Percy the mockingbird spent the whole warm season chirping and twittering. When it began to get colder Percy visited a squirrel and sang a song for her, expecting to get some of the squirrel's sunflower seeds in return. However, the squirrel was very disappointed in him.

"You are a terrible singer!" she yelled. "I'm not giving you any of my wheat."

A tear rolled down Percy's cheek, and he vowed to give up singing for good.

*Literal-Similarity Match*

A magpie named Sam sang all summer. When winter came he paid a visit to a chipmunk and performed a ballad for her, hoping she would give him some nuts in return. However, the chipmunk was not at all pleased.

"You don't deserve any nuts of mine!" she exclaimed. "Your song was terrible."

*Analogy Match*

Sam travelled all over the world buying beautiful things. When he ran out of money he paid a visit to his mother and gave her a gift he bought while in Tibet, hoping she would give him a loan in return. However, his mother was not at all pleased.

"You don't deserve any money of mine!" she exclaimed. "This is a piece of junk!"

*Surface-Similarity Match*

A magpie named Sam sang all summer. When winter came he paid a visit to a chipmunk. However, the chipmunk was not at all pleased with Sam.

"You have wasted the summer while I have been hard at work!" she said. Sam performed a ballad for her hoping she would give him some nuts in return. But she was still not pleased. "I will not give you any of my nuts!" she exclaimed.

*FOR Match*

Sam travelled all over the world buying beautiful things. When he ran out of money he paid a visit to his mother. However, she was not at all pleased with him.

"While I have been hard at work you have been wasting your time," she said. Sam gave her a gift he bought in Tibet, hoping she would give him a loan in return. But she was still not pleased. "I will not give you any of my hard-earned money!" she exclaimed.

*Objects-Only Match*

One unusually warm spell in February Sam the magpie thought "This is my chance." He stood up on the edge of his nest and trilled proudly. His song was so loud and cheerful that it woke a nearby chipmunk. The chipmunk asked for another song. He was so moved by Sam's talents that he forgot it was still winter and decided to go looking for nuts to store.

## REFERENCES

Anderson, J. R., Farrell, R., & Sauers, R. (1984). Learning to program in LISP. *Cognitive Science*, **8**, 87–129.

Bareiss, R., & King, J. A. (1989). Similarity assessment in case-based reasoning. In *Proceedings: Case-Based Reasoning Workshop*, (pp. 67–71). Pensacola Beach, FL: Kaufmann.

Barsalou, L W. (1982). Context-independent and context-dependent information in concepts. *Memory and Cognition*, **10**, 82–93.

Bassok, M., & Holyoak, K. J. (in press). Pragmatic knowledge and conceptual structure: Determinants of transfer between quantitative domains. In D. K. Detterman, & R. J. Sternberg (Eds.), *Transfer on Trial*. Norwood, NJ: Erlbaum.

Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology*, **20**, 493–523.

Carbonell, J. G. (1986). Derivational analogy: A theory or reconstructive problem solving and expertise acquisition. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 2, pp. 371–392). Los Altos, CA: Kaufmann.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Carey, S. (1991). Knowledge Acquisition: Enrichment or Conceptual Change? In S. Carey & R. Gelman (Eds.), *The Epigenesis of Mind: Essays on Biology and Cognition* (pp. 257–291). Hillsdale, NJ: Erlbaum.

Catrambone, R. (in preparation). The use of labelled subgoal structure in solving probability problems.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **15**, 1147–1156.

Chen, Z., & Daehler, M. W. (1989). Positive and negative transfer in analogical problem solving by 6-year-old children. *Cognitive Development*, **4**, 327–344.

Chi, M. T. H. (1978). Knowledge structures and memory development. In R. S. Siegler (Ed.), *Children's thinking: What develops?* (pp. 73–96). Hillsdale, NJ: Erlbaum.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, **5**, 121–152.

Clement, C. A., & Gentner, D. (1991). Systematicity as a selection constraint in analogical mapping. *Cognitive Science*, **15**, 89–132.

Clement, C. A., Mawby, R., & Giles, D. E. (1991, November). *The effects of manifest relational representation on analog retrieval*. Paper presented at the meeting of the Psychonomics Society, San Francisco, CA.

Clement, J. (1982). Analogical reasoning patterns in expert problem solving. *Proceedings of the fourth meeting of the Cognitive Science Society* (pp. 79–81), Ann Arbor, MI.

Clement, J. (1986). Methods for evaluating the validity of hypothesized analogies. *Proceedings of the Eighth Annual conference of the Cognitive Science Society* (pp. 223–234), Amherst, MA. Hillsdale, NJ: Erlbaum.

Collins, A., & Gentner, D. (1987). How people construct mental models. In D. Holland & N. Quinn (Eds.), *Cultural models of language and thought* (pp. 243–265). Cambridge, England: Cambridge University Press.

