Running Head: segregation effect

The Segregation Effect, Time Pressure and Other Factors Affecting Argument Strength in Categorical Inductive Inference

David Leiser

Limor Biton-Bereby

Ben-Gurion University of the Negev

The Segregation Effect, Time Pressure and Other Factors Affecting Argument Strength in Categorical Inductive Inference

Abstract

This work investigates category-based induction in adults. Two main experiments showed that it is possible to predict with very high accuracy the strength of a complex argument (two premises), from the strength of its component arguments and the similarity between the premises. The stronger of the simple arguments has a larger effect than the other, but both do contribute to the judged overall strength of the argument. We also demonstrate a novel effect, segregation: adding premises can decrease the strength of the complex argument, when the two premise categories are very similar to one another. Time pressure does not affect the result pattern. No signs of reliance on pre-existing hierarchical categorical structure were found. These results are confronted with contrasting predictions derived from four existing accounts of inductive inference.

The Segregation Effect, Time Pressure and Other Factors Affecting Argument Strength in Categorical Inductive Inference

Over the last 25 years, there have been many studies of inductive reasoning, and especially of how people respond to various kinds of categorical inductive arguments. These are probabilistic arguments, and subjects are typically requested to evaluate the likelihood of the argument's conclusion if certain premises are assumed. For example, subjects could be asked, "Assume that Hyenas and Cows both have the property P. What is the probability that Wolves have that same property?"

Categorical arguments may be separated in two classes, in keeping with the number of categories involved. (1) *Simple arguments*, of the form (F → T), are based on two categories only, as in questions of the type: Assume that category F has property P, how likely is it that category T has property P too? (2) *Complex arguments*, of the form F_1, F_2, …, F_n → T), involve three (or more) categories. Questions may be of the form: Assume that category F_1, F_2, …, F_n have property P, how likely is it that category T also has property P?. We will be concerned with the relation between these two types of arguments, and investigate how the relations of generalizability between pairs of categories predict the strength of complex arguments.

In an extensive review, Heit (2000) lists the main "touchstone results" on these two classes of categorical arguments. The main results relating to simple arguments are that similarity between premise and conclusion categories promotes induction, that typicality of the premise category promotes induction. and homogeneity of the conclusion category promotes induction. Concerning complex arguments, the question is: What makes a set of cases generalizable? A set of the most induction-promoting cases does not necessarily lead to the strongest possible ensemble of cases. Two important results are: (1) Greater *diversity* of observations, or premises, promotes induction. This effect was called *premise diversity* by Osherson, Smith, Wilkie, Lopez and Shafir, (1990). Thus, the

argument (Chimpanzee Fox→ Bear)[1] will be judged stronger than the argument (Sheep Fox→ Bear). (2) Greater *number* of premises promotes induction. This is called *premise monotonicity* by Osherson et al. (1990), and the "premise diversity effect" by others. For example, the argument (Pig Wolf Fox→ Gorilla) will be judged stronger than the argument (Pig Wolf→ Gorilla). However, certain circumstances result in a violation of this rule, an exception called *nonmonotonicity* by Osherson et al (1990). Specifically, when an "improper premise" is added, the argument will be judged as weaker. An "improper premise" is a member of a category different from that of the premise or the conclusion. Thus, the argument (Mosquito→ Bee) will be judged as stronger than the argument (Mosquito Gorilla→ Bee), because Gorilla is an improper premise in this context.

Models of inductive arguments evaluation

Several models have been offered to account for adult induction behavior. We next review the main models and derive contrasting predictions made by the different models, in an attempt to decide between them. A set of experiments will then presented, in an effort to obtain the relevant empirical data

1. Osherson et al: a two-component model

The most extensive and influential work on induction from multiple categories was conducted by Osherson et al. (1990). Osherson et al. (1990) identified a range of characteristic effects in adult subjects' probability judgments of categorical arguments and formulated the "Similarity Coverage Model" (SCM), which purports to account for these effects. Their model is based on two components that reflect similarity and category membership respectively. The components are (a) the extent of similarity between the premises' categories and the conclusion, .and (b) the extent to which the premises' categories cover their *superordinate category*, defined as the lowest level category that includes all the categories in the premises and the conclusion. For example, in the argument (Robin Duck → Sparrow) subjects may conclude that Sparrows have the relevant

predicate since (a) it resembles Robins (induction based on similarity) and (b) since they infer that Birds in general have that predicate (induction based on category membership). According to SCM, *the similarity component* is computed by comparing the degree of similarity between each of the premise categories and the conclusion category, and selecting the largest value. The larger that value, the stronger the argument. In the previous example, assuming that Robins are more similar to Sparrows than are Ducks, the similarity component of (Robin, Duck → Sparrow) = *similarity* (Robin, Sparrow).

To define the *coverage component*, Osherson et al. (1990) introduce the concept of the *superordinate category*. In the previous argument (Robin Duck→ Sparrow), the superordinate category is the Birds category, since it comprises both the categories in the premises (Robin and Duck) and the conclusion category (Sparrow). The superordinate category is the *lowest* inclusive category in the hierarchical system. For instance, in the argument (Tiger, Jaguar → Cat) the superordinate category would be "Felines" whilst in the argument (Tiger, Jaguar→ Giraffe) the superordinate category is "mammals". Individual categories "cover" a higher order category to the extent that, between them, they resemble the other members of that category. Coverage increases to the extent that the premises' categories resemble different elements of the superordinate category (visually, they cover different surfaces in the area of the superordinate category). The larger the coverage of the superordinate category by the premises' categories, the stronger an argument will be judged. It is this component that accounts for the diversity effect mentioned earlier. Premises that are more diverse cover their superordinate category better, thus strengthening the argument in which they serve. The similarity and coverage components of the model are combined additively, with different weights expressing the relative importance given to the two components, though the type of data collected in the past does not allow computing those weights.

2. HEIT: a Bayesian analysis

Heit (1998) offers an alternative model of category-based induction, in the rational analysis perspective. He focuses on the predicate, P and considers every new premise of the form "predicate P is true of category C" as an occasion for belief revision about P. Evaluating an inductive argument is conceived of as estimating the range of P, and this warrants the application of a Bayesian model. Further, he supposes that people assume that novel properties will be distributed like already-known properties. With these and similar assumptions, Bayes's formula readily explains the similarity effect. This is most easily seen with a concrete example. People will be more willing to project a novel property from Cows to Sheep than from Cows to Ferrets whenever people already know of more common properties for Cows and Sheep than for Cows and Ferrets. Heit can also account for diversity effects, by noting that if two very different categories share a property, it is unlikely that they will be the only ones to have this property. Thus, "the properties shared by Cows and Ferrets, such as having fur and being warm-blooded, tend to be true of other mammals as well. Hence a novel property of Cows and Ferrets is likely to be distributed the same way, and thus true of other mammals."

3. Sloman: a processing model

Another model that is broadly similar is that by Sloman (1993). However, whereas Heit's model is a "computational-level" account, that is, a description of the task that is performed in evaluating inductive arguments, Sloman's (1993) model is a processing model. In his model, every category is characterized by a vector of features, and each feature possessed by a given category is represented by a position in its associated features' vector. When a predicate is asserted of a category, the features of the category are associated with the predicate. The model's ability to perform inductive reasoning is based on explicit updating of information on the predicate involved. Every encounter with a category said to possess the predicate updates what is known about the

predicate, and specifically the strength with which various features are associated with possessing the predicate.

To implement this, Sloman introduces a feature vector for the predicate itself too. That vector is updated whenever the predicated is asserted of an additional category, by a straightforward connectionist mechanism. To judge the strength of an argument, premises are encoded by associating the features of the premises to the predicate and then testing the degree of activation of the predicate upon presentation of the conclusion category features vector. Argument strength can be roughly defined as the proportion of features in the conclusion category that are also in the premise categories. The argument's strength increases with the extent of shared features between premise and conclusion categories and decreases with the extent of features that appear exclusively in the conclusion category.

In his model, features of the premises and those of the conclusion categories are not treated symmetrically: features of the premises categories that are not found in the conclusion are simply ignored by the model. The rationale for this is that an argument is strong to the extent that its conclusion is "supported" by its premises relative to the total amount that must be explained or justified. By the same logic, a premise only increases belief in a conclusion to the extent that the features its category shares with the conclusion category are not redundant, i.e. they are not also shared by other premise categories.

4. Tenenbaum and Griffiths: A more complex Bayesian model

A more elaborate model of inference or generalization in the Bayesian tradition was recently proposed by Tenenbaum and Griffiths (2001). Their proposal may be seen as an extension of the influential "universal law of generalization" introduced by Shepard (1987). Shepard observed that across a wide variety of experimental situations, including both human and animal subjects, generalization gradients tend to taper off approximately exponentially with distance in an

appropriately scaled psychological space (as obtained by multidimensional scaling). He offered a rational probabilistic argument for the origin of this universal law, starting with some basic assumptions about natural kinds, and derived an exponential decay function as the form of the universal generalization gradient. That region around the focus of the generalization is called the *consequential region*.

