A QUANTITATIVE ANALYSIS OF ANALOGY BY SIMILARITY

Stuart J. Russell
Department of Computer Science
Stanford University
Stanford, CA 94305

ABSTRACT

In the absence of specific relevance information, the traditional assumption in the study of analogy has been that the most similar analogue is most likely to provide the correct solutions; a justification for this assumption has been lacking, as has any relation between the similarity measure used and the probability of correctness of the analogy. We show how a statistical analysis can be performed to give the probability that a given source will provide a successful analogy, using only the assumption that there are some relevant features somewhere in the source and target descriptions. The predicted variation of the probability with source-target similarity corresponds closely to empirical analogy data obtained by Shepard for human and animal subjects for a wide variety of domains. The utility of analogy by similarity seems to rest on some very fundamental assumptions about the nature of our representations.*

I INTRODUCTION

Analogical reasoning is usually defined as the argument from known similarities between two things to the existence of further similarities. Formally, we can define it as any inference following the schema

$$P(S,W),\ P(T,W),\ Q(S,Y)\ \xrightarrow{\text{anal}}\ Q(T,Y)$$

(see [Russell 86b]) where T is the *target*, about which we wish to know some fact Q (the *query*); S is the *source*, the analogue from which we will obtain the information to satisfy Q by analogy; P represents the known similarities given by the shared attribute values W. P and Q can be arbitrary predicate calculus formulae.

An innumerable number of inferences have this form but are plainly silly; for example, both today and yesterday occurred in this week (the known similarity), yet we do not infer the further similarity that today, like yesterday, is a Friday. The traditional approach to deciding if an analogy is reasonable, apparently starting in [Mill 73], has been to say that each similarity observed contributes some extra evidence to the conclusion; this leads naturally to the assumption that the most suitable source analogue is the one which has the greatest similarity to the target. Thus similarity becomes a measure on the *descriptions* of the source and target. However we define the similarity measure, it is trivially easy to produce counterexamples to this assumption. Moreover, Tversky's studies [Tversky 73] show that similarity does not seem to be the simple, two-argument function this naïve theory assumes. One can convince oneself of this by trying to decide which day is most similar to today.

The theory of *determinations* ([Davies & Russell 86], [Russell 86b]) gives a first-order definition to the notion of the *relevance* of one fact to another. given that the known similarities are (partially) relevant to the inferred similarities, the analogical inference is guaranteed to be (partially) justified. The fact that P is relevant to Q is encoded as a determination, written as $P(\underline{x},\underline{w}) \succ Q(\underline{x},\underline{y})$ and defined as

$$P(\underline{x},\underline{w}) \wedge P(\underline{z},\underline{w}) \wedge Q(\underline{x},\underline{y}) \Rightarrow Q(\underline{z},\underline{y}).$$

With this information, the overall similarity becomes irrelevant.

When the similarity is insufficient to determine the query at hand, i.e., we have no idea which of the known facts might be relevant, the theory does not apply. However, it still seems plausible that the most similar source is the best analogue. What has been lacking in previous theories of analogy by similarity is any attempt to justify this assumption; the analysis in this paper hopes to rectify this situation. Since an inference by analogy is still an inference, the justification must take the form of an argument as to why a conclusion from similarity is any better than a random guess; better still, the theory should be able to assign a probability to the conclusion given the truth of the premises. The object of this paper is thus to compute (or at least sketch) the relationship between the measure of similarity between two objects, and the probability that they share a further, specified similarity.

The principal problems which need to be solved before such a theory can be constructed are:

1) A reasonable way must be found to circumscribe the source and target descriptions. Without this, the sets of facts to be compared are essentially without limit.
2) A similarity measure must be defined in such a way as to be (as far as possible) independent of the way in which the source and target are represented.
3) We must identify the assumptions needed to relate the similarity measure to the desired probability.

The precise similarity measure itself is not important; in fact, it is essentially meaningless. If we have a different similarity measure, we simply need to relate it in a different way to the probability of correctness of the analogy. Thus we will *not* be attempting to define a similarity measure that is more plausible than those proposed previously.