Duncker, K. (1945). On problem-solving. *Psychological Monographs*, **58**, (5, Whole No. 270).

Ellis, H. C. (1965). *The transfer of learning*. New York: MacMillan.

Elio, R., & Anderson, J. R. (1981). The effects of category generalizations and instance similarity on schema abstraction. *Journal of Experimental Psychology: Human Learning and Memory*, **7**, 397–417.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1986). The structure-mapping engine.

*Proceedings of the Fifth National Conference on Artificial Intelligence* (pp. 272–277), Philadelphia, PA. Los Altos, CA: Kaufmann.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41, 1–63.

Faries, J. M., & Reiser, B. J. (1988). Access and use of previous solutions in a problem solving situation. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 433–439), Montreal. Hillsdale, NJ: Erlbaum.

Forbus, K. D., & Gentner, D. (1986). Learning physical domains: Toward a theoretical framework. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 2, pp. 311–348). Los Altos, CA: Kaufmann.

Forbus, K. D., & Gentner, D. (1989). Structural evaluation of analogies: What counts? *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 341–348), Ann Arbor, MI. Hillsdale, NJ: Erlbaum.

Forbus, K. D., & Oblinger, D. (1990). Making SME greedy and pragmatic. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 61–68), Cambridge, MA. Hillsdale, NJ: Erlbaum.

Gentner, D. (1980). *The structure of analogical models in science* (Tech. Rep. No. 4451). Cambridge, MA: Bolt Beranek and Newman, Inc.

Gentner, D. (1982). Are scientific analogies metaphors? In D. S. Miall (Ed.), *Metaphor: Problems and perspectives* (pp. 106–132). Brighton, Sussex: Harvester Press.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155–170.

Gentner, D. (1988a). Analogical inference and analogical access. In A. Prieditis (Ed.), *Analogica* (pp. 63–88). Los Altos, CA: Morgan Kaufmann.

Gentner, D. (1988b). Metaphor as structure mapping: The relational shift. *Child Development*, 59, 47–59.

Gentner, D. (1989a). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 199–241). New York: Cambridge University Press.

Gentner, D. (1989b, May). *In Proceedings: Case-Based Reasoning Workshop*. By the Defense Advanced Research Projects Agency Information Science and Technology Office, Pensacola, Florida.

Gentner, D. (1991). Similarity is like analogy. Paper presented at the San Marino conference on similarity, San Marino, CA.

Gentner, D., & Clement, C. (1988). Evidence for relational selectivity in the interpretation of analogy and metaphor. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 22, pp. 307–358). New York: Academic Press.

Gentner, D., Falkenhainer, B., & Skorstad, J. (1988). Viewing metaphor as analogy. In D. H. Helman (Ed.), *Analogical reasoning: Perspectives of artificial intelligence, cognitive science, and philosophy* (pp. 171–177). Dordrecht, The Netherlands: Kluwer.

Gentner, D., & Forbus, K. D. (1991). MAC/FAC: A model of similarity-based access and mapping. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, Chicago, IL.

Gentner, D., & Forbus, K. D. (in preparation). A computational simulation of a two-stage model of similarity in transfer.

Gentner, D., & Gentner, D. R. (1983). Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 99–129). Hillsdale, N.J.: Erlbaum.

Gentner, D., & Landers, R. (1985). Analogical reminding: A good match is hard to find.

*Proceedings of the International Conference on Cybernetics and Society* (pp. 607–613), Tucson/New York: IEEE.

Gentner, D., & Rattermann, M. J. (1991). Language and career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on Thought and Language: Interrelations in Development*, (pp. 225–277). London: Cambridge University press.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, **10**, 277–300.

Gholson, B., Eymard, L. A., Long, D., Morgan, D., & Leeming, F. C. (1988). Problem solving, recall, isomorphic transfer, and nonisomorphic transfer among third-grade and fourth-grade children. *Cognitive Development*, **3**, 37–54.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, **12**, 306–355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, **15**, 1–38.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, **91**, 1–67.

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, **97**, 3–18.

Goldstone, R. L., Gentner, D., & Medin, D. L. (1989). Relations relating relations. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 131–138), Ann Arbor, MI. Hillsdale, NJ: Erlbaum.

Goldstone, R. L., & Medin, D. L. (in press). Similarity, interactive-activation and mapping. In K. J. Holyoak and J. A. Barnden (Eds.) *Advances in Connectionist and Neural Computation Theory: Vol. 2. Analogical Connections*. Norwood, NJ: Ablex.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the non-independence of features in similarity judgments. *Cognitive Psychology*, **23**, 222–262.

Hall, R. P. (1989). Computational approaches to analogical reasoning: A comparative analysis. *Artificial Intelligence*, **39**, 39–120.