This original formulation applied only to the ideal case of generalization from a single encountered stimulus to a single novel stimulus, and for stimuli that can be represented as points in a continuous metric psychological space. Tenenbaum and Griffiths (2001) recast Shepard's theory in a Bayesian framework which enabled them to extend Shepard's approach to the case of generalizing from multiple premises or stimuli. Their main advance over Shepard's original analysis comes in introducing the *size principle* for judging hypotheses about the consequential region based on their size, or specificity. In their conceptualization, the subject has to decide, on the basis of information about one or several stimuli, what the consequential region is, i.e., to what other stimuli the novel information can generalize. This analysis leads Tenenbaum (1999) to the size principle, according to which smaller consequential regions are judged as more probable than larger ones, even when both are equally consistent with the observed consequential stimulus. The demonstration of the size principle assumes that the stimuli presented are sampled randomly from the true concept. This is called "strong sampling", as opposed to "weak sampling" that assumes that examples are generated by an arbitrary process independent of the true concept. Weak sampling is a natural model in some cases (e.g. when the learner himself or a random process chooses objects, and these are labeled later as "positive" or "negative" by a teacher). However, when the data is presented as a source of information, the assumption of strong sampling seems appropriate – as the subject is entitled to expect that the information presented is relevant, in keeping with Grice's (1989) "cooperative principle" (whose first pragmatic maxim is the maxim of relation: Be relevant).

Contrasting predictions

Insert Figure 1 about here

We will now derive predictions made by each model, and first present our notational convention. Consider Figure 1. Arguments go **F**rom premises **T**o conclusion. Accordingly, let us note premise categories by F_1, F_2 etc., and the conclusion category by T. We can then note as $F_i \rightarrow T$ the simple argument with F_i as premise and T as conclusion, and as $F_1F_2 \rightarrow T$ the complex argument with $F_1$and F_2 as premises. Lastly, we will use the convention that F_1 will be the premise category that is closest to T, so that argument $F_1 \rightarrow T$ is always, by definition, at least as strong as $F_2 \rightarrow T$. The first set of predictions concern the effect on $F_1F_2 \rightarrow T$ of the similarity pattern between the three categories F_1 F_2 and T.

The contribution of the second premise

Of course, all four models predict the similarity effect for simple arguments: similarity between premise and conclusion categories promotes induction. The more similar F and T, the stronger the inference $F \rightarrow T$. However, they differ when it comes to complex arguments, and especially regarding the effect of $F_2 \rightarrow T$. It will be remembered that SCM specifically excludes $F_2 \rightarrow T$ from the similarity component, since it is only the premise most similar to the conclusion category that affects the similarity component. Moreover, the similarity of F_2 and T is excluded from the coverage component as well, since the conclusion category does not figure in the coverage component as such, as SCM considers the relation between F_1 and F_2, on the one hand, and the superordinate category spanned by F_1 F_2 and T, on the other. In contrast, the other three models use both $F_1 \rightarrow T$ and $F_2 \rightarrow T$ to update the degree of belief in the conclusion. Other things being equal, both premises increase confidence in the conclusion, albeit to a different extent.

Diversity and Segregation

Let us now turn to the premise diversity effect, and begin by pointing out that the effect should be separated in two. According to the premise diversity effect, adding an additional premise fairly different from the first will increase the strength of the conclusion. This effect has been very widely studied and was even dubbed "one of the hallmarks of human induction" (Heit and Hahn, 2001). However, an opposite line of reasoning (suggested in passing in Heit, 2000) must also be considered. Suppose F_1 and F_2 are close to one another, relatively to T, as in the argument (Mouse Hamster → Horse). That argument may actually appear less convincing than (Mouse → Horse), because the property asserted to be true of Mouse and Hamster may be supposed to be idiosyncratic and restricted to rodents, a conjecture that does not come to mind when Mouse and Horse only are presented. Let us call this putative effect *segregation*. We define segregation as cases where the introduction of an additional premise F_2 to an argument F_1→T weakens the conclusion when F_1 and F_2 are similar. Both effects refer to an inverse relation between premise similarity and strength of the conclusion, but whereas the diversity effect posits F1→T as its lower boundary, the segregation effect has no such boundary.

SCM explains readily the diversity effect by the coverage model. Other things being equal, the more distant F_1 and F_2 are from one another and the more extensive their coverage of the superordinate category. What about segregation? According to SCM, adding an additional premise F_2 to an argument of the form F_1 → T can never decrease the strength of the conclusion: the similarity component will either increase, in case F_2→ T is stronger than F_1→T, or it will remain the same otherwise, whereas joint coverage can only increase. The one exception allowed is the case of non-monotonicity, described earlier, where coverage decreases when the superordinate category is adjusted.

Heit's account likewise predicts the diversity effect. Further, it too predicts that adding a premise won't decrease the strength of the conclusion. This is because, in his Bayesian terms, additional evidence increases belief in the conclusion if it is relevant and in the worst case, the additional premise is irrelevant, and so does not modify it. To use one of his examples: "Intuitively, cows and horses are so similar that finding out that horses have some property adds little information when you already know that cows have the property." Segregation therefore is not predicted.

Sloman's model too accounts for the diversity effect. When updating the predicate's features vector, the model takes into account the extent to which the new property is actually novel with respect to those features already known to be associated with the predicate: the more novel the premise, the more extensive the updating. More diverse premises cover the space of features better than more similar premises because their features are not redundant and are therefore more likely to overlap with features of the conclusion category. Again, this model does not predict segregation, because features that are shared by the premises but are not shared with the conclusion are simply ignored by the model, whereas the segregation effect is precisely defined by reference to features shared by the premises but not be found in the conclusion category.

The one model that predicts a weakening of the argument when a similar premise is added is that of Tenenbaum and Griffiths. As they demonstrate, the tapering off of consequential regions becomes more marked when more example cluster close together. More formally: the probability that the consequential region C extends a distance d beyond the examples rapidly *decreases* as the number of examples *increases* within a fixed range. This notable result comes about because for most previous models "the generalization gradients produced by multiple examples of a given consequence are essentially just superpositions of the exponential decay gradients produced by each individual example" whereas their analyses computes the joint consequential region of several

premises in a more sophisticated way, and explain how encountering additional consequential stimuli can cause the probability of generalizing to some new stimulus to decrease. Table 1 summarizes the differing predictions so far.

Insert Table 1 about here

The role of the concepts hierarchy

Beyond the contrasting predictions made by the models, there is an additional difference between them. SCM relies explicitly on the hierarchical structure of concepts, inasmuch as coverage is defined in terms of the lowest-level common category in the hierarchy. The other models make do without this concept. For Sloman, all features are of equal value, and the property of belonging to a given superordinate category is not treated differently than possessing any other specific attribute. For Hayes, every premise category is considered by itself, without reference to a hierarchy, and the same is true of Tenenbaum and Griffiths. Their concept of consequential region is very different from a pre-existing superordinate categories scheme. There is such a region "around" each category, (at least if it is a natural kind), and for any combination of categories, their joint consequential region can be computed by a suitable algebra.

This is of course not to say that any of these researchers would deny that there are superordinate categories, at least in certain domains, but only that hierarchical structure has not special status in deriving conclusions that do not explicitly refer to them. Osherson et al (1990) introduced the distinction between general and specific arguments. In *general* arguments, the conclusion category is superordinate to at least some of the premises categories (Hyenas Cows → Mammals). In *specific* arguments, all the categories involved belong to the same hierarchical level (Hyenas Cows → Wolves, all Mammals). In our experiments, we will accordingly use specific arguments, and pit against one another effects of similarity and effects of hierarchy. According to

SCM, effects of hierarchy should be found, whereas for the other three models, similarity alone should explain the patterns of results.

Effects of Pace

It may be possible to shed some additional light on the nature of the cognitive processes that implement categorical induction by studying the effects of time pressure on performance. To predict such effects, let us consider the two reasoning systems that have been discussed by many authors, the one inductive, associative and the other deductive, analytical, "rule-based" (Sloman, 1996; Sloman and Rips, 1998; Osherson, et al., 1998; Rips 2001). These are two distinct ways to evaluate an argument. The associative inductive response is produced automatically once the problem is presented; the symbolic, deductive process may take longer and, if used, can delay the production of the response. The faster functioning of the associative system has been observed for instance in studies on the effect of time pressure on categorization (e.g. Ward, 1983).

The distinction between the similarity component and the coverage component proposed by SCM may map onto these two systems. Carey (1985), for example, suggested that the similarity component is more primitive, and therefore accessible to younger children, whereas components that rely on hierarchical structures are only used by older children. Similarly, Lopez et al (1992) failed to find diversity effects with 5-year old children, and found diverse effects with 9-year-olds. This pattern of results was explained by children's deficient mechanism for generating superordinate categories. Similar results were reported by Gutheil & Gelman (1997), but interpreted by them as a difficulty in using diversity information in their inductive judgments, and not just as difficulty in generating superordinate categories. Lo, Sides, Rozelle, Sides, and Osherson (2002) gave 9 and 11-years-olds the task of comparing arguments and selecting those set of premises that best support conclusions about superordinate categories and found no awareness of the uses of diversity, not even in those cases were they were found by Lopez et al. (1992). In contrast, Heit & Hahn (2001)

did find diversity effect in children, and ascribed the differences in their findings, when compared to that of their predecessors, to the contents used.

