Proceedings of CogSci89

Structural Evaluation of Analogies: What Counts?

Kenneth D. Forbus Dedre Gentner
Qualitative Reasoning Group Psychology Department

Beckman Institute, University of Illinois

Abstract: Judgments of similarity and soundness are important aspects of human analogical processing. This paper explores how these judgments can be modeled using SME, a simulation of Gentner's structure-mapping theory. We focus on structural evaluation, explicating several principles which psychologically plausible algorithms should follow. We introduce the *Specificity Conjecture*, which claims that naturalistic representations include a preponderance of appearance and low-order information. We demonstrate via computational experiments that this conjecture affects how structural evaluation should be performed, including the choice of normalization technique and how the systematicity preference is implemented.

1 Introduction

Judging soundness and structural similarity are important aspects of human analogical processing. While other criteria (such as factual correctness and relevance to current goals) are also important, they cannot replace structural evaluation. For example, neither factual correctness or relevance are enough when an analogy is used to make an argument; the claimed consequences must legitimately follow from the analogy or the argument will be rejected. The importance of structural evaluation is even clearer when one considers the use of analogy to discover new ideas: the learner must have some means of judging the comparison without knowing in advance if its implications are correct or relevant.

We have suggested that human structural evalution of analogies depends largely on the degree to which the analogs share systematic relational structure (i.e., share systems of relations governed by common higher-order relations) [8]. There is psychological evidence supporting this position as a descriptive account [10]. SME[5,6], our simulation of Gentner's structure-mapping theory [7,8], includes a structural evaluator which appears to match psychological data on analogical soundness judgments reasonably well [17]. In this paper we use a combination of theoretical argument and sensitivity analyses to probe more deeply into the issues surrounding structural evaluation.

Section 2 begins with a brief overview of SME and outlines some constraints on psychologically plausible algorithms for structural evaluation. Section 3 summarizes psychological results concerning analogical soundness, and shows how our prior simulation experiment provides a framework for sensitivity analyses. Section 4 proposes that the representations used in AI and cognitive simulation tend to be unrealistically sparse (the *Specificity Conjecture*). The next two sections demonstrate how this conjecture constrains structural evaluation algorithms. Two design dimensions are considered. Section 5 compares alternate normalization strategies (i.e., how evidence is combined). Section 6 compares our original cascade-like technique for implementing the systematicity preference (*trickle-down*) with another technique, *order-scoring*. We conclude that trickle-down with result normalization provides the best fit to human data. We close by considering the broader implications of the Specificity Conjecture.

2 The Structure-Mapping Engine

SME was designed to provide an *accountable* simulation of Gentner's structure-mapping theory. By accountable, we mean that processing choices not explicitly constrained by the theory must be easily changable, so that dependence on alternate choices can be explored. To achieve accountability, SME's input includes two sets of rules which construct and evaluate local matches. By varying these rules SME can be programmed to emulate all the comparisons of structure-mapping, as well other matchers consistent with its assumptions [4]. Here we use this programmability to perform sensitivity analyses to rule out certain processing choices as being unable to account for human data.

Given *base* and *target* descriptions to match, SME produces a set of *Gmaps*, representing the possible interpretations of the comparison. Each Gmap includes a set of correspondences between the items (objects and propositions) in the base and target, the set of *candidate inferences* sanctioned by the match (i.e., knowledge about the base conjectured to hold in the target by virtue of the correspondences), and a *structural evaluation score* (SES) indicating the "quality" of the match.

SME begins by computing local *match hypotheses* (MH's) involving pairs of items from base and target. The construction rules guide this process[1]. At this stage the match is incoherent, in that the set of match hypotheses collectively can contain many-to-one mappings. Local constraints, such as one-to-one mappings and structural consistency (see [6] for details) are enforced next. These constraints rule out match hypotheses which cannot be part of any legal interpretation, and note which pairs of match hypotheses cannot consistently be part of the same interpretation. Gmaps are built by finding the maximal structurally consistent collections of local matches, and using the computed overlap to determine what non-overlapping aspects of the base can be postulated to hold in the target (i.e., the candidate inferences). The structural evalutation score is computed last. First, the evaluation rules are run to provide a score for each match hypothesis. The SES of each Gmap is computed by adding the scores of its match hypotheses.