Hammond, K. J. (1989). *Case-based planning: Viewing planning as a memory task*. Boston: Academic Press.

Hammond, K. J., Siefert, C. M., & Gray, K. C. (1991). Functionality in analogical transfer: A hard match is good to find. *The Journal of the Learning Sciences*, **1**, 11–152.

Hayes-Roth, F., & McDermott, J. (1978). An interference matching technique for inducing abstractions. *Communications of the ACM*, **21**(5), 401–411.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, **16**, 96–101.

Hintzman, D. (1986). 'Schema abstraction' in a multiple-trace memory model. *Psychological Review*, **93**, 411–428.

Hofstadter, D. R. (1984). *The Copycat project: An experiment in nondeterministic and creative analogies* (M.I.T. A.I. Laboratory Memo 755). Cambridge, MA: M.I.T.

Hofstadter, D., & Mitchell, M. (in press). The copycat project: An overview. In K. Holyoak and J. Barnden (Eds.) *Advances in Connectionist and Neural Computation Theory, Vol. 2: Connectionist Approaches to Analogy, Metaphor, and Case-Based Reasoning*. Norwood, NJ: Ablex.

Holyoak, K. J. (1985). The pragmatics of analogical transfer. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 19, pp. 59–87). New York: Academic Press.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, **15**, 332–340.

Holyoak, K. J., & Thagard, P. (1989a). Analogical mapping by constraint satisfaction. *Cognitive Science*, **13**, 295–355.

Holyoak, K. J., & Thagard, P. R. (1989b). A computational model of analogical problem solving. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 242–266). New York: Cambridge University Press.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, 2, 208–233.

Kass, A. (1989). Strategies for adapting explanations. In *Proceedings, Case-Based Reasoning Workshop*, (pp. 119–123). Pensacola Beach, Florida. Morgan Kaufmann Publishers, Inc.

Kass, A., Leake, D., & Owens, C. (1987). SWALE, a program that explains. In R. Schank (Ed.), *Explanation patterns: Understanding mechanically and creatively*. Hillsdale, NJ: Erlbaum.

Keane, M. (1985). On drawing analogies when solving problems: A theory and test of solution generation in an analogical problem-solving task. *British Journal of Psychology*, 76, 449–458.

Keane, M. T. (1987). On retrieving analogues when solving problems. *Quarterly Journal of Experimental Psychology*, 39, 29–41.

Keane, M. T. (1988). *Analogical problem solving*. Chichester: Ellis Horwood (New York: Wiley).

Kedar-Cabelli, S. (1988). Toward a computational model of purpose-directed analogy. In A. Prieditis (Ed.), *Analogica* (pp. 89–107). Los Altos, CA: Kaufmann.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Kolodner, J. L. (1984). *Retrieval and organizational structures in conceptual memory: A computer model*. Hillsdale NJ: Erlbaum.

Kotovsky, L., & Gentner, D. (in preparation). Progressive alignment: A mechanism for the development of relational similarity.

Lockhead, G. R., & King, M. C. (1977). Classifying integral stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, 3, 436–443.

Markman, A. B., & Gentner, D. (1990). Analogical mapping during similarity judgments. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 38–44), Cambridge, MA. Hillsdale, NJ: Erlbaum.

Markman, A. B., & Gentner, D. (in press). Structural alignment during similarity judgments. *Cognitive Psychology*.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: Freeman.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1990). Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science*, 1, 64–69.

Medin, D. L., Goldstone, R. L., & Gentner, D. (in press). Respects for similarity. *Psychological Review*.

Medin, D., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). New York: Cambridge University Press.

Medin, D. L., & Ross, B. H. (1989). The specific character of abstract thought: Categorization, problem-solving, and induction. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 5, pp. 189–223). Hillsdale, NJ: Erlbaum.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85, 207–238.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92, 289–316.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**, 510–520.

Ortony, A. (1979). Beyond literal similarity. *Psychological Review*, **86**, 161–180.

Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. *Psychological Review*, **56**, 132–143.

Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology*, **9**, 441–474.

Palmer, S. E. (1978). Structural aspects of visual similarity. *Memory & Cognition*, **6**, 91–97.

Palmer, S. E. (1989). Levels of description in information-processing theories of analogy. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 332–345). New York: Cambridge University Press.

Palmer, S. E., & Kimchi, R. (1985). The information processing approach to cognition. In T. Knapp & L. C. Robertson (Eds.), *Approaches to cognition: Contrasts and controversies* (pp. 37–77). Hillsdale, NJ: Erlbaum.

Pirolli, P. (1985). *Problem solving by analogy and skill acquisition in the domain of programming*. Unpublished manuscript.

Pomerantz, J. R., Sager, L. C., Stoever, R. J. (1977). Perception of wholes and of their component parts: Some configural superiority effects. *Journal of Experimental Psychology: Human Perception and Performance*, **3**, 422–435.

