Intermediate Features and Informational-level Constraint on Analogical Retrieval

Mark Alan Finlayson
Patrick Henry Winston
Computer Science and Artificial Intelligence Laboratory,
Massachusetts Institute of Technology,
32 Vassar St., Cambridge, MA 02139 USA
{markaf, phw}@csail.mit.edu

Abstract

Two different sorts of retrieval in analogical tasks, novice-like and expert-like, have been demonstrated in psychological experiments. With recent computational work in object recognition as an inspiration, we propose the computation of intermediate features, and their use as triggers for retrieval, as the relevant constraint at the informational level of characterization of the retrieval process as seen in experts. We conduct computational experiments which show that, in conjunction with a feature-matching retrieval mechanism, features of an intermediate size and complexity give the strongest analogical retrieval.

An Intriguing Difference in Retrieval

The retrieval of relevant precedents is commonly considered a critical first step in the analogical reasoning process (French, 2002). Psychologists have studied retrieval in detail, using a straightforward experimental paradigm; in this paradigm, referred to by some as the "And-now-for-something-completely-different approach" (Brown, 1989), subjects are given some set of source tasks (such as memorizing a set of stories, or solving a problem), and then after some delay are given a second set of target tasks that are related to the source tasks in an analogical fashion. The experiments then test if the subjects are able to perform the target tasks by analogy with the source tasks.

Studies of this sort have characteristically provided strong evidence that most people do not retrieve analogically profitable items from memory, even when delays between source and target tasks are short, or when analogical relationships are especially strong. The seminal demonstration of this was by Gick and Holyoak (1980, 1983) who reported that nearly two-thirds of their subjects were unable to spontaneously retrieve analogous source problems. Since then, a wide variety of studies have provided strong evidence for people's inability to spontaneously recall relevant analogs. Gentner and Landers (1985) conducted a story recall experiment from which they concluded that, rather than retrieving on the basis of analogical relatedness, people retrieve on the basis of surface similarity—i.e., the characteristics or properties of actors and objects involved in the description. Rattermann and Gentner (1987) went on to show that object-descriptions and first-order relations between objects promote retrieval, but higher-order relations do not, and that preference of retrieval and ratings of stories for inferential soundness are negatively correlated. Other problem solving and retrieval studies showed that subjects needed to be explicitly informed of the relation between two problems before they were able to apply analogical inferences, and that recall is heavily dependent on surface semantic or syntactic similarities between representations (Reed, Ernst, & Banerji, 1974; Reed, Dempster, & Ettinger, 1985; Ross, 1984, 1987). The pattern of retrieval shown by these experiments is clear: in uncontrolled populations, analogically related items are not preferred in retrieval.

In contrast, evidence drawn from the literature on experts suggests that a high level of skill and training in a particular domain can allow for the recall of analogically related knowledge. For example, in a classic study, Chi and colleagues (1981) demonstrated that physics experts (advanced graduate students in the area) categorize on the basis of abstract physics principles, whereas novices categorize on the basis of literal features. Shafto and Coley (2003) have shown similar effects with college students versus commercial fisherman in the categorization of marine creatures. Novick (1988) showed that experts compared to novices are more likely to demonstrate spontaneous analogical transfer when problems share structural features. These categorization experiments are not exactly parallel to the retrieval experiments described above; the single study on expert retrieval we were able to locate is by Shneiderman (1977) which showed that expert computer programmers recalled computer programs primarily based on the purpose of the code, but not on its specific form, while novices were heavily influenced by trivial syntactic constructions. For the purposes of this paper, we will take as a given the natural conjecture that experts are better than novices at retrieving useful analogical precedents within their domain of expertise, with the caveat that this supposed difference may be upended by future surprising experimental results.

Explanation at the Informational Level

This supposed difference in recall between experts and novices is intriguing. How might it be explained? A simple hypothesis is that immediate, visceral recall from memory is a relatively automatic process, and the actual mechanism does not vary substantially from person to person. Instead, the difference between expert-like and novice-like retrieval would be in the parameters of the process; in other words, the algorithm is the same, but the constraints on algorithm are different. This so-called *informational* or *computational* level treatment of the problem, an approach outlined explicitly by Marr (1982), has already been profitably applied in the study of analogy, in particular to models of the construction of an analogical mapping (Keane, Ledgeway, & Duff, 1994; Palmer, 1989). With respect to retrieval, we ask the question, 'what are the information-level constraints that allows experts to efficiently retrieve analogs?' That is, what is it that experts are doing (or computing) which allows them to effect analogical retrieval? Our proposal is that retrieval occurs by means of a feature-matching process, and that when a person shows expert-like retrieval, what they are doing is calculating and using what we will call *intermediate features*. What are intermediate features? They may be explained best, perhaps, by an analogy with previous work.