SME provides a process model for structure-mapping. The goal is to achieve sophisticated results using computationally simple techniques. We believe that combining local match hypotheses into coherent global interpretations is a psychologically plausible aspect of SME [9]. However, not every aspect of SME is equally plausible psychologically. For example, we do not believe people necessarily compute all interpretations, although for experimental purposes we generally have SME compute the complete set of Gmaps to gain more insight into the match. A second limitation is that SME models only the structural component of match quality. Contextual and pragmatic factors can also play a role in match evalution. However, understanding those factors involves simulating larger pieces of the overall processing system, with a subsequent increase in the number of free parameters. By understanding structural evaluation in isolation we hope to tightly constraint that aspect of the system.

Structure-mapping postulates that *systematicity* is preferred in structural evaluations [8]; i.e., a Gmap involving a larger connected system of relations, particularly higher-order relations[2], should have a higher SES than one involving a smaller, or disconnected, system of relations. The systematicity constraint is stated at the information processing level (as defined by Marr [14]); additional principles are needed to provide constraint at the algorithm and implementation levels. This paper focuses on the algorithm level, importing only the most general constraints from the prospect of highly parallel, neural-like implementations. Our current *implementation* is serial, but that is an accident of technology – the SME algorithm lends itself naturally to a variety of parallel implementations [6][3].

The score associated with a match hypothesis indicates how strongly the correspondence between the base and target items it connects is preferred on structural grounds. We restrict evaluation rules to use only local, structural properties in assigning scores. For example, MH's receive some initial score based on the kinds of items matched (relation, function, or attribute). Under structure-mapping only propositions involving identical relations or attributes match[4], so the same initial score is used for all relations and attributes (i.e., matches involving relations such

[1] Which pairs of items are hypothesized to match and the structural constraints defining consistent global interpretations are fixed by structure-mapping theory.

[2] Structure-mapping defines the *order* of an item in a representation as follows: Objects and constants are order 0. The order of a predicate is one plus the maximum of the order of its arguments. Thus GREATER-THAN(x,y) is first-order if x and y are objects, and CAUSE[GREATER-THAN(x,y), BREAK(x)] is second-order. Examples of higher-order relations include CAUSE and IMPLIES.

[3] We view Holyoak and Thagard's ACME [11] as evidence that SME could be implemented in at least a localist connectionist framework, since there is substantial overlap in the information processing and algorithm levels between SME and ACME.

[4] We assume a decompositional semantics, so that synonyms are translated into some common form (c.f. [1,3]). This allows similarity to be reduced to partial identity. The alternative course of allowing similar predicates to match requires one to define similarity by invoking it.

as CAUSE are given the same score as matches involving relations such as IMPLIES or LEFT-OF). This parameter is called Same-Predicate. The parameter Same-Function is used for identical functions[5]. This part of the structural evaluation can proceed in parallel with match hypothesis construction.

At first glance systematicity might appear to be an inherently global concept, requiring difficult computations to enforce. We implemented it locally via *trickle-down*, a cascade-like model [12]. In trickle-down, a match hypothesis MH adds its score, scaled by the parameter Trickle-Down, to the match hypotheses linking the arguments of the items matched by MH. Thus in a deep system of relations the scores will cascade down, providing high scores for the object correspondences supporting the system (and thus for the system as a whole). This computation, too, can proceed in parallel, taking $\mathcal{O}(log(N))$ for a match hypothesis set of size N [6].

The structural evaluation system thus has three parameters: Same-Predicate, Same-Function, and Trickle-Down[6]. Once the propagation of local scores for all match hypotheses is complete, the SES of an interpretation (Gmap) is computed by summing of the scores of its constituent match hypotheses. In the original version of SME, scores were represented and combined using the Dempster-Shafer formalism [16,2]. We do not normalize Gmap scores, since doing so would only introduce further parameters without theoretical motivations. We also wish to avoid arbitrary assumptions concerning the scaling of human soundness judgments. Consequently, our conclusions will be based completely on ordinal comparisons between scores, never on the actual magnitude of scores themselves.