Ratcliff, R., & McKoon, G. (1989). Similarity information versus relational information: Differences in the time course of retrieval. *Cognitive Psychology*, **21**, 139–155.

Ratcliff, R., & Murdock, B. B., Jr. (1976). Retrieval processes in recognition memory. *Psychological Review*, **83**, 190–214.

Rattermann, M. J., & Gentner, D. (1987). Analogy and similarity: Determinants of accessibility and inferential soundness. In *Proceedings of the Ninth Annual Conference of the Cognitive Science Society* (pp. 23–35).

Read, S. J. (1983). Once is enough: Causal reasoning from a single instance. *Journal of Personality and Social Psychology*, **45**, 323–334.

Read, S. J., (1984). Analogical reasoning in social judgment: The importance of causal theories. *Journal of Personality and Social Psychology*, **46**, 14–25.

Reed, S. K. (1987). A structure-mapping model for word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **13**, 124–139.

Reed, S. K., Ernst, G. W., & Banerji, R. (1974). The role of analogy in transfer between similar problem states. *Cognitive Psychology*, **6**, 436–450.

Reisbeck, C. K., & Schank, R. C. (1989). *Inside case-based reasoning*. Hillsdale, NJ: Erlbaum.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59). New York: Cambridge University Press.

Ross, B. H. (1984). Remindings and their effects in learning a cognitive skill. *Cognitive Psychology*, **16**, 371–416.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **13**, 629–639.

Ross, B. H. (1989). Remindings in learning and instruction. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 438–469). New York: Cambridge University Press.

Ross, B. H. (1991). Some influences of instance comparisons on concept formation. In D. H. Fisher, Jr., M. J. Pazzani, & P. Langley (Eds.), *Concept Formation: Knowledge and Experience in Unsupervised Learning*, (pp. 207–236). San Mateo, CA: Kaufmann.

Ross, B. H. (in press). Access and use of relevant information: A specific case and general issues. In R. Freedle (Ed.), *AI and the future of testing*. Hillsdale, NJ: Erlbaum.

Schank, R. C. (1982). *Dynamic memory*. New York: Cambridge University Press.

Schumacher, R. M., & Gentner, D. (1987, May). *Similarity-based remindings: The effects of similarity and interitem distance.* Paper presented at meeting of the Midwestern Psychological Association, Chicago, IL.

Schumacher, R. M., & Gentner, D. (1988a). Remembering causal systems: Effects of systematicity and surface similarity in delayed transfer. *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 1271–1275). Anaheim, CA. Santa Monica, CA: Human Factors Society.

Schumacher, R. M., & Gentner, D. (1988b). Transfer of training as analogical mapping. *IEEE Transactions on Systems, Man, and Cybernetics,* **18,** 592–600.

Schumacher, R. M., & Gentner, D. (in preparation). Analogical access: Effects of repetition, competition, and instruction.

Seifert, C. M., McKoon, G., Abelson, R. P., & Ratcliff, R. (1986). Memory connections between thematically similar episodes. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **12,** 220–231.

Simon, H. A., & Hayes, J. R. (1976). The understanding process: Problem isomorphs. *Cognitive Psychology,* **8,** 165–190.

Skorstad, J., Falkenhainer, B., & Gentner, D. (1987). Analogical processing: A simulation and empirical corroboration. *Proceedings of the Sixth National Conference on Artificial Intelligence* (pp. 322–326), Seattle, WA. Los Altos, CA: Morgan Kaufmann.

Skorstad, J., Gentner, D., & Medin, D. (1988). Abstraction processes during concept learning: A structural view. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 419–425), Montreal, Canada.

Thagard, P. (1988). Dimensions of analogy. In D. H. Helman (Ed.), *Analogical reasoning: Perspectives of artificial intelligence, cognitive science, and philosophy* (pp. 105–124). Dordrecht, The Netherlands: Kluwer.

Thagard, P., Holyoak, K. J., Nelson, G., & Gochfeld, D. (1990). Analog retrieval by constraint satisfaction. *Artificial Intelligence,* **46,** 259–310.

Thorndike, E. L. (1903). *Educational psychology.* New York: Lemcke and Buechner.

Wharton, C. M., Holyoak, K. J., Downing, P. E., Lange, T. E., & Wickens, T. D. (1991). Retrieval competition in memory for analogies. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society,* (pp. 528–533). Hillsdale, NJ: Erlbaum.

Wharton, C. M., Holyoak, K. J., Downing, P. E., Lange, T. E., & Wickens, T. D. (1992). The story with reminding: Memory retrieval is influenced by analogical similarity. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society.*

Winston, P. H. (1982). Learning new principles from precedents and exercises. *Artificial Intelligence,* **19,** 321–350.

(Accepted January 5, 1993)