Following this logic generates another contrasting prediction. SCM leads us to expect that time pressure will change the relative weights for the two components: the similarity component should become more important under time pressure, whereas the coverage component, which relies on a more analytic exploitation of the hierarchical categorization of the domain, will matter less. Conversely, if time pressure does not lead to a different pattern of inferences, this will lend at least some support to the other models that posit no such distinction in processing.

Points of method

In this context, the selection of the specific task and the design of the procedure may be of especial significance. Previous researchers have typically assumed that the different experimental paradigms address the same underlying processing (Heit 2000). Yet it is more likely that special instructions affect the mode of processing (analytical as opposed to associative). In particular, tasks requiring meta-cognitive awareness may encourage a more analytic approach. This has consequences for the experimental procedure. Asking subjects to compare the strength of two arguments encourages a reflective, "formal operational" stance in Piagetian terminology, because the arguments are not actually asserted, but evaluated and compared as mental objects. By contrast, when the premises are asserted and the likelihood of the conclusion is to be evaluated, the arguments are not so much considered as processed. This could for instance explain the difficulties of the young children in the study by Rozelle, Sides, and Osherson (1999), when compared with those of Lopez et al (1992), while the same line of argument may explain in part the unusual success of young subjects in Heith and Hahn (2001).

Most previous studies used a forced-choice paradigm, where subjects compared the strengths of two arguments. If we note premise categories by F_1, F_2 etc., and the conclusion category

by T, we may state that most previous studies compared the strength of two arguments $F_1 \rightarrow$ T and $F_1F_2 \rightarrow$T (usually finding that $F_1F_2 \rightarrow$T is judged stronger than $F_1 \rightarrow$T); or compared $F_1F_2 \rightarrow$ T and $F_1F_2' \rightarrow$ T, that is, one of the premises is replaced by another and the effect on argument strength was assessed. The comparison was done by presenting such pairs of arguments to groups of subjects, and determining which of the two is judged as a stronger argument by more subjects. Our study relies on direct evaluation of individual arguments. Further, rather than inviting judgments about the (comparative) strength of the arguments, we will ask them to evaluate the strength of their conclusions. Subjects will be told to assume that the premise(s) is/are true, and requested to judge how likely it is that the conclusion is true too.

We will not only obtain such judgments for the conclusion of the complex argument, but also collect information on every pair of categories involved in the arguments. Thus, for every argument of the form $F_1F_2 \rightarrow$T, we will request a judgment on $F_1F_2 \rightarrow$T (i.e., how likely is it that T has property P if we assume that F_1 and F_2 have it). But we will also elicit the subjects' judgment on all the simple relations involved: $F_1 \rightarrow$T (i.e., how likely is it that T has property P if we assume that F_1 has it?), $F_2 \rightarrow$T, $F_1 \rightarrow F_2$ and $F_2 \rightarrow F_1$ (see Figure 1). This will enable us to perform within-subjects comparisons and analyses throughout.

The literature includes extensive discussion of the influence of specific predicates on categorical induction (Sloman, 1994; Heit & Rubinstein; 1994, Sloman, 1997; Heit & Hahn, 2001; Lo et al, 2002) We are interested in the way consequential categories are used and combined, not in the added complexity introduced by the contents of the predicate and their interactions with the categories. Since the predictions for the different models can be made without reference to specific relations, we avoided the known but incompletely understood additional complexities of interactions with known predicates. Our study will not be concerned with this aspect of the question and relies on so-called "blank" properties, such that the subjects cannot have prior knowledge about them.

Insert Figure 2 about here

Experiment 1 - Similarity Evaluation

In order to construct the stimuli for the first main experiment, we needed to collect information on the similarity structure of the animal categories for our subjects and ran a preliminary experiment.

Method

Subjects. Ninety students, freshmen in a college psychology program, served as subjects for this preliminary experiment.

Procedure. The subjects completed a questionnaire that asked for their judgment about the similarity between pairs of animals on a five-point scale. The instructions were very terse: please indicate on the scale besides each pair of animals the extent to which you judge them similar. The columns were labeled 1-5, and in addition, the words *very low / very high similarity*[2] were written under "1" and "5".

We took animals from four main categories: mammals, insects, reptiles, and birds. The *Mammals* included the following animals, regrouped here according to sub-categories when appropriate: Tiger, Cat, Lion (*Felines*); Horse, Zebra, Donkey (*Equines*); Hamster, Squirrel, Mouse (*Rodents*); Rhinoceros, Hippopotamus (*Pachyderms*); Rabbit, Giraffe, Pig; Bat, and Dolphin. The *Reptiles* were Snake, Lizard, Alligator, and Turtle. The *Birds* included Falcon, Hawk, Eagle (*Birds of Prey*); Wagtail; Goose, Chicken (Fowl). The *Insects* were Fly, Bee, Mosquito, Grasshopper, and Beetle. (The words in italics appeared nowhere in the questionnaire.)

Subjects were presented with pairs of animal names. In view of the very large number of possible pairs, we restricted the study to two still large sub-sets as follows. All the possible pairs of animals inside each main category were presented, (156 pairs in all). In the comparison across

different categories, some animals were removed: the rhinoceros and the hippopotamus were not included (hence were only compared to other mammals). Further, each bird was compared to all the animals in the bird, reptile and insect categories, but only to some of the animals in the mammal category (namely: Horse, Tiger, Hamster, Mouse, Rabbit, Pig, and Bat). This yielded 266 between-categories pairs. These 422 pairs were randomly divided between three different groups of 30 subjects each.

Results

The subjects' evaluations of similarity between animals in the various groups were combined to compute overall, across groups mean similarity between pairs of animals. This was done for the two sets of data for which we collected all the necessary information, one without the birds, the other with the birds but without some of the other animals. The data concerning both sets of animals was subjected to a multi-dimensional scaling algorithm from which a two dimensional map was derived (Figure 3 shows the resulting map for the first set of animals.). The map reflects the animal taxonomy (divided into mammals, insects, and reptiles). Animals that belong to the same higher order category are grouped together. Likewise, in the subjects' similarity evaluation we find secondary divisions of the general categories. For instance, in the mammal category there is a distinction between the rodents cluster (Hamster, Mouse, Squirrel) and the felines cluster (Lion, Tiger, Cat), while in the insect category one can identify a cluster of flying insects (Mosquito, Fly, Bee). The overall disposition on the map is consistent with several properties. For instance, larger animals are located to the left, smaller ones to the right. Habitat, gross morphology, etc. all seem to play some role. In particular, note how certain animals (Worm, Bat, and Dolphin) are close to animals from a taxonomically different category, a feature we exploit in Experiment 2.

Insert Figure 3 and Figure 4 about here

We also ran a hierarchical cluster analysis of similarity among the sets of animals. We used Ward's method to build the clusters (maximizing the between-clusters variance, while minimizing the within cluster variance in every stage of the analysis). This analysis resulted in findings similar to the MDS, as illustrated in the dendogram of Figure 4 for the set including the birds. The main categories are found (mammals, birds, reptiles, and insects), and the secondary division of the categories is apparent too (e.g., rodents). Lastly, some animals are judged as more similar to animals of a different category, and therefore grouped with them (e.g., bat, judged as more similar to birds than to other mammals), a feature we will make use of below.

Experiment 2

This experiment investigated how subjects evaluate the likelihood of a conclusion based on inductive arguments with animals, and the effect of time pressure on their performance. The stimuli were constructed with reference to the categorical structure identified in Experiment 1.

Method

Subjects. Forty-eight college freshmen served as subjects as a course requirement.

Design. The categorical arguments involved animals chosen from the "animal taxonomy" built in Experiment 1. The properties used were "blank" (invented by us for this experiment), hence the subjects couldn't have had previous experience with them. The properties are: G.L.D. mechanism, Goldner mechanism, Turner mechanism, PERI mechanism; Burner capillaries, Ackslen capillaries, Verner capillaries; Ashter tissue, Sufnir tissue, Leshem tissue, Shiffonit tissue; Golder system, Syndon system, MEDI system, S.P.I. system.[3] These properties were randomly placed in the arguments.

In order to evaluate the role of the categorical hierarchy on complex induction cases, stimuli were sampled from the three possible patterns of category membership (hereafter PCM)[4]: *AAA*: both premises and the conclusion category belong to the same superordinate category (e.g., Horse,

Cat → Rabbit i.e., *Mammal Mammal → Mammal*); *ABB*: the two premises are members of two different intermediate categories and the conclusion belongs to one of these (e.g., Cat, Snake→Lizard i.e., *Mammal Reptile → Reptile*). *AAB*: both premises are members of a same superordinate category and the conclusion category belongs to a different one (e.g., Fly, Grasshopper → Tiger i.e., *Insect Insect → Mammal*).

We thought it likely that subjects vary in their judgments of the relative similarity of different categories, and therefore needed to be in a position to perform all the analyses within-subject. Since a large number of trials was needed, we separated the subjects at random into two groups (24 subjects each) and every group answered about half the complex induction cases (involving two premises) as well as the four simple induction cases derived from them, to enable within-subject analyses. Thus, if a subject received a given complex argument of the form $F_1F_2 \rightarrow T$, we presented him or her also with the simple arguments: $F_1 \rightarrow T$, $F_2 \rightarrow T$, $F_1 \rightarrow F_2$, and F2 $\rightarrow F_1$. Specifically, Group 1 was presented with forty-one complex and their eighty-four constituent simple arguments, and Group 2 with forty-eight complex and eighty-nine simple arguments (the difference in numbers stemming from the slightly different pattern of overlap possible for each set of questions.)