3 Modeling Soundness Judgments

To perform a sensitivity analysis one must have a standard for comparison. We use the cognitive simulation experiment described in [17], which showed that SME could replicate aspects of human soundness judgments demonstrated empirically [10,15]. In the psychological studies, subjects first read a large set of stories. In a subsequent session, they were shown similar stories and tried to retrieve corresponding original stories (an access measure). Afterwards, subjects were asked to judge the inferential soundness of pairs of stories. What was varied was the kind of similarity between pairs of stories; some cases shared only relational structure (i.e., were analogous), some only shared object similarities (i.e., appearance matches), and some shared both (i.e., were literally similar). Subjects rated literal similarity and analogy pairs as signficantly more sound than appearance matches.

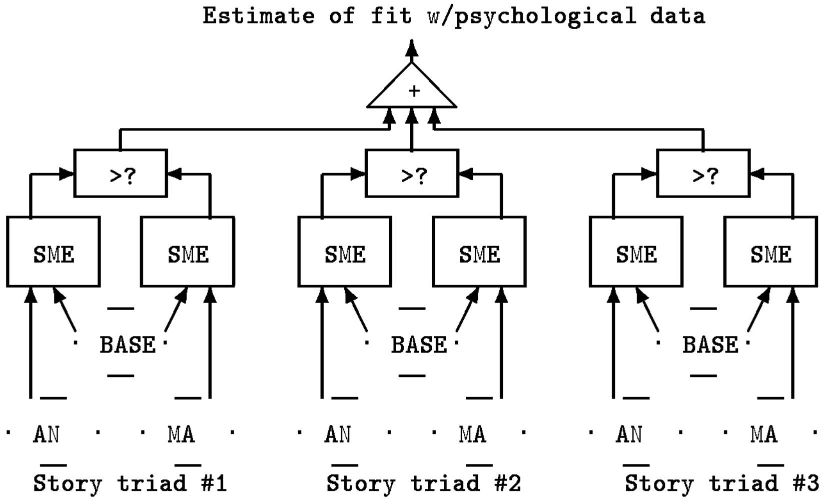
In the original simulation study, five triads of stories were encoded, each consisting of the base story (Base), an analogous story with different surface structure but similar relational structure (AN), and a story with surface similarities but different relational structure (MA). We asked whether SME's structural evaluation system could model these judgments. That is, if we interpret the SES as an indication of the soundness rating a subject would give, then to match the human data the score computed by SME for the Base/AN match should be higher than the score for the Base/MA match. As predicted, SES$(Base/AN) >$ SES$(Base/MA)$.

This experiment provides a useful framework for carrying out sensitivity analyses. Suppose we have N triads of stories. For any particular collection of parameters – numerical, symbolic, or algorithmic – we can define the fit with human performance to be the number of triads for which SES$(Base/AN) >$ SES$(Base/MA)$. By analyzing how the fit varies we can determine how sensitive the results are to each choice. Figure 1 depicts how this design can be viewed as an experimental apparatus. In the analyses which follow, three triads of stories were used each time. For each manipulation, this ESENSE apparatus was run over a sample of the numerical parameter space to estimate what fraction of the space provides results which fit the psychological data. The

[5] Functions are treated differently from predicates, since derived matches between non-identical functions are allowed if the structure above them matches (e.g., Temperature to Pressure).

[6] The original structural evaluation rules [6] had eight parameters; the reduction to three required some theoretical analysis and about two Symbolics-days of numerical sensitivity analyses. Although our initial use of eight parameters may seem large, it should be noted that simulations often have far more: ACME, for instance, relies on a numerical "similarity score" being available for each pair of predicates, hence the number of parameters is at least as large as the square of the number of predicates in the underlying representation language. Ascertaining the dependence of ACME's performance on its parameters via sensitivity analysis would appear to be a rather formidable task.

Figure 1: ESENSE Experimental setup for sensitivity analyses
The previous simulation experiment can be viewed as an apparatus which, for any particular combination of parameters, representations, and algorithm provides an estimate of fit to psychological data (here, an integer ranging between 0 and 3). Running this apparatus over alternate choices provides insight about how each aspect of the system accounts for the fit.



desirable outcome is that some portion, but not all, provides such a fit to the human data. If the whole space fits, then the parameters are irrelevant. If none of the space fits, then clearly that combination cannot account for human soundness judgments.