The essence of our approach is to show that analogy to a maximally similar source can be justified in the absence of any applicable determination by showing that such a source is the most likely to match the target on the properties which are relevant to the query (*even though the identity of these*

* This work was performed while the author was supported by a NATO studentship from the UK Science and Engineering Research Council, and by ONR contract N00014-81-K-0004. Computing support was provided by the SUMEX-AIM facility, under NIH grant RR-00785.

properties is unknown). If a source matches the target on all relevant features, an analogy from that source is assumed to be correct.

We first calculate the probability of such a match for the simple case of an attribute-value representation in which the relevance of any attribute is equally likely *a priori*; initially this is done assuming a fixed number of relevant features, and then we incorporatè the assumption of a probability distribution for the number of relevant features. The result of the analysis is a quantitative prediction of the probability of correctness of an analogy to a given source as a function of the overall similarity of that source to the target. The prediction bears a very close resemblance to the empirical 'stimulus generalization probability' (the psychological term for the probability we are trying to calculate) measured in animal and human experiments.

In a subsequent section we attempt to relax the simple representational assumptions to allow the theory to apply to the general case. We conclude with a discussion of the difficulties inherent in such a task, and an indication of how similarity can be combined with determination-based reasoning to create a more general theory of analogy.

II THE SIMPLE MODEL

A simplified model for analogy in a database is this: we have a target T described by m attribute-value pairs, for which we wish to find the value of another attribute Q. We have a number of sources $S_1 \dots S_n$ (analogues) which have values for the desired attribute Q as well as for the m attributes known for the target.

Define the similarity s as the number of matching attribute values for a given target and source. The difference $d = m - s$.

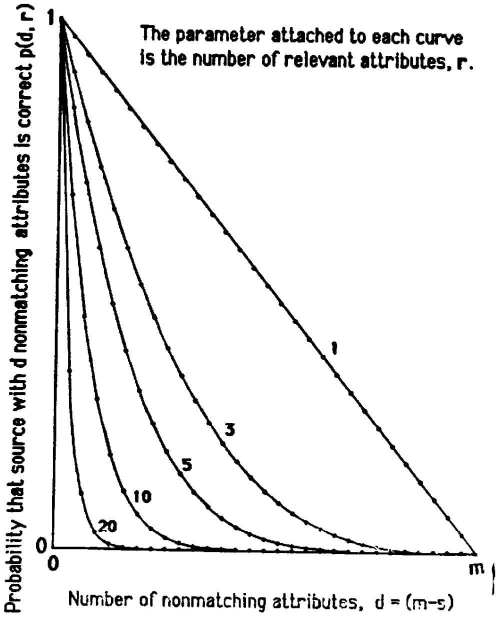


Fig. 1 $p(d,r)$ for $r = 1,3,5,10,20$

Assume that there are r attributes relevant to ascertaining the value of Q, and that *the relevant attributes are all included somewhere in the target descriptions.* This is equivalent to saying that the conjunction of all the attributes in the description is sufficient to determine the query (but since not all the attribute values match we cannot use this to conclude the desired similarity with certainty). Thus the solution to the problem of circumscribing the source and target descriptions is to limit them to the attributes contained in the left-hand side of the least specific determination available for the query at hand.

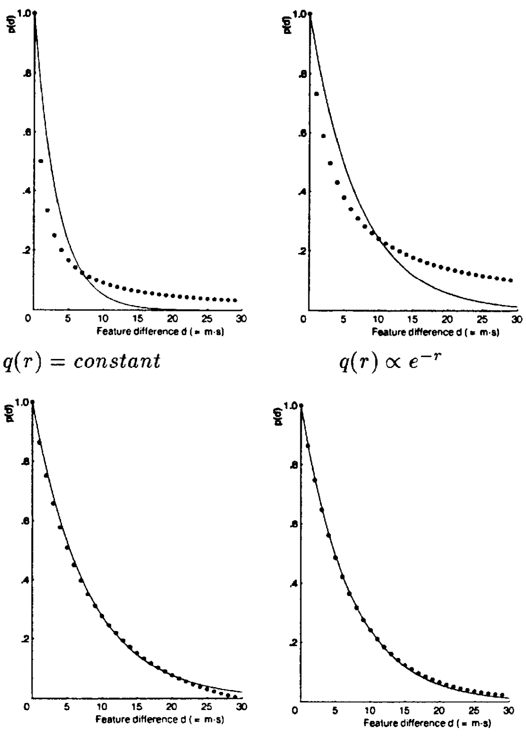
Define $p(d,r)$ to be the probability that a source S, differing from the target on d attributes, matches it on the r relevant attributes. In the first instance, we assume that *all attributes are equally likely to be relevant.* We can thus calculate $p(d,r)$ using a simple combinatoric argument: the number of choices of which attributes are relevant such that S matches T on those attributes is $(m-d)$ choose r; the total number of choices of which attributes are relevant is m choose r; the value of $p(d,r)$ is the ratio of these two numbers:

$$p(d,r) = \binom{m-d}{r} \bigg/ \binom{m}{r} \quad (r \geq 1)$$

For any r, this function drops off with d (= m-s), monotonically and concavely, from 1 (where d=0) to 0 (where d > m-r). Thus the most similar analogue is guaranteed to be the most suitable for analogy. Figure 1 shows $p(d,r)$ for values of r of 1, 3, 5, 10, 20 with the total number of attributes $m = 30$. As we would expect, the curve narrows as r increases, meaning that a higher number of relevant attributes necessitates a closer overall match to ensure that the relevant similarities are indeed present.

III ALLOWING r TO VARY

The assumption of a fixed value for the number of relevant features seems rather unrealistic. The most general assumption we can make is that r follows a probability distribution



$q(r) = constant$

$q(r) \propto e^{-r}$

$q(r) \propto re^{-r}$

$q(r) = N(4,2)$

Figure 2 $p(d)$ given various assumptions about $q(r)$

$q_Q(r)$ which depends on the type of the query Q. Thus, for example we could assume that there are equally likely to be any number of relevant features, or that three or four seems reasonable whilst 25 is unlikely. Although this introduces an extra degree of freedom into the theory, we find that the results are almost independent of what we assume about q. We calculate the probability of successful analogy now as a function of the source-target difference d only:

$$p(d) = \sum_{r=0}^{m} q(r)p(d,r)$$

using the above formula for $p(d,r)$. For any reasonable assumption about the shape of q(r), the variation of p(d) with d remains approximately the same shape.

For $q(r) = constant$, $p(d) \sim 1/(d+1)$

For $q(r) \propto e^{-r}$, $p(d) \sim e^{-d}$ for low d

For $q(r) \propto re^{-r}$, $p(d) \sim e^{-d}$ except at large d

For $q(r) = Normal(\mu = 4, \sigma = 2)$, $p(d) \sim e^{-d}$

The first two assumptions are somewhat unrealistic in that they assign significant probability to there being no relevant features. When this possibility is discounted, the curves come much closer to being exponential. In figure 2 we show values of p(d) (plotted as dots) computed using these four assumptions of $q(r)$, with a simple exponential decay ($p(d) \propto e^{-d}$, solid line) superimposed.

IV EMPIRICAL DATA ON STIMULUS GENERALIZATION

Psychological experiments on *stimulus generalization* are highly relevant to the study of analogy by similarity. In these experiments, a (human or animal) subject is given an initial stimulus, to which it makes a response. If necessary, the correct response is confirmed by reinforcement. This original stimulus-response pair is the *source* in our terms. Then a second stimulus is given, which differs from the original. This represents the *target* situation, for which the subject must decide if the original response is still appropriate. The empirical probability that the subject makes the same response (*generalizes* from the original stimulus) is measured as a function of the difference between the stimuli. This probability is essentially what we are predicting from rational grounds in the above analysis.

Early results in the field failed to reveal any regularity in the results obtained. One of Shepard's crucial contributions ([Shepard 58]) was to realize that the similarity (or difference) between the stimuli should be measured not in a *physical* space (such as wavelength of light or pitch of sound) but in the subject's own *psychological* space, which can be elicited using the techniques of multi-dimensional scaling ([Shepard 62]). Using these techniques, Shepard obtained an approximately exponential stimulus generalization gradient for a wide variety of stimuli using both human and animal subjects. Typical results, reproduced, with kind permission, from Shepard's APA presidential address ([Shepard 81]), are shown in figure 3. His own recent theory to explain these results appears in [Shepard 84], and has a somewhat similar flavour to that given here.