The two paces (Slow/Fast) were blocked, and the blocks were counterbalanced. Every block had two parts, first *all* the simple (single premise) arguments, then all the complex (two premises) arguments for that subject. Each experimental part began with three practice trials. A typical experimental session could therefore consist of: speeded-simple, speeded-complex, slow-simple, and slow-complex, each preceded by the appropriate three practice trials.

Procedure. The arguments were presented on a computer screen to subjects sitting in individual booths. Every argument was shown in two steps: first its premise(s), then the conclusion. Figure 5 illustrates the appearance on screen of the argument: Hyena, Cow → Wolf (which represents the question: "*If Hyenas and Cows have the Burner system, how likely is that Wolf do too?*"). Every

argument was presented twice – once at a fast pace and once at a slow pace. The procedure at the fast pace counted three steps (see Figure 5a). The first screen (with premises and property) was shown for 2 sec. The second screen (adding the conclusion category) was then shown, and the subjects had 2 seconds to answer. Finally, a tone was sounded to indicate the end of the trial. The instructions given to the subject were: "Answer quickly from the moment of presentation of the conclusion category and until the tone is heard. Late responses will be invalidated".

Insert Figure 5 about here

The procedure for the slow pace was as follows (see Figure 5b): The first screen (with premises and property) was shown for 4 seconds, whereupon the second screen (adding the conclusion category) was shown, also for 4 seconds. A tone was sounded, indicating to the subjects that they were now allowed to answer. They then had 8 seconds to answer, and the trial was terminated. The instructions were: "You may answer only after the tone is heard (not before) and until the stimulus disappears. Early responses will be invalidated." The time intervals were selected after extensive pre-testing. In particular, those for the slow pace were derived from subjects' behavior with no time limit imposed on their answers. Single premise arguments were presented in like manner, with the lone premise centered horizontally. Subjects answered by pressing one of the keys on a five-key custom-made keyboard; they were instructed to press the leftmost key when they thought the conclusion was not very likely, the rightmost key when they felt it was very likely, and intermediate keys to express intermediate values.

The experimental session ended with the verification of the categorical membership attribution by the subjects. The SCM refers to the categorical membership of animals and the similarity between them. Since we make special use of those animals that belong to one category but are similar to animals in another, it was important to make sure that the category membership of the

animals is known to the subjects. Subjects were requested to classify the animals mentioned in the experiment into four categories, by means of a straightforward forced-choice paper questionnaire.

Results

Category Membership Subjects' responses show a high percent of correct answers, between 87.5% -100%. The lower values were those where similarity and categorical membership were in conflict, as for example, when they categorize Bats with Birds, or Worms as Reptiles. In analyzing the data, responses to arguments that involve misclassified animals were discarded. For example, if a subject classified *Bats* as *Birds*, all that subject's judgments relative to arguments involving bats were discarded.

Evaluating Categorical Arguments First, let us repeat an important notational convention. We will compare the evaluation of complex induction argument ($F_1F_2 \rightarrow T$) and those of the constituent single premise arguments ($F_1 \rightarrow T$; $F_2 \rightarrow T$). The single premise argument that was evaluated as the strongest was called $F_1 \rightarrow T$, and the other $F_2 \rightarrow T$. For every subject and every complex argument, $F_1 \rightarrow T$ is therefore higher than $F_2 \rightarrow T$ by definition. To illustrate: in the argument (Horse Hamster → Mouse), if the subject judged the simple argument (Hamster → Mouse) as stronger than (Horse → Mouse), the indices were assigned as follows: F_1 $F_2 \rightarrow T$ (Horse Hamster → Mouse); $F_1 \rightarrow T$ (Hamster → Mouse); $F_2 \rightarrow T$ (Horse → Mouse). The rationale for this is that the two premises of the complex arguments always appeared simultaneously so that there is no intrinsic order. On the other hand, being the "closest neighbor" is an important characteristic of such argument, whether in the pioneering work by Carey (1985), in SCM, or indeed, as the foundation of "nearest neighbor algorithms" in general (Duda, Hart, & Stork, 2001). We made an exception for arguments of the form ABB (e.g., Gorilla Fly → Bee), where the interest centers on whether the strength of the conclusion of the simple inference Fly → Bee is actually weakened by the addition of the additional

premise Gorilla. Accordingly, these cases were coded F_2=Gorilla, F_1 =Fly and T=Bee, so that F_1 and T belong to the same sub-class, and F_2 will always be the "inappropriate" category.

In what follows, we will use the same codes to refer both to strength of beliefs in the conclusion of arguments, and to the arguments themselves, and rely on the context to disambiguate. Using this convention, we operationalized premise diversity as $\frac{1}{2}[F_1 \rightarrow F_2 + F2 \rightarrow F_1]$, that is, as the average of the belief in F_2, given F_1, and that of the belief in F_2, given F_1 (see Figure 2). In view of the very high correlation known to exist between projectibility and similarity, this seems a reasonable operationalization, which enabled us to perform within-subject statistical analyses. Our notation for this variable will be F_1—F_2 (a symmetric relation).

Predicting the Strength of a Complex Argument:

How is the strength of a complex argument related to the strength and the diversity of its constituents? Our first step was to compute a regression analysis, to predict the strength of the conclusion of the complex inference $F_1F_2 \rightarrow T$ from that of the four related simple inferences. This was done separately for the two paces, slow and fast, and Table 3 presents the results of the analysis.

Insert Table 3 about here

Subjects consider both premises in evaluating the conclusion. Remember that we defined F_1 and F_2 so that F_1 is the premise from which the subject in question was most willing to draw the conclusion of the complex argument. As is consistent with all models, the strongest predictor is $F_1 \rightarrow T$. However, the second premise ($F_2 \rightarrow T$) also plays a non-negligible role, and the third variable (F_1—F_2) is significant too and explains half of the (little) residual variance. Together, the three variables account for virtually all of the variance in the slow pace condition (R^2=0.945).

Time pressure had no material effect on the regression equation. In the time pressure condition, all three variables enter the regression equation, in the same order as with the slow pace,

and with practically the same beta values. The proportion of the variance explained remains very high (R^2=0.825), even if somewhat lower than in the absence of time pressure.

We return below to the findings of Experiment 2, but first report a replication of the main findings in a different substantive domain.

Experiment 3: The Languages Domain

Classical cognitive anthropology studied folk classification of plants and animals (Berlin, Breedlove, and Raven 1973; Berlin 1992; Ellen 1993) on the assumption that the mind approaches biological classifications and that of other objects in the same way. However, Atran's cognitive anthropological work (1990) developed the view that folk-biological knowledge was based on a domain-specific approach to living things characterized by specific patterns of categorization and of inference. The predictions he derived from this view have been mostly confirmed (Medin and Atran, 1998). For the sake of generality, we accordingly studied whether the same patterns of inferences would obtain in a different domain, and selected that of the relations between languages. People perceive languages as being more or less similar to one another and as belonging to some wider categories, commonly called families, such as the Latin, Germanic and Semitic language families. At the same time, whatever may have been the evolutionary pressure responsible for the special human cognitive preparedness to classify living things, it is highly unlikely that it would be relevant to meta-linguistic awareness of the sort required to compare and classify languages.

Method

Subjects. Fifty college freshmen, all born in Israel and native Hebrew speakers, served as subjects to fulfill a course requirement.

Design and Procedure. The study was based on three questionnaires, all involving relations between languages. The set of languages consisted of Bulgarian, Japanese, Spanish, Portuguese, Italian, and German. The first questionnaire asked about the *similarity* of every pair of languages in

the set. Subjects were required to indicate the degree of similarity on a 7-point scale (with the endpoints marked as *not at all similar — extremely similar*). The second questionnaire returned to the same pairs, and assessed *projectibility* of blank predicates from *single premises*. A set of pseudo-linguistic properties was contrived ("*adverbial cleft backposition*", "*verb anteposition*" and the like). Questions were of the form: *Assume that language L_1 has property P (where P is one of the invented properties). How likely is it that language L_2 has that property? Language L_3? Language L_4, etc.* Projectibility was assessed in one direction only, and subjects answered on a 7-point scale (*not at all likely — extremely likely).* The third questionnaire was devoted to complex categorical inference $F_1F_2 \rightarrow T$, and assessed *projectibility from two premises*. Subjects were presented with a series of questions of the form: *Assume that languages L_1 and L_2 share property P (P was again one of the invented properties). How likely is it that language L_3 has that property? Language L_4? etc.* and answered on the same 7-point scale. There were 23 such complex arguments, and all were answered by all subjects.

Results

First, we made certain that the languages domain as perceived by our subjects had the anticipated hierarchical structure. To this end, we subjected the data collected from the second questionnaire to a hierarchical cluster analysis. We computed the closeness of any two pairs of languages by the mean degree of projectibility between them. The resulting clusters are shown in Figure 6. A cluster analysis based on similarity (as elicited by the first questionnaire) yielded the same hierarchical cluster structure, which is unsurprising in view of the high correlation we found between similarity and projectibility (Pearson r=0.92).