Intermediate Features in Visual Recognition

Our inspiration for pursuing intermediate features for analogical recall was recent work on visual image classification by Ullman, Vidal-Naquet, and Sali (Ullman, Vidal-Naquet, & Sali, 2002). Ullman and coworkers noted that the human visual system assembles complex representations of images from simpler features, but that it is still an open question how these complex representations are used in visual processing. With this in mind, they showed that, from an information-theoretic point of view, features of an intermediate size and complexity are best for the basic visual task of classification; for example, to identify a face in an image, looking for face fragments of intermediate size (such as a pair of eyes) is more useful than looking for small features (an eye) or large features (a complete face).

In their experiment they searched a database of approximately 180 faces and automobiles for approximately 50 selected face fragments, and measured the mutual information delivered by each fragment. In this context, a feature that yields a great deal of mutual information was a feature which, if present, was a good indicator that the class was present as well.

From this perspective, Ullman and coworkers found that fragments of an intermediate size and complexity maximized the mutual information. Leveraging this knowledge, they designed a detection scheme that weighted matches of intermediate features more heavily than matches of either small or large features, and produced a 97% face detection rate in novel images, with only a 2.1% false positive rate.

They intuitively account for their detection scheme's impressive results by noting that small, blurred features produce many false positives (a blurred eye often matches a random image feature by chance) and that large, complex features produce many false negatives (a detailed image of a face rarely matches anything in a small collection of stored faces). It is rather the features that are only somewhat blurred and of a intermediate size that are most useful for identification.

Intermediate Features in Symbolic Representations

By analogy with visual recognition, we can think of features in the symbolic domain as portions of a description. Our proposal is thus that retrieval occurs by a feature-matching process (Holyoak & Koh, 1987) and that it is the size of the features which vary between experts and novices. If we think, as is common in computational cognitive modelling, of the cognitive descriptions as graph-like representational structures, a feature would be a collection of nodes from that description, and an *intermediate* feature would be a collection of nodes which was not too small or too large relative to the whole. This notion is directly related to Gentner's notion of zero-, first- and higher-order nodes in a symbolic representation (Gentner, 1983). 'Small' features would have nodes of low order in them; 'intermediate' features might contain first- and second- order nodes plus their descendants (i.e., they would be first- and second- order systematic representations, as Gentner might say), and large features would include high-order nodes plus descendants. A few examples of fragments of various sizes are shown in Figure 1, where the description of the orbit of a planet around a sun is split into small fragments (such as *sun* and *planet*), intermediate fragments (such as the *greater-than* relation and its children nodes) and large fragments (such as the whole description).

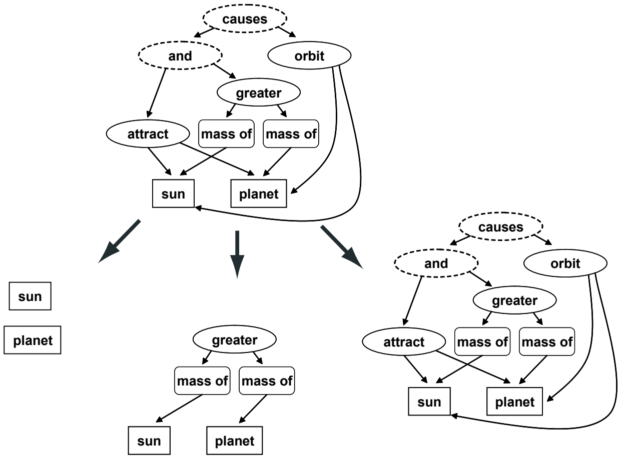

Figure 1: Small, intermediate, and large-sized features in a symbolic representation. Two features of small size, drawn from the overall description, appear on the left. A feature of intermediate size, shown in the middle, would in this case consist not only of two objects but also a relation between them. A feature of large size, shown on the right, might be the whole description. We argue that features of intermediate size are essential when modelling precedent retrieval at the level of human experts.