V GENERALIZING THE MODEL

In principle, we can make the simple model analyzed above applicable to any analogical task simply by allowing the 'attributes' and 'values' to be arbitrary predicate calculus formulae and terms. The assumption that each of these new 'attributes' is equally likely to be relevant is no longer tenable, however. In this section we will discuss some ways in which the similarity measure might be modified in order to allow this assumption to be relaxed. The idea is to reduce each attribute to a collection of uniform mini-attributes; if the original assumptions hold for the mini-attributes, our problem will be solved. Unfortunately, the task is non-trivial.

The first difficulty is that we can only assume equal relevance likelihood if the *a priori* probabilities of a match on each attribute value are equal; in general, this will not be the case. In the terms of [Carnap 71], the *widths* of the regions of possibility space represented by each attribute are no longer equal. Accordingly, the simple notion of similarity as the number of matching attributes needs to be revised. If the cardinality of the range of possible values for the i^{th} attribute is k_i, then the probability p_i of a match (assuming uniform distribution) is $1/k_i$. Although k will vary, we can overcome this by reducing each attribute to $\log_2 k$ mini-attributes, for which the probability of a match will be uniformly 0.5. If the original distribution is not uniform (for example, a match on the NoOfLegs attribute with value 2 is much more likely than a match with value 1), a similar argument gives the appropriate contribution as $-\log_2 p_i$ mini-attributes. This refinement may underlie the intuition that 'unusual' features are important in metaphorical transfer and analogical matching ([Winston 78], [Ortony 79]).

In [Russell 86b], the notion of one value 'almost matching' another is taken into account by supposing that determinations are expressed using the 'broadest' attributes possible, so that precise attributes are grouped into equivalence classes appropriate to the task for which we are using the similarity.In other words, similarities are re-expressed as commonalities. In the current situation, however, we will not know what the appropriate equivalence classes are, yet we still want to take into account inexact matches on attribute values; for example, in heart disease prognosis a previous case of a 310-lb man would be a highly pertinent analogue for a new case of a 312-lb man. If the weight attribute was given accurate to 4 lbs, these men would weigh the same; thus in general an inexact match on a scalar attribute corresponds to an exact match on less fine-grained scale, and the significance of the 'match' is reduced according to the log of the accuracy reduction (2 bits in this case).

A consequence of this view of the significance of an attribute leads to a constraint on the possible forms of $q(r)$: if we assume that the relevant attributes must contain at least as much information as the attribute Q whose value they combine to predict, then we must have $q(r) = 0$ if r is less than the significance value of Q. Here r, as well as the total 'attribute count' m and the similarity s, are all measured on a scale where a one-bit attribute has a significance of 1. At first sight, it seems that we have succeeded in breaking down our complex features into uniform elements, all of which are equally likely to be relevant, so all the earlier results should still apply.

However plausible this may seem, it is simply false. The base of the logarithms chosen is of course totally arbitrary — we would still have uniform mini-attributes if we used $\log_4$. This would mean halving our values for m, r and s; but the formula for $p(d,r)$ contains combinatoric functions, so it will not scale. Hence our predicted probability will depend on the base we choose for the logarithms! This is clearly unsatisfactory. What we have done is to neglect an important assumption made in using the combinatorial argument, namely that the relevant information consisted of a set of *whole features.* If we allow it to consist of a collection of sub-elements of various features, then clearly there are many more ways in which we can choose this set. The plausibility of the simple model rests in our unstated assumption that the attributes we use carve up the world in such a way as to correctly segment the various causal aspects of a situation. For example, we could represent the fact that I own a clapped-out van by saying

$$OwnsCar(SJR, 73DodgeSportsmanVanB318)$$

using one feature with a richly-structured set of values; but for most purposes a reasonable breakdown would be that I own a van (for other people's moving situations), that it's very old (for long-distance trip situations), that it can seat lots of people (for party situations), that it's a Dodge (for frequent repair situations) and that it's virtually worthless (for selling situations). Few situations would require further breakdown into still less specific features. In some sense, therefore, we will require a theory of natural kinds for features as well as for objects.