Insert Figure 6 about here

Next, we turned to the replication of the regression results. The questionnaires were devised in such a way that for every judgment of every complex inference $F_1F_2 \rightarrow T$ by a subject, we have also that subject's judgment of $F_1 \rightarrow T$, $F_2 \rightarrow T$ and F_1—F_2. Since F_1—F_2 and $F_1 \rightarrow T$ are highly

correlated in the data (Pearson r =.69, p=.0002), ridge regression analysis was used. (Ridge regression artificially decreases the correlation coefficients so that more stable -- if biased -- beta coefficients can be computed.) The results of the ridge regression (lambda=.1) with F_1 F_2 →T as dependent variable are shown in Table 4. The fundamental pattern of Experiment 2 was therefore replicated. The proportion of the variance explained was lower, R^2=.65, presumably due to the much smaller number of data points. However, all three variables contribute significantly, and the relative size and directions of the Beta coefficients are comparable to those found in Experiment 2.

Insert Table 4 here

Segregation

We now turn to the other contrasting predictions of the models. To do so, we need a rich domain, and the language domain was not sufficiently complex to enable the required testing. We must therefore rely only on the findings of Experiment 2.

To test the segregation effect, we first compare the AAB and the AAA cases. To recall, AAB cases are arguments where F_1 and F_2 belong to the same superordinate category and T belongs to another one, whereas in AAA cases all three categories belong to the same superordinate category. ABB cases are the very ones where non-monotonicity is predicted by SCM. The segregation effect should be found in AAB cases, and not in AAA cases. Figure 7 shows the relevant means. The important comparison concerns the interaction of argument (F_1→T vs. F_1F_2 →T) and case (AAA vs. AAB). It may be seen that in the AAA cases, F_1 F_2 →T is not materially different from F_1→T, whereas in the AAB cases F_1F_2 →T is lower than F_1→T. A planned comparison analysis of variance confirmed this difference. There are 17 cases of AAB and 27 cases of AAA. F(1,41)= 16.3, p=0.00023. A planned comparison of the two cases AAA and AAB showed that these two PCM cases differ significantly [F(1,41)=16.31 p=.00023]. Pace did not interact significantly with these

results [F(1,41) = 0.26]. Again, there is no evidence of an effect of working pace on the pattern of results.

Insert Figure 7 and Figure 8 about here

Another place to look for the segregation effect is *within* different subsets of AAA cases, for instance, within the class of Mammals. We are interested in comparisons such as (Hamster, Cat→ Horse) and (Hamster Mouse → Horse). All the animals involved are of course mammals, but Hamster and Mouse are especially similar to one another, being both rodents. Starting from the same simple argument (Hamster →Horse) we will be interested in contrasting two comparisons: adding an unrelated premise (Hamster *Cat* → Horse) and adding a related premise (Hamster *Mouse* →Horse). The segregation effect would consist in a weakening of the argument when the related premise is added, and no such weakening when the unrelated premise is added. We built nine such cases, and, after controlling for misclassification, used the data from 45 subjects. The relevant means are presented in Figure 8. Planned comparison shows no effect of adding the unrelated premise ("baseline cases") whereas the argument strength decreases when the additional, related premise is added ("special cases") [F(1,44)=5.8 p=.02]. Again, pace did not interact with the type of case. We computed accordingly a planned comparison of the type of case (baseline or special) across Slow and Fast pace, and this too proved significant [F(1,44)= 5.80, p=.020].

We conclude that the segregation effect was demonstrated. The results of the regression analysis indicated that the more distant the premises are from one another, the stronger the argument. We have now further established that this effect goes beyond what was allowed for by SCM, Heit's and Sloman's model, and that when premises are close enough, the argument is actually weakened.

The Role of the Concepts Hierarchy

An additional difference between SCM and the other models is the role of the hierarchical category structure. SCM gives an essential role to that hierarchy, which is the basis of the coverage component. The other models confer no special role to superordinate category, and rely entirely on similarity or generalizability to predict argument strength. To decide between the two approaches, we return to the data of Experiment 2. Recall the "non monotonicity" phenomenon where adding a seemingly "improper" premise (e.g. Horse, *Salmon* →Bat) was found to weaken the argument (Horse →Bat). This was due, according to SCM, to the widening of the superordinate category. The relevant data of Experiment 2 are the cases that we called ABB, where A is the "improper" premise (Figure 7 above).

Insert Figure 9 about here

Such a pattern of results should be interpreted with caution. Remember that the experimental procedure called for asking all the questions about single premises first, followed by all the questions about complex arguments. This order was chosen in order to prevent subjects from making deliberate comparisons between these two parts of the experimental session, and so obtain independent estimates of the strength of the conclusion in each case. Under these conditions, each block induces a range of variability, which is mapped by the subject to the five keys of the keyboard (see Goldstone, Steyvers, Spencer-Smith and Kersten, 2000). It may therefore be misleading to compare absolute values concerning simple and complex arguments. We cannot evaluate the effect of non-monotonicity on the basis of our paradigm, since by rights it supports only the analysis of contrasts between cases and conditions. Still, it is possible to exploit such a contrast between different cases of predicted non-monotonicity to arbitrate between the accounts of categorical induction. The SCM's hierarchy-based account predicts that adding an improper premise, one forcing an adjustment of the superordinate category, should decrease the strength of the argument.

According to the other models, the strength of the argument depends on the similarity of the additional premise category to the conclusion category. This led us to investigate whether the non-monotonicity effect would be weaker when the "improper premise" is similar to the conclusion, even when belonging to a different sub-category. To test this argument, we built specific pairs of arguments of the ABB type, $F_1F_2 \rightarrow T$ and $F_1F'_2 \rightarrow T$, in such a way that the strength of $F_2 \rightarrow T$ is greater than $F'_2 \rightarrow T$, F_1 and T belong to the same category B, and F_2 and F'_2 both belong to another category, A. Consider for example the argument (Horse→Bat), where both are mammals (category B). What would be the effect of adding either Vulture or Chicken, both birds (category A)? If the feature-based account is correct, one can expect that adding a premise such as Vulture (relatively similar to Bat) would weaken the argument less than adding a premise more different from the conclusion (such as Chicken).

We constructed five such contrasting pairs of complex arguments in all, using the relevant information from Experiment 1. The analysis involves 33 subjects, again excluding those subjects who misclassified the animals involved (usually Bats). We ran a two-way within subject ANOVA. The independent variables were working pace (Fast/Slow) and argument type, the latter had also two levels: (1) "special" arguments with the "improper premise" similar to the conclusion (e.g. Eagle Horse→ Bat), and (2) "baseline" arguments in which the similarity is much smaller (e.g., Chicken Horse→Bat). We found a significant effect of argument type [$F(1, 55)= 27.88$, $p = 0.000002$] (see Figure 9). Once again, Pace had no effect, whether singly or in interaction (both $F(1, 55)<1$).

The SCM account in terms of the forced change of superordinate category cannot explain these findings. The relevant superordinate category, that would have to include Birds, is vast (Vertebrates, perhaps). Surely, the coverage of that large and varied category by Horse together with either Eagle or Chicken is less extensive than that of Mammals by Horse. In fact, when the added

premise was similar to the conclusion, even though not belonging to the same category, the argument was significantly strengthened.

Conclusions

It is possible to predict with very high accuracy the strength of a complex argument, from the strength of its component arguments and the similarity between the premises. This was demonstrated on two very different substantive domains: the animal kingdom and that of language families.

The stronger component of the simple arguments has a larger effect than the other, but both do contribute to the judged overall strength of the argument. This finding contradicts the SCM model. The model may be rescued by assuming that the values of $F_1 \rightarrow T$ and $F_2 \rightarrow T$ are determined probabilistically, and are moreover so variable that sometimes $F_1 \rightarrow T > F_2 \rightarrow T$ will hold, and sometimes the converse it true. However, in view of the relative sizes of the Beta's (that of $F_2 \rightarrow T$ being about 2/3 of that of $F_1 \rightarrow T$), this explanation assumes such large variability as to make any model testing very problematic. The finding is consistent with the other models.

Next, we replicated the well-known finding that increasing the diversity of the premises increases the strength of the argument, an effect predicted by all models. But we also found a hitherto undocumented effect: adding premises can actually decrease the strength of the complex argument, when the two premise categories are very similar to one another. We called this novel effect *segregation*, and were able to demonstrate it in the two distinct contexts where it was expected to manifest itself: in a comparison of AAA and AAB cases, and in contrasting two types of AAA cases, when the two premises belong to the same sub-category of the superordinate category. This finding is only consistent with the Tenenbaum and Griffiths model but not with the others.

When we pitted category membership and similarity against one another, we found no signs of reliance on pre-existing hierarchical categorical structure. It seems that under the conditions of our experiments, judgments were based only on the similarity-based component, whereas category inclusion information and rule-based considerations were neglected.[5] Lastly, we found that forcing the subjects to work at a fast pace, or allowing them to work more leisurely, did not materially affect the patterns of results. These two results militate against the SCM model, and are consistent with models that do not posit two components.