4 The Specificity Conjecture

Representational choices are often the most difficult issue in cognitive simulation. Rarely does a theory completely constrain the representational format, and while many choices are logically equivalent, even small changes can yield very different performance for a particular algorithm. Often there is no agreement (and sometimes intense disagreement) on what representations are reasonable. The typical solution is to test programs on a variety of examples to ensure generality. We believe that content variations alone are not always enough. Varying more global representational assumptions, such as the amount of perceptual information, can also be crucial.

The representations used in cognitive simulation tend to have much in common with those used in AI. They focus on the important aspects of what is to be represented, leaving out "irrelevant" information. Consequently they tend to be rather sparse. Such representations are fine if the only purpose is to compute a particular kind of answer (such as how to fix a broken car), and surely some human representations are like that. But is it reasonable to assume that most are? We suspect not. A person solving a problem or reading a story builds an internal representation from a variety of sources. This can include rich visual and auditory information about appearances (possibly including mental imagery) from which the relevant factors must be extracted. In fact, the more realistic the problem-solving scenario, the more irrelevant information there tends to be. While an expert may have an intricate theory of the situation, it is far from clear that the theory, as a percentage of the total number of propositions in the representation, dominates. And a novice faced with the same domain may have no applicable abstract knowledge, and thus can only encode observable properties.

Let us use *top-heavy* to refer to descriptions where most of the information is abstract, with very little information about appearances or basic object properties, and *bottom-heavy* for descriptions in which appearance information dominates (there may be just as much relational structure as top-heavy descriptions, as long as there is even more appearance information). Based on the observation that we can see far more than we can explain, we make the *Specificity Conjecture:* bottom-heavy descriptions are very common in human memory, perhaps outnumbering top-heavy descriptions. If this conjecture is correct, it is important to test simulations on bottom-heavy descriptions as well as the top-heavy descriptions which have been the favorite of experimenters.

How does the Specificity Conjecture affect structural evaluation of analogy? Structurally, top-heavy descriptions have a preponderance of higher-order relations, while bottom-heavy descriptions have many more attributes and first-order relations (e.g., LEFT-OF, BELOW). Consider the relative number of match hypotheses in the Base/AN and Base/MA comparisons described above. All else being equal, given a top-heavy representation the Base/AN comparison will have more match hypotheses than the Base/MA comparison, since there is more higher-order structure than appearance information. Conversely, in a bottom-heavy representation the Base/MA comparison will have more match hypotheses than the Base/AN comparison, since there is more appearance information to match than higher-order structure. Thus in top-heavy representations $\texttt{SES}(Base/AN) > \texttt{SES}(Base/MA)$ will tend to be true even with uniform MH scores, assuming that the higher-order structures do in fact match. But in bottom-heavy representations the tendency is towards $\texttt{SES}(Base/AN) < \texttt{SES}(Base/MA)$, due to the predominance of appearance information. In this case trickle-down plays a crucial role, to prevent the inferentially important comparison from being "swamped" by the surface comparison. People apparently have the ability to find structural commonalities even when they have bottom-heavy representations[7]. By looking for swamping over a space of numerical parameters and representation choices (i.e., by varying the amount of appearance information), we have a more subtle probe for exploring structural evaluation.

5 Analyzing normalization strategies

Any physically realizable computing scheme must include elements of finite dynamic range, and hence there will always be some normalization scheme which ensures that scores are within that range. The ability of trickle-down to prevent swamping depends in part on the normalization strategy used in computing scores. We can divide such strategies into two broad classes: *result normalization*, and *contribution normalization*. Connectionist models tend to use result normalization; a unit's inputs are multiplied by a set of coefficients, added, and then scaled by some non-linear function [13]. Formalisms for probabilistic reasoning tend to use contribution normalization; MYCIN's certainty factors, for instance, scale every contribution to belief in a proposition by the percentage of uncertainty remaining for that belief. Which kind of strategy, when plugged into the ESENSE apparatus, provides a better fit to the data?