If it is the case that humans have succeeded in developing such well-tuned representations, then it is indeed reasonable for us to assume that the relevant information, which corresponds to the part of the real-world situation which is responsible for

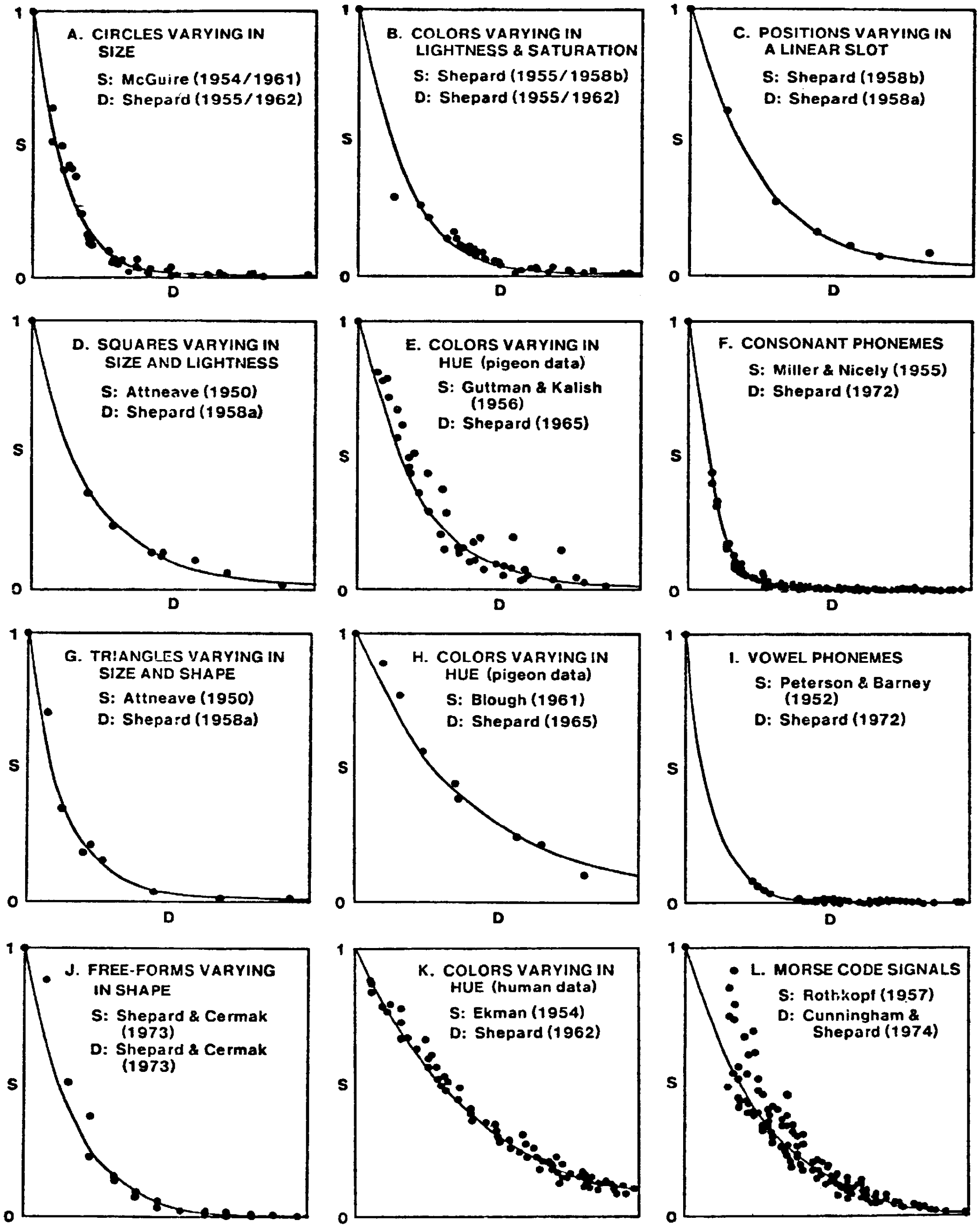


Fig. 3 Plots of analogical response probability (S) against source-target difference (D) from [Shepard 81].

determining the queried aspect, will consist of a set of discrete features corresponding to the various possible causal factors present. This of course raises a vast throng of questions, not least of which is that of how an AI system is to ensure that its representation has the appropriate properties, or even it can know that it does or doesn't. The subject of the semantic implications of using a particular representation is also touched upon in [Russell 86a], where we tie it in to the process of vocabulary acquisition; a real understanding is still far beyond our reach, but an appreciation of the problem, and the areas on which it impinges, is a first step.