Also by analogy with Ullman's visual recognition scheme, we can conceive of a feature-matching retrieval scheme as one which takes a feature of a description (i.e., a collection of nodes) and "looks" for this feature in other description graphs in memory. Those sources in memory which best match the set of input features would be retrieved most strongly. We hypothesize that to achieve either novice-like or expert-like retrieval, we merely change the average size of the feature used for retrieval. Novice-like retrieval would use small features, and expert-like retrieval would use intermediate features. In other words, novice-like and expert-like retrieval are achieved with the same mechanism, but with different constraints.

Experimental Results

To lend support to our hypothesis we have constructed an implementation of a feature-matching retrieval scheme in which the feature size can be varied, and ran computational experiments that have shown that analogical retrieval occurs preferentially when intermediate features are used. Our expected result is to show that, with a generalized feature-matching algorithm, if one retrieves with small features, one achieves a novice-like pattern of retrieval, and if one retrieves with intermediate-sized features, one achieves an expert-like pattern of retrieval.

Outline of a Test Algorithm

The algorithm that was run is outlined in Figure 2. Each feature is a subtree of the description graph, that is, a node plus descendants. Match scores for feature pairs are produced by a simple tree alignment, which starts by matching the head nodes, measuring their semantic similarity, and then proceeding recursively down the tree, stopping when two nodes are incompatible (e.g., have different numbers of arguments) or the features are exhausted. The threshold in the algorithm is used to vary the average size of the features that are involved in matching. It does not control the size of the features directly but rather eliminates feature pairings based on the pair's match score. When two features match well, their match score is proportional to their size, so the threshold loosely controls the size of the features which participate in matching. This technique was used so as to eliminate highly-uninformative features from the match pool at low thresholds.

This algorithm is intended to be a generalized feature-matching algorithm. We do not claim that this is *the* algorithm used for the process, but rather it is our aim to demonstrate that by using a feature-matching algorithm and concentrating on intermediate features, one can move from novice-like to expert-like retrieval.

Producing a Test Dataset

Like previous work from our laboratory on analogical reasoning, our implementation takes near-natural English paraphrases of situations and automatically produces a description graph on which all subsequent algorithms are run (Winston, 1980, 1982). In these graphs, nodes represent objects or concepts, and edges represent a simple "argument-of" relation. The stories averaged 16 sentences (deviation of 5.3) and contained on average 65 nodes (deviation 20.2). Our representation incorporates a rough model of semantics called *thread memory* (Vaina & Greenblatt, 1979), where to each node is attached a collection of ordered lists (*threads*) of hierarchy terms to which the object or concept represented by the node belongs. In other words, every node, whether it represents an object, a relation, or something else, maintains one or more sequences of class membership strings, and each such sequence is called a thread. For example, a person might be described with the thread `ambassador--diplomat--politician--human--primate--animal--thing` which indicates that the most specific class of which they are member is the `ambassador` class, and the most general class the `thing` class. Thus, comparing with another person, say a fireman: `fireman--rescue-worker--human--primate--animal--thing` we find that they match on the last four classes, but not the others. By counting the number of elements in common between two threads we can get a rough measure of their semantic similarity. [1] As can be seen, our representation is highly similar to others used in research on analogy that it encompasses both episodic memory implemented as a graph (here, nodes and their graph structure), and a type of semantic memory implemented as frames attached to nodes in that graph (here, thread memory) (Thagard, Holyoak, Nelson, & Gochfeld, 1990).

Basic Feature-matching Retrieval Test Algorithm

```
1. Break the target $p$ and source $t$ into all possible
   features $\{F_P\}$ and $\{F_t\}$.
2. $\forall f_p \in F_P$, from largest to smallest
   (a) Measure the pairwise match score $\forall(f_p, f_t) : f_t \in F_t$.
   (b) Take the best feature match and, if it is above
       the threshold, add its score to the total for $t$, and
       remove the corresponding $f_t$ from $F_T$.
3. Return the total score for $t$.
```

Figure 2: Description of the algorithm which is used to demonstrate the utility of intermediate features for retrieving analogical sources from memory. The algorithm takes as input a target and source description and a threshold, and returns a retrieval score. To judge what is retrieved from a source set given a target description, the algorithm is run on each source in the set and the scores are compared.