To answer this question we set up the following experiment. First, we modified the encoded stories of the original simulation experiment to produce three sets of stories: one consisting of top-heavy descriptions, one consisting of bottom-heavy descriptions (i.e., twice as many match hypotheses for the Base/MA comparison as for the Base/AN comparison) and one "neutral" set, where the number of match hypotheses for the Base/MA and Base/AN comparisons were exactly equal. Then, we implemented a representative algorithm for each type of normalization. For the result normalization case we used the following rule:

$$\texttt{AddMax}: W_{i+1} = Min(1.0, W_i + C_i)$$

where W_i, W_{i+1} are the MH's score before and after the contribution, C_i is the amount contributed, and $W_0 = 0.0$. For the contribution normalization case we used the Dempster/Shafer code from the original SME structural evaluator. We then ran the ESENSE apparatus over every set of stories using each normalization strategy, varying the numerical parameters over a broad range, to see how these choices interacted to affect the fit with human performance.

[7] For example, in the experiment described above people gave higher structural evaluations to analogies than to appearance matches. Yet we can infer that they must have stored the stories with a great deal of low-order information, because their memory access was better for appearance matches than for analogical matches.

Table 1: Summary of fit as a function of representation and normalization
This table shows, for each combination of representation type and normalization algorithm, how much of the sampled parameter space can completely account for the data. That is, a value of $X\%$ indicates that given any parameter setting in that fraction of the space, SME's performance will exactly match the original human data. Dempster/Shafer cannot account for the data unless top-heavy representations are assumed.

	Top-heavy	Neutral	Bottom-heavy
Dempster/Shafer	4.1%	0.0%	0.0%
AddMax	75.8%	41.1%	18.1%

One complication in setting up the experiment is that these strategies differ in the ranges of parameters they allow. In Dempster/Shafer all parameters must be between zero and one. In AddMax allowing Same-Predicate or Same-Function to be one or greater is equivalent to just counting match hypotheses, so we restrict these to be less than one. Trickle-Down, on the other hand, can be greater than one, since the other parameters could be substantially less than one. (Even if the product is larger than one it makes no difference for AddMax, although it would violate the fundamental assumptions of Dempster/Shafer.) For AddMax we varied the three numerical parameters over the following ranges: Same-Predicate and Same-Function over (0.0, 10^{-4}, 10^{-3}, 0.01, 0.1, 0.3, 0.9) and Trickle-Down over (0.0, 0.5, 1.0, 2.0, 4.0, 8.0, 16.0). For Dempster/Shafer we varied all three parameters (Same-Predicate, Same-Function, and Trickle-Down) over the same set of values: (0.0, 10^{-4}, 10^{-3}, 0.01, 0.1, 0.3, 0.9). The number of samples for each algorithm is thus 7^3, or 343 points. To compute whether or not a point fits requires running each structural evaluator six times for each story set (i.e., to do the Base/AN and Base/MA comparison for each of three story triads in a set). Thus with three story sets 6,174 structural evaulations were required.

Table 1 summarizes the results by showing what percentage of the sampled parameter space yields a perfect fit of the data, as measured by the ESENSE apparatus. Dempster/Shafer clearly allows swamping as the number of attribute matches is increased. Thus it cannot explain the data, unless attention is restricted to top-heavy representations. AddMax, by contrast, can be tuned to fit the data for each type of representation. Can a single setting of parameters suffice? That is, are there subsets of the sample space in which AddMax fits the data for all three types, or are the subsets which fit the data for each type of representation disjoint? Yes, there is a single subset which fits the data. The boundary of this region appears complicated, and the coarseness of our sampling precludes a detailed description of it. However, it is reasonably large, indicating that the algorithm is not overly sensitive to particular choices of parameters. For example, within the ranges Same-Predicate $\in [10^{-3}, 0.01]$, Same-Function $\in [10^{-4}, 0.01]$, and Trickle-Down $\in [4, 16]$, every point fits perfectly. The regions which are clearly outside are interesting: Trickle-Down values of 1.0 or less, values of Same-Predicate of 0.3 or more, and values of Same-Function 0.9 or higher. Intuitively, what seems to be happening is this: Unless Trickle-Down is sufficiently high, not enough score cascades down to overcome the swamping effect of the large number of attribute matches. For the same reason, the "baseline activation" for each MH must be kept small; otherwise the cascade effect will be blocked by normalization.