One may wonder about the contradictory results found here and by Osherson and his collaborators. Two explanations may be suggested. First, we used a different experimental paradigm. A task requiring subjects to compare the strength of two arguments encourage a reflective and more analytic approach than ours, where the argument is merely presented to the subject, who then evaluates the likelihood of the conclusion. Our paradigm would also seem to be more ecologically valid. Second, we collected a much wider range of cases and measures, and obtained these in a way that allowed intra-subject analyses. We could therefore be more accurate in evaluating the exact contribution (or non-contribution) of each to the strength of the conclusion.

In any event, we do not contest the existence of stable categories in certain substantive domains such as the animal kingdom, nor do we doubt that these may be used in the course of some types of argument comparisons. When a person is confronted with a categorical argument, they have distinct ways to evaluate it (Rips, 2001). Our more modest claim is that categorical induction does not habitually involve the rule-based approach. Allowing a leisurely pace of work does not induce subjects to switch to the more demanding rule-based processing. This thesis is supported by an additional finding by Sloman (1988), who found extensive support for the inclusion similarity phenomenon: while arguments of the form Animals→Mammals are judged as stronger than those of the form Animals→Reptiles, they are not judged maximally strong; remarkably, if a premise is

added stating explicitly *"All Mammals are Animals"* then the argument *is* judged perfectly strong. This is presumably because that additional premise elicits the use of the rule-like approach. Without it, subjects use the similarity-based processing.

The one model that predicts all four findings (see Table 1) is that by Tenenbaum and Griffiths. However, this model is a computational level account that makes no claims about the processing steps involved. Further, our study involved specific arguments and blank predicates only. This is a serious restriction, in view of the recent work suggesting that the crucial issue to be captured by modeling is that of the interaction of knowledge and induction (Heit and Hahn, 2001). Reasoning is a general-domain skill, of the type analyzed by Piaget, but specific knowledge has an essential influence on the way that skill is conducted. Modeling the process by which these two components are combined remains very much a challenge.

References

Biton-Bereby, L. (1999). *Categorization under time pressure.* Unpublished MA Thesis, Ben-Gurion University, Beer-Sheva. (Hebrew)

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

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification.* New York: Wiley.

Eysenck, M.W. & Keane, M.T. (1990). *Cognitive Psychology.* Hove, UK: Erlbaum

Goldstone, R. L., Steyvers, M., Spencer-Smith, J., & Kersten, A. (2000). Interactions between perceptual and conceptual learning. In E. Dietrich & A. B. Markman (Eds.), *Cognitive Dynamics: Conceptual and representational changes in humans and machines* (pp. 191-228). Mahwah, NJ: Erlbaum.

Grice, H. P. (1989). *Studies in the Way of Words.* Cambridge: Harvard University Press.

Gutheil, G., & Gelman, S. A. (1997). Children's use of sample size and diversity information within basic-level categories. *Journal of Experimental Child Psychology, 64*, 159–174.

Heit, E. (1997, November 28-30). *Features of Similarity and Category-Based Induction*. Paper presented at the Interdisciplinary Workshop on Similarity and Categorisation, Edinburgh, Scotland.

Heit, E. (2000) Properties of inductive reasoning. *Psychonomic Bulletin & Review* 7 (4) 569-592.

Heit, E., & Hahn, U. (2001). Diversity-based reasoning in children. *Cognitive Psychology, 43*, 243-273.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Language, Memory and Cognition,* **20**(2), 411-422.

Lo, Y., Sides, A., Rozelle, J., & Osherson, D. (2002). Evidential diversity and premise probability in young children's inductive judgment. *Cognitive Science, 26*(2), 181-206.

Lopez, A., Gelman, S. A., Gutheil, G., & Smith, E. E. (1992). The development of category-based induction. *Child Development, 63*(Oct), 1070-1090.

Medin, D. & Atran, S. (1998). *Folk Biology*. Cambridge: MIT Press.

Osherson, D., Perani, D., Cappa, S., Schnur, T., Grassi, F., & Fazio, F. (1998). Distinct brain loci in deductive versus probabilistic reasoning. *Neuropsychologia*, **36**, 369-376.

Osherson, D.N., Smith, E.E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category based induction. *Psychological Review,* **97(2)**, 185-200.

Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science,* **12**, 129-134.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, **237**, 1317-1323.

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology,* **35(1)**, 1-33.

Sloman, S. A., & Rips, L. J. (1998). Similarity as an explanatory construct. *Cognition,* **65**, 87-101.

Sloman, S.A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin,* **119**(1), 3-22.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in Neural Information Processing Systems* (Vol. 11,): Cambridge, MA: MIT Press, 1999.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, Similarity, and Bayesian Inference. *Behavioral and Brain Sciences*, **24**(2), 629-641.

Ward, T.B. (1983). Response tempo and separable-integral responding: Evidence for an integral-to-separable processing sequence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance,* **9**, 103-112.

Table 1

Contrasting predictions of the four models.

	F_1 T	F_2 T	F_1 F_2	
			Diversity	Segregation
SCM	+	–	+	–
Heit	+	+	+	–
Sloman	+	+	+	–
Tenenbaum and Griffiths	+	+	+	+

Table Note: + : predicted ; – : not predicted

Table 2

ANOVA of main variables

Effect	df Effect	MS Effect	df Error	MS Error	F	p-level
Question	2	101.77	94	.53	191.17	.000000*
Speed	1	12.48	47	.90	13.85	.000527*
PCM	3	98.21	141	.55	177.53	.000000*
Question * Speed	2	1.06	94	.15	6.80	.001736*
Question * PCM	6	2.37	282	.20	11.75	.000000*
Speed * PCM	3	5.90	141	.14	39.73	.000000*
Question * Speed * PCM	6	2.18	282	.11	18.97	.000000*

Table 3

Regression results – Experiment 2

		ß	St.err of ß	B	St.err of B	*p*	Cumulative R^2
SLOW							
	F1T	0.679	0.046	0.614	0.042	0.0000	0.886
	F2T	0.443	0.057	0.490	0.063	0.0000	0.905
	F1F2	-0.277	0.042	-0.219	0.033	0.0000	0.945
FAST							
	F1T	0.603	0.083	0.614	0.085	0.0000	0.743
	F2T	0.428	0.085	0.569	0.113	0.0000	0.804
	F1F2	-0.162	0.060	-0.144	0.054	0.0097	0.825

Table 4

Regression results – Experiment 3

	ß	St. Err. ß	B	St. Err. of B	*p*
F1T	.57	.20	.59	.21	.011
F2T	.42	.16	.53	.20	.016
F1F2	-.45	.19	-.36	.16	.033

List of Figures

Figure 1 Hierarchical categories

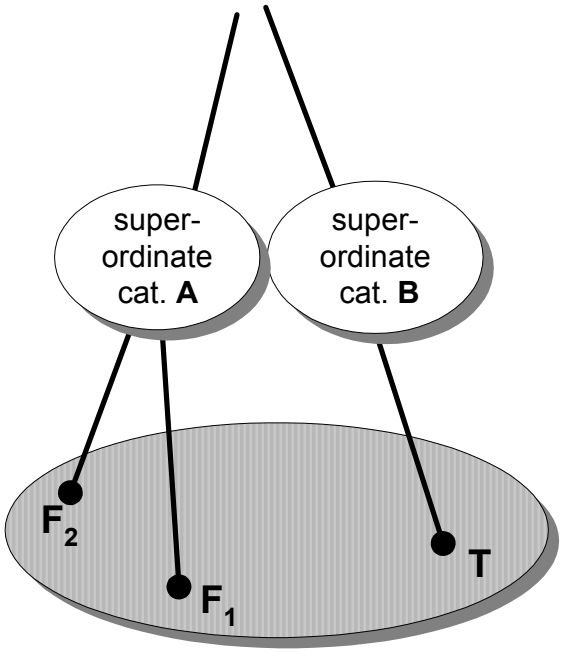


Figure 2 Complex inference and single premise inferences associated with it.

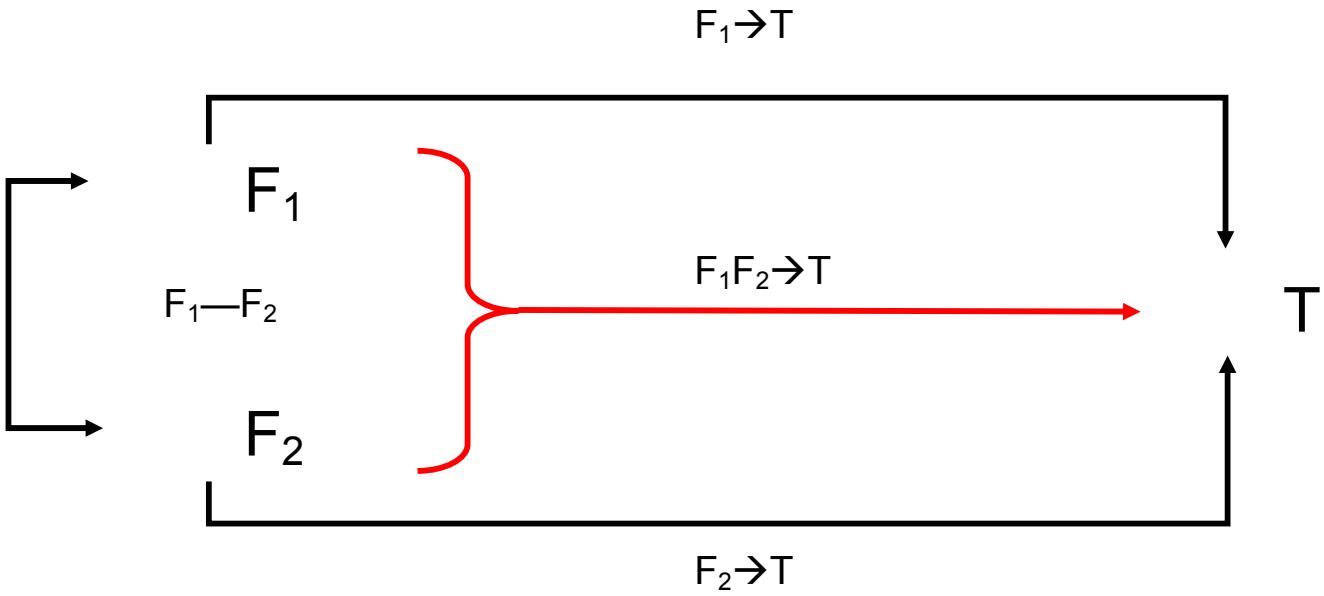


Figure 3 MDS of judged similarity between mammals, reptiles, and insects.