To guarantee that our source memory contains matches to our targets of the proper character, that is, those that resemble the sorts of sources used in psychological experiments, we synthesized sources by systematic transformation of nodes and threads in the tar-

[1] In particular, our algorithm takes the number of thread elements in common over the number of distinct thread elements between two nodes as the semantic similarity between two nodes, a number that runs between 0 and 1.

get. For each target, we made four sources from its description: an analogically related match (AN), a less-analogically related match (LAN), a mere-appearance match (MA), and a literally-similar match (LS).[2] For example, suppose we began with a target description "The boy ran away to the rock and hid because the girl threatened him." To make a literally-similar match to this target, we replace the objects in the situation with nearly similar or identical objects, while leaving the structure unchanged. In our implementation we replaced each object with another object which matches on all but the last thread element. Thus we might replace *boy* with *man* and *girl* with *woman* and *rock* with *boulder* to produce "The man ran to the boulder and hid because the woman threatened him." To obtain a merely-apparently similar source, we leave the objects unchanged, but scramble the higher-order structure. This means that we take higher-order nodes of the same order and randomly switch their subjects and objects. This might produce "The girl threatened the boy because he ran to the rock and hid." An analogical match involves different objects, but the same sorts of relations. To effect this we replaced all the objects in the target with objects which matched on only highest membership classes, while leaving the higher-order structure unchanged. Thus a generated analogical source might be "The army retreated to the castle and dug in because the enemy army approached." To make what we call a less-analogically similar match, we transform as if to make an analogy, but we mix up the subjects and objects of some fraction of higher-order relations as is done for a mere-appearance match.[3] This might produce "The army returned to their castle, but the enemy only approached when they dug in." Table 1 summarizes the different sorts of source types and their examples.

Type	Example
Probe	The boy ran away to the rock and hid because the girl threatened him.
LS source	The man ran to the boulder and hid because the woman threatened him.
MA source	The girl threatened the boy because he ran to the rock and hid.
AN source	The army retreated to the castle and dug in because the enemy army approached.
LAN source	The army returned to their castle, but the enemy only approached when they dug in.

Table 1: Examples of systematic derivations of literally similar (LS), merely-apparently similar (MA), analogical (AN), and less-analogical (LAN) source matches to a target.

Following on Gentner's retrieval results (Rattermann & Gentner, 1987), we expect that novice-like retrieval will result in LS probes being most preferred, followed by MA, AN and LAN in that order. Thus the novices prefer superficially-similar stories (MA) to analogically-related stories (AN and LAN). Expert-like retrieval would also place LS probes first, but would prefer AN and LAN probes before MA probes.

Experiment 1

Experiment 1 demonstrates that both novice-like and expert-like retrieval can be achieved with variation of the feature-size parameter of a feature-matching mechanism. In this experiment we run our demonstration algorithm and vary its single parameter, the threshold, which loosely controls the size of the features used in matching.

Our dataset consisted of fourteen story descriptions provided in near-natural simple English. The fourteen story descriptions were first parsed from simple English into our graph representations. These descriptions were used as the targets. Each description was used to generate the four related descriptions (literally-similar, mere-appearance, analogical, and less-analogical), and these 56 descriptions were used as sources. The retrieval algorithm was run with each target, and the retrieval score was measured between the target and its related sources. The results of each source type was averaged across all the targets for each threshold and normalized, and the order of retrieval was compared to the predicted order for novice-like (LS > MA > AN > LAN) and expert-like (LS > AN > LAN > MA) retrieval, resulting in two confidence curves shown in Figure 3.

The curves are calculated as follows: the novice or expert retrieval patterns differ in the position of the mere-appearance source. Each source score which was in the correct order relative to its associated mere-appearance source score was given a confidence of 1 (i.e., a correct prediction). If in the incorrect order, it was given a confidence of 0. If the scores were equal, they were given a confidence of 0.5 (fifty percent chance of choosing the correct order). These three confidence values were then averaged to obtain the probability of making a correct order prediction given the retrieval scores assigned by the algorithm.