This experiment suggests an interesting possibility. Since any physical computation scheme incorporates elements of limited dynamic range, for any set of parameters there will be some maximum depth beyond which additional systematicity cannot be distinguished, since the processing elements will have reached their maximum scores. This limit may be so high as to be irrelevant for human representations, or may show up as an "order cutoff" in failing to distinguish one comparison as more sound than another if both are extremely intricate.

6 Trickle-Down versus Order-Scoring

An interesting alternative to trickle-down for implementing systematicity is *order-scoring.* Consider a large relational structure which is shared by both base and target in some interpretation of an analogy. This structure will have a number of match hypotheses involving

Table 2: Results of Order Scoring on the story sets

This describes the percentage of the sampled points which perfectly fit the data for the three story sets described above. We repeat the trickle-down AddMax data for easy comparison. Order-scoring fails to account for human data, assuming the Specificity Conjecture holds.

	Top-heavy	Neutral	Bottom-heavy
Order-Scoring	69.1 %	47.2 %	0.0%
Trickle-Down	75.8%	41.1%	18.1%

relational items of high order. To satisfy structural consistency, the arguments of each such item must themselves have correspondences in the interpretation. Hence its mere presence in the interpretation indicates the existence of matches "all the way down" to object matches. Order scoring simply scales the score given to each match hypothesis by the order of the items involved, instead of passing scores downward as in trickle-down.

On computational grounds, we find order-scoring less preferable to trickle-down. First, to satisfy our constraints order must itself be computed locally. This is not difficult, if one allows information to propagate "upwards" from match hypotheses between entities (which have order zero) to match hypotheses which include them as arguments, and so on. However, this explicit computation of order seems inelegant, since, to paraphrase [12], it requires "more complex currency" than simply propagating local scores. A second difference is that order-scoring directly signals the existence of higher-order relations, and only indirectly signals the connectivity of a system of relational matches. Trickle-down, on the other hand, directly signals connectivity, leaving order implicit. Intuitively, connectivity seems a better structural reflection of coherence and inferential power than simply the existence of higher-order relations. Thus trickle-down has greater theoretical appeal as a way of deriving a structural evaluation.

But intuitions can be misleading. To see whether order-scoring could account for the data, we implemented a set of evaluation rules using this strategy. The contribution of order was defined by the function $\mathcal{OF}$:

$$\mathcal{OF}\ (MH) = Min(1.0, C \times [1 + Order(MH) \times \texttt{Order-Bias}])$$

where C is either Same-Predicate or Same-Function as appropriate. We sampled this parameter space in the same way as in the earlier analysis: i.e., Same-Predicate and Same-Function ranged over (0.0, 10^{-4}, 10^{-3}, 0.01, 0.1, 0.3, 0.9) and Order-Bias over (0.0, 0.5, 1.0, 2.0, 4.0, 8.0, 16.0). Table 2 summarizes the results. Clearly, $\mathcal{OF}$ is not a viable candidate for implementing the systematicity preference, since it is swamped on bottom-heavy representations. We suspect that this result will hold for all order-scoring algorithms. Even when Order-Bias is high, local normalization prevents the score of any particular MH becoming too high. Trickle-down avoids this limitation by co-opting all the lower-order structure matches under the high-order match, thus providing better resistance to swamping.

7 Discussion

Previous work demonstrated that structural critera are important in judging the relative soundness of analogical comparisons. This paper explores the relationship between the data and simulation in detail, making explicit the principles which constrain the space of algorithms we allow, and using the previous experiments to provide a framework for sensitivity analyses (the ESENSE apparatus) that help provide a deeper account of structural evaluation. In particular, we introduced the Specificity Conjecture, which suggests that in mental representations appearance and other low-order information is likely to dominate. If true, our experiments indicate that (a) normalization of scores for local matches should occur by a result normalization strategy rather than by contribution normalization and (b) trickle-down provides a better implementation of the systematicity preference than order-scoring.