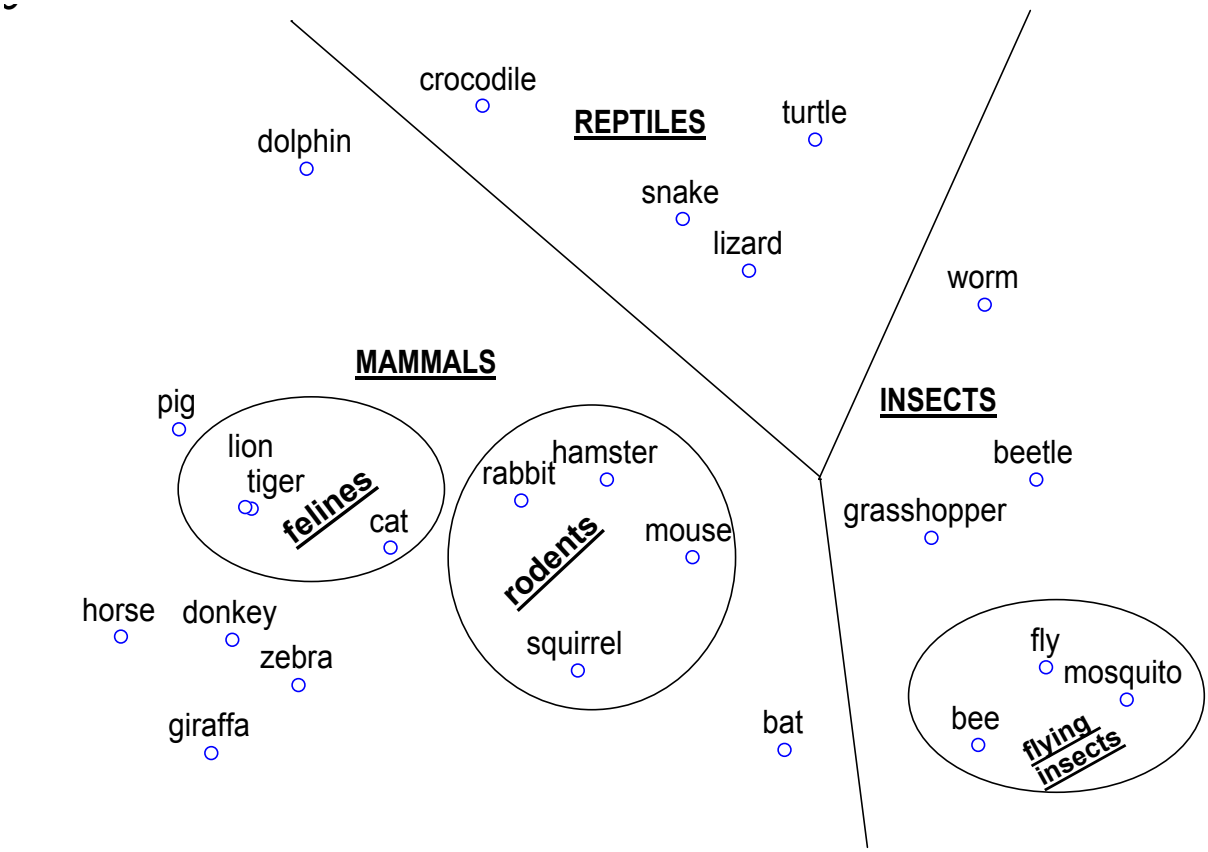


Figure 4 Taxonomy of the judged similarity between birds and animals from other categories.

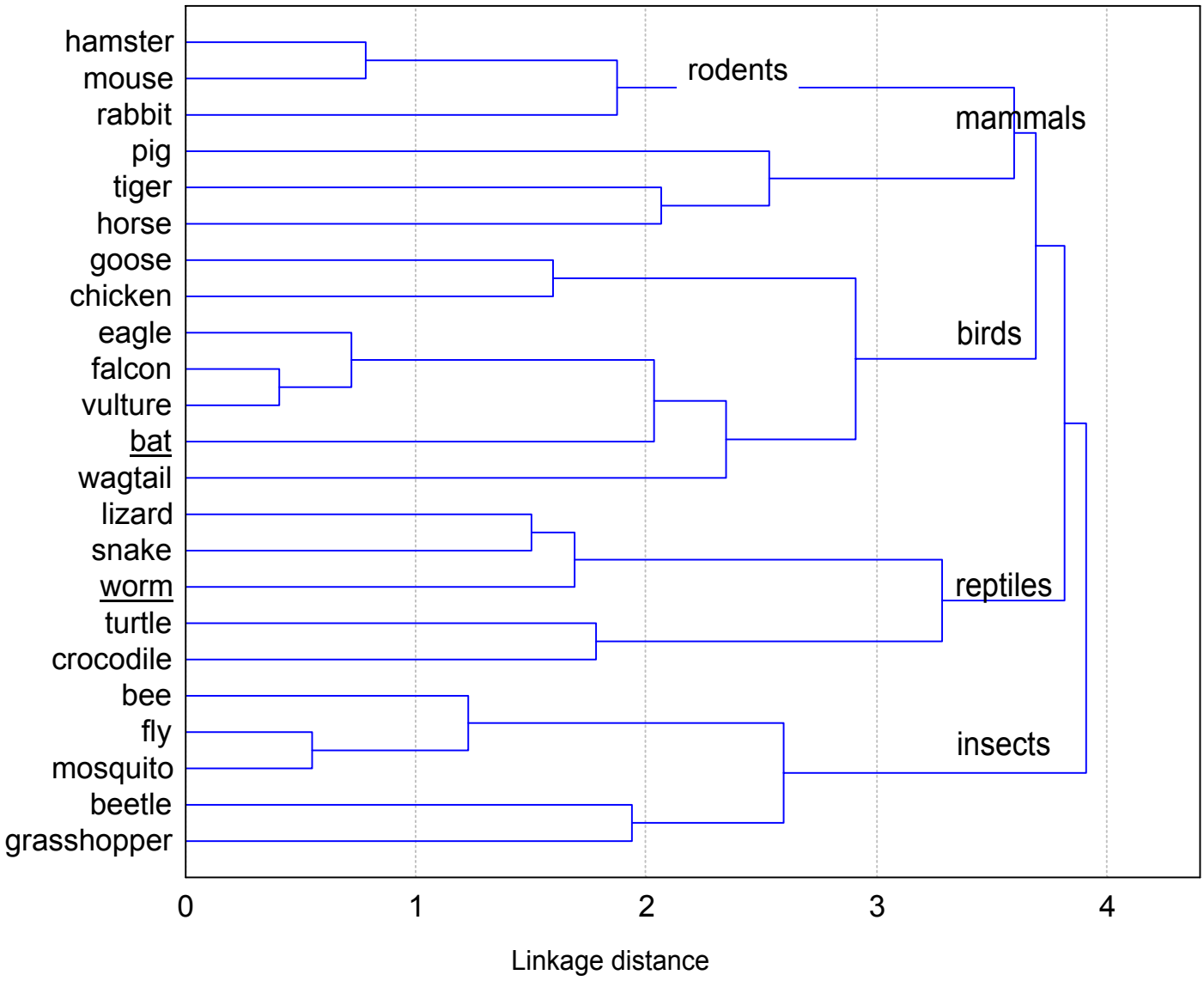
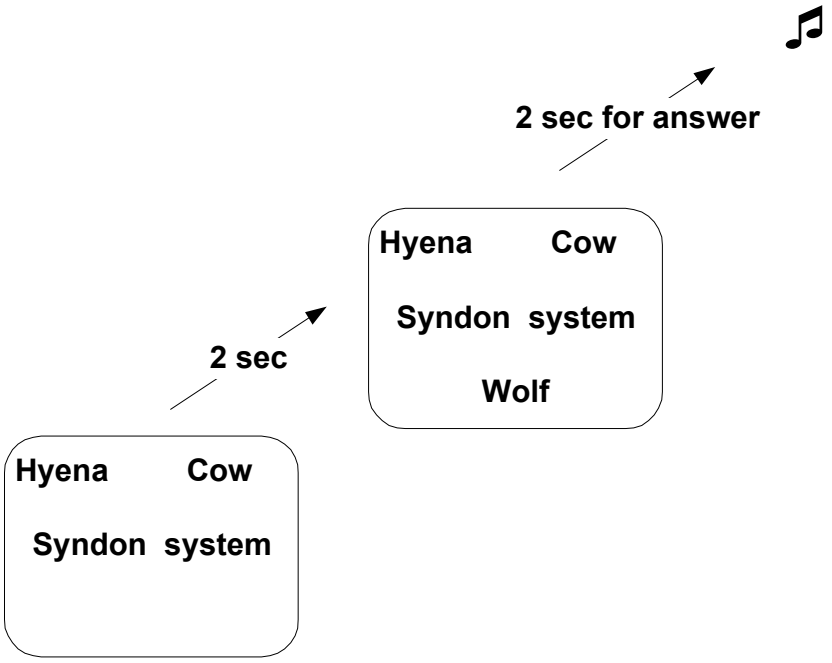