VI CONCLUSIONS

The first steps toward a quantitative analysis of the probability of correctness of an analogy as a function of the source-target difference have been presented, giving the first justification for the maximal similarity heuristic. Although several difficult problems remain, it may be possible to define a representation-independent similarity measure on reliably circumscribed object descriptions. The empirical verification of the theory by Shepard's results is extremely good, in the sense that it shows that humans and animals possess a rational ability to judge similarity which has evolved, presumably, because of the optimal performance of its predictions given the available information. Shepard's explanation of the results and our own are somewhat complementary in that he deals with unanalyzed stimuli whereas our model assumes a breakdown into features. Given the usual nature of AI representations, this is well-suited for our purpose of constructing a computational theory of analogy and a generally useful analogy system for AI. We intend to further explore the implications and loose ends of the theory by performing large numbers of analogies in an AI database of general knowledge (Lenat's CYC system; see [Lenat et al 86]). A further goal is to integrate analogy by similarity with the determination-based analogical reasoning theory. We anticipate three forms of integration:

1) overconstrained determinations will circumscribe broad classes of potentially relevant features; we reason by similarity within these constraints if no exact match can be found;
2) probabilistic determinations can add weights to the contributions of individual attributes to the overall similarity total;
3) observation of an unexpectedly high similarity can initiate a search for a hitherto unknown regularity to be encoded as a new determination.

When intelligent systems embodying full theories of limited rationality are built, an ability to perform analogical reasoning using both determinations and similarity will be essential in order to allow the system to use its experience profitably. Analogy by similarity also seems extremely well suited to the task of producing reliably fast, plausible answers to problems, particularly in a parallel environment. It is hoped that the ideas in this paper have gone some way towards realizing this possibility, although it is clear that more questions have been raised, some of them for the first time, than have been answered.

ACKNOWLEDGEMENTS

I would like to thank my advisors Doug Lenat and Mike Genesereth for fruitful discussions of these ideas, Benjamin Grosof and Devika Subramanian for helpful comments on an earlier draft, and Roger Shepard for making these ideas possible.

References

[Carnap 71]
Carnap, Rudolf. A Basic System of Inductive Logic, Part I. In R. Carnap and R. C. Jeffrey (Eds.) *Studies in Inductive Logic and Probability* Vol I. Berkeley, CA: University of California Press; 1971.

[Davies & Russell 86]
Davies, Todd & Stuart Russell. *A Logical Approach to Reasoning by Analogy.* Stanford CS Report (forthcoming) and Technical Note 385, AI Center, SRI International; June, 1986.

[Lenat et al 86]
Lenat D., Mayank P. and Shepherd M.. CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks. *AI Magazine* Vol. 6 No. 4; Winter 1986.

[Mill 73]
Mill, J. S. *System of Logic* Book III Ch XX 'Of Analogy' in Vol. VIII of *Collected Works of John Stuart Mill.* University of Toronto Press; 1973.

[Ortony 79]
Ortony A.. Rôle of Similarity in Similes and Metaphors in Ortony A. (ed.) *Metaphor and Thought.* Cambridge University Press; 1979.

[Russell 86a]
Russell, Stuart J. "Preliminary Steps toward the Automation of Induction". In *Proceedings of the National Conference on Artificial Intelligence.* Philadelphia: AAAI; 1986.

[Russell 86b]
Russell, Stuart J. *Analogical and Inductive Reasoning.* Ph. D. thesis. Stanford University; 1986.

[Shepard 58]
Shepard R. N. Stimulus and Response Generalization: Deduction of the Generalization Gradient from a Trace Model. *Psychological Review* Vol. 65; 1958.

[Shepard 62a,b]
Shepard R. N. The analysis of proximities: multidimensional scaling with an unknown distance function (Parts I and II). *Psychometrika* Vol. 27; 1962.

[Shepard 81]
Shepard, Roger. APA Division 3 Presidential Address, Los Angeles, August 25, 1981.

[Shepard 84]
Shepard R. N. Similarity and a law of universal generalization. Paper presented at the annual meeting of the Psychonomic Society, San Antonio, TX; November, 1984.

[Tversky 77]
Tversky, Amos. Features of Similarity. *Psychological Review* Vol. 84, No. 4; 1977.

[Winston 78]
Winston, Patrick H. Learning by Creating and Justifying Transfer Frames. *Artificial Intelligence* Vol. 10, No. 4; April, 1978.