These curves show the probability of predicting the correct human retrieval order given the retrieval scores provided by the algorithm. As can be seen, the novice order is well predicted by the scores produced at a low threshold, that is, at a low feature size. The expert order is well predicted at intermediate feature sizes. We see the effect anticipated, namely a novice-like retrieval pattern at low feature sizes, and an expert-like retrieval pattern at intermediate features sizes.

Experiment 2

Experiment 2 shows that higher-order features do not contribute significantly to novice-like retrieval, and fur-

[2] Note that the AN, MA, and LS match types are not unique to our work, but follow on source types established in the analogy and retrieval literature (Gentner, 1983; Gentner & Landers, 1985).

[3] For the experiments presented in this paper, the fraction was approximately one-third.

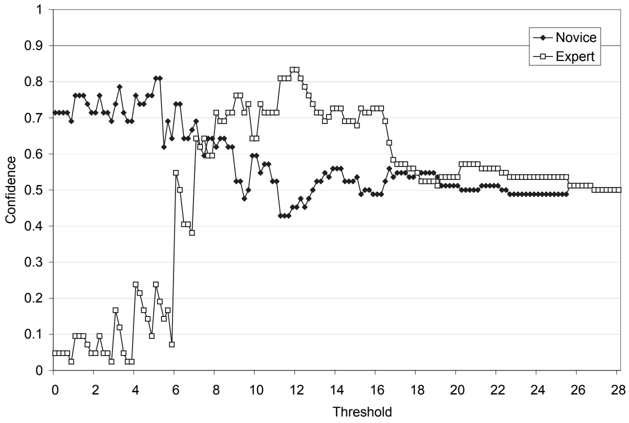

Figure 3: Probability of predicting the correct ordering (either Novice or Expert) averaged over the dataset, against threshold. The behavior of our test program becomes more like that of human experts when features of small drop out allowing features of intermediate size to dominate matching, at a threshold of about eight, but then the behavior degrades when intermediate features drop and only larger features participate, at a threshold of about 18. On the other hand, the behavior of our test program is most like that of a human novice when features of only small size participate in matching. The abscissa runs until the threshold is larger than the score of any feature pair.

thermore that intermediate-features are alone responsible for expert-like retrieval. In the previous experiment, all feature pairs which have score above a certain threshold are allowed to count toward a source's total retrieval score. Thus, as the threshold is raised, so too is the average score, and the average feature size. However, when the threshold is low, higher scoring matches also contribute to the retrieval scores. According to our conception of the intermediate features constraint, novices are characterized by their inability to access, index, form, or use these higher-order feature pairs. Thus this experiment investigates whether higher-order features effect the novice-like retrieval pattern. This experiment used the same dataset and procedure as the previous two experiments, with the exception that the algorithm was changed slightly to reject matches with a score *higher* than the threshold, so that the threshold could be dropped from above and we could investigate the retrieval pattern as participating feature matches were restricted to smaller and smaller features.

As can be seen, the retrieval pattern of the algorithm is novice-like until features of small size no longer participate in matching, allowing intermediate features to dominate. As the threshold drops from above, the novice retrieval pattern is maintained until features with small scores begin to be discarded, at which point the pattern becomes degraded and extremely noisy. This confirms that higher-order features do not significantly contribute to the novice-like retrieval pattern. Furthermore, because intermediate-features and small-features together do not produce expert-like retrieval (Experiment 2), and neither large-features (Experiment 1) nor small features (Experiment 2) alone do not produce expert-like retrieval, we conclude that intermediate features alone are responsible for expert-like retrieval.

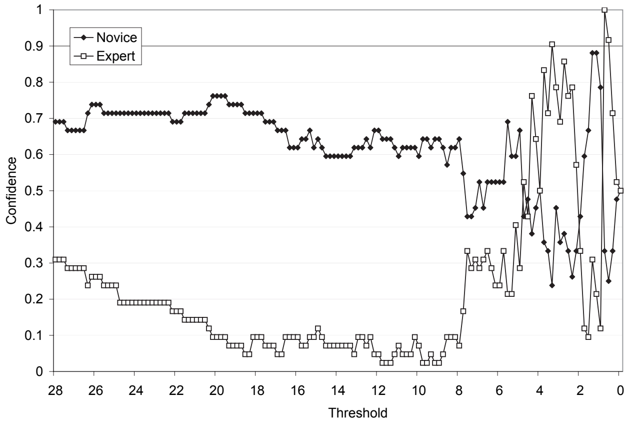

Figure 4: Probability of predicting the correct ordering averaged over the dataset, as the threshold is dropped from above. Note that the abscissa is reversed so that the left side of the figure indicates the same condition as the left side of Figure 3, i.e., no features excluded. As expected, the behavior of our test program mirrors human novices until features of small size no longer participate in matching, at which point the results fluctuate widely because relatively few features participate in the match.