Figure 5 (a) Fast and (b) Slow pace procedure

(a)



(b)

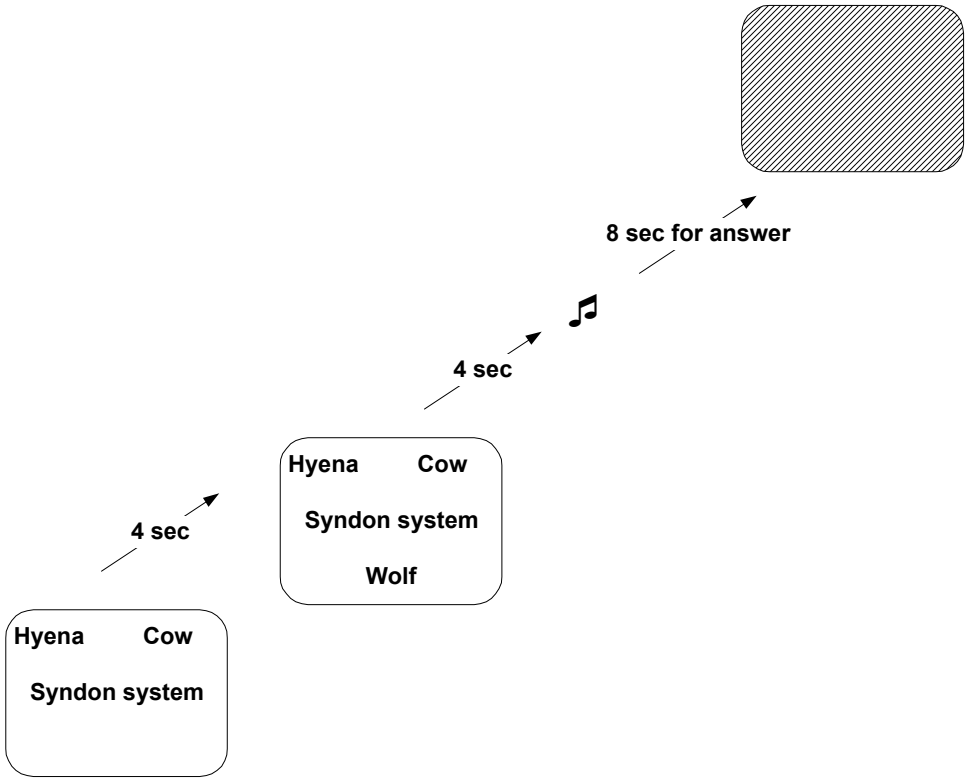


Figure 6 Hierarchical Clustering of languages by projectibility (Complete Linkage)

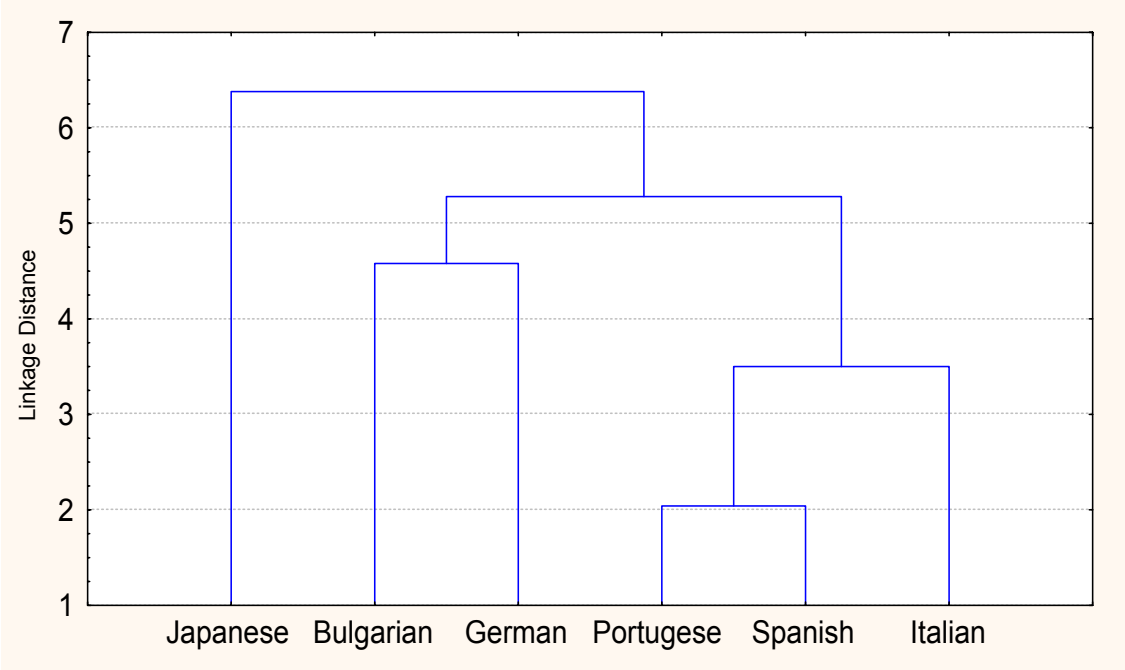


Figure 7 Segregation effect in AAB cases

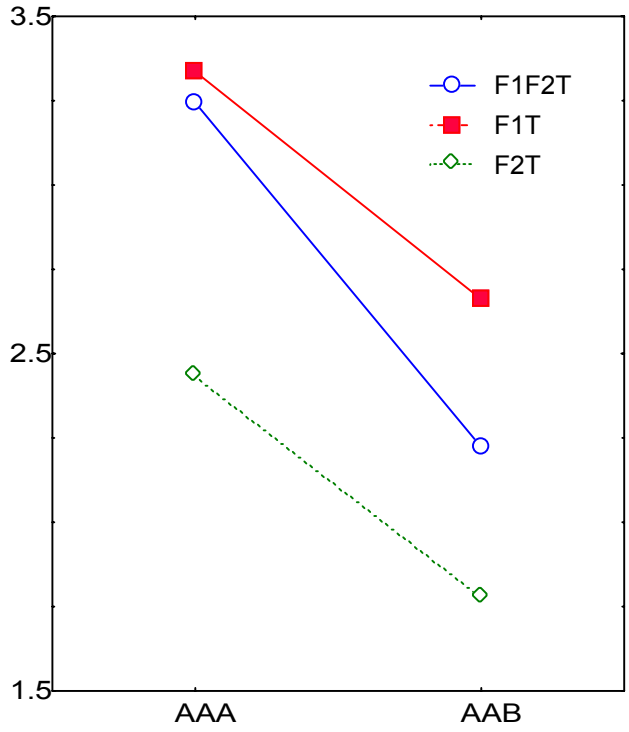


Figure 8 segregation effects among AAA cases

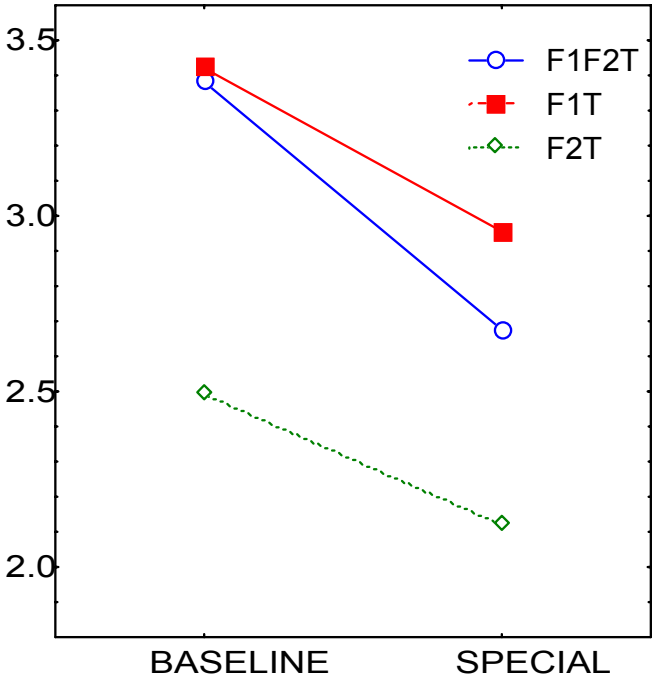