We believe the Specificity Conjecture has important general ramifications for cognitive simulation. The aesthetic for good representations in AI is driven by the desire to solve particular kinds of problems. Since AI workers tend to do more explicit formal representations than workers in other areas of Cognitive Science, their aesthetic tends also to be inherited by other areas, even when it may not be appropriate. There is very little direct evidence about the format and statistical properties of mental representations (i.e., when they tend to be top-heavy versus bottom-heavy). Still, the fact that humans have powerful perceptual systems which deliver a rich assortment of information regardless of whether they know much else about what they are seeing argues for the importance of testing simulations with bottom-heavy representations. Showing that a simulation works in different content areas is now common. This is clearly important, but we now believe that it is not enough. One must explore how well a simulation performs with a range of representations that captures plausible intuitions about what the range of mental representations might be like.

8 Acknowledgements

Brian Falkenhainer provided valuable advice and algorithms for carrying out parts of the sensitivity analyses. This paper benefited from discussions with John Collins, Brian Falkenhainer, Rob Goldstone, Art Markman, Doug Medin, Janice Skorstad, and Ed Smith. This research was supported by the Office of Naval Research, Contract No. N00014-85-K-0559, an NSF Presidential Young Investigator award, and an equipment grant from IBM.

References

[1] Burstein, M. H. (1983). A model of learning by analogical reasoning and debugging. In *Proceedings of the National Conference on Artificial Intelligence*, Washington, D. C.

[2] Falkenhainer, B., Towards a general-purpose belief maintenance system, in: J.F. Lemmer (Ed.), *Uncertainty in Artificial Intelligence, Volume II*, 1987. Also Technical Report, UIUCDCS-R-87-1717, Department of Computer Science, University of Illinois, 1987.

[3] third stage in the analogy process: Verification-Based Analogical Learning, Technical Report UIUCDCS-R-86-1302, Department of Computer Science, University of Illinois, October, 1986. A summary appears in *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, Milan, Italy, August, 1987.

[4] Falkenhainer, B., The SME user's manual, Technical Report UIUCDCS-R-88-1421, Department of Computer Science, University of Illinois, 1988.

[5] Falkenhainer, B., K.D. Forbus, D. Gentner, The Structure-Mapping Engine, *Proceedings of the Fifth National Conference on Artificial Intelligence*, August, 1986.

[6] Falkenhainer, B., Forbus, K., Gentner, D. The Structure-Mapping Engine: Algorithm and examples *Artificial Intelligence*, to appear, 1989.

[7] Gentner, D., The structure of analogical models in science, BBN Tech. Report No. 4451, Cambridge, MA., Bolt Beranek and Newman Inc., 1980.

[8] Gentner, D., Structure-mapping: A theoretical framework for analogy, *Cognitive Science* **7**(2), 1983.

[9] Gentner, D., Mechanisms of analogical learning. To appear in S. Vosniadou and A. Ortony, (Eds.), *Similarity and analogical reasoning.* Presented in June, 1986.

[10] Gentner, D., & R. Landers, Analogical reminding: A good match is hard to find. In *Proceedings of the International Conference on Systems, Man and Cybernetics.* Tucson, Arizona, 1985.

[11] Holyoak, K. & Thagard, P. Analogical mapping by constraint satisfaction, to appear in *Cognitive Science.*

[12] McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review,88*(5), 375-407.

[13] Rumelhart, D. and McClelland, J. *Parallel Distributed Processing, Volumes 1 & 2*, The MIT Press, 1986.

[14] Marr, D. *Vision*, W. H. Freeman and Company, San Francisco, 1982.

[15] Rattermann, M.J., and Gentner, D. Analogy and Similarity: Determinants of accessibility and inferential soundness, *Proceedings of the Cognitive Science Society*, July, 1987.

[16] Shafer, G., *A mathematical theory of evidence*, Princeton University Press, Princeton, New Jersey, 1976.

[17] Skorstad, J., Falkenhainer, B., Gentner, D., Analogical Processing: A simulation and empirical corroboration, in: *Proceedings of the Sixth National Conference on Artificial Intelligence*, Seattle, WA, August, 1987.