Discussion

Our work speaks to a computational-level account of the retrieval phenomenon and does not commit us to a particular implementation at the algorithmic level. For example, we can readily imagine efficient processes that select appropriate intermediate-level features and compress them into single nodes; this would bring our model into alignment with fast algorithms based on feature-vector comparisons (Forbus, Gentner, & Law, 1994). Such an approach would explain why novices cannot simple tell themselves to retrieve on intermediate size chunks; they lack the apparatus for selecting and compressing intermediate features.

Alternatively, in an implementation based on a constraint-satisfaction network (Thagard et al., 1990), intermediate-sized representation pieces could be implemented by applying a feature-size filter to the construction of the nodes in the constraint network. Then, when the network is run, intermediate features would dominate the retrieval of sources, and the system would accomplish expert-like retrieval.

For a hybrid system such as Kokinov's (Kokinov, 1994), our results might indicate the proper balance between the amount of structure and amount of semantics used in the construction of a representation intended for expert-like retrieval. For a system based on dynamic binding of representational structures (Hummel & Holyoak, 1997), intermediate features might indicate the level at which representational elements should first be synchronized.

Contributions

First, at the information-level, we supported the view that both novice-like and expert-like retrieval are manifestations of a single, parameterized feature-matching mechanism.

Second, we implemented a representative algorithm that embodied, in a transparent fashion, the basic informational-level principles at the root of the hypothesis.

Finally, we conducted experiments with that algorithm that showed that it can achieve both novice-like retrieval via small features and expert-like retrieval via intermediate-sized features.

Acknowledgments

Thanks to all the members of the Genesis group for their assistance and constructive critiques. This work is supported in part by the National Science Foundation under Grant Numbers 0218861 and IIS-0413206.

References

Brown, A. L. (1989). Analogical learning and transfer: What develops? In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning.* Cambridge: Cambridge University Press.

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.

Forbus, K. D., Gentner, D., & Law, K. (1994). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, *19*, 141-205.

French, R. M. (2002). The computational modeling of analogy-making. *Trends in Cognitive Sciences*, *6*, 200-205.

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

Gentner, D., & Landers, R. (1985). Analogical reminding: A good match is hard to find. In *Proceedings of the international conference on systems, man and cybernetics* (p. 607-613). Tucson, AZ: IEEE.

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.

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

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*, 427-466.

Keane, M. T., Ledgeway, T., & Duff, S. (1994). Constraints on analogical mapping: A comparison of three models. *Cognitive Science*, *18*, 387-438.

Kokinov, B. N. (1994). A hybrid model of reasoning by analogy. In J. A. Barnden & K. J. Holyoak (Eds.), *Analogical connections* (Vol. 2, p. 247-318). Norwood, NJ: Ablex Publishing.

Marr, D. (1982). *Vision.* W.H. Freeman and Company.

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

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

Rattermann, M. J., & Gentner, D. (1987). Analogy and similarity: Determinants of accessibility and inferential soundness. In *Proceedings of the annual conference of the cognitive science society* (Vol. 9, p. 23-35). Lawrence Erlbaum Associates.

Reed, S. K., Dempster, A., & Ettinger, M. (1985). Usefulness of analogous solutions for solving algebra word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 106-125.

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

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.

Shafto, P., & Coley, J. D. (2003). Development of categorization and reasoning in the natural world: Novices to experts, naive similarity to ecological knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 641-649.

Shneiderman, B. (1977). Measuring computer program quality and comprehension. *International Journal of Man-Machine Studies*, *9*, 465-478.

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

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, *5*, 682-687.

Vaina, L., & Greenblatt, R. (1979). *The use of thread memory in amnesia aphasia and concept learning* (Working Paper No. No. 195). MIT Artificial Intelligence Laboratory.

Winston, P. H. (1980). Learning and reasoning by analogy. *Communications of the ACM*, *23*, 689-703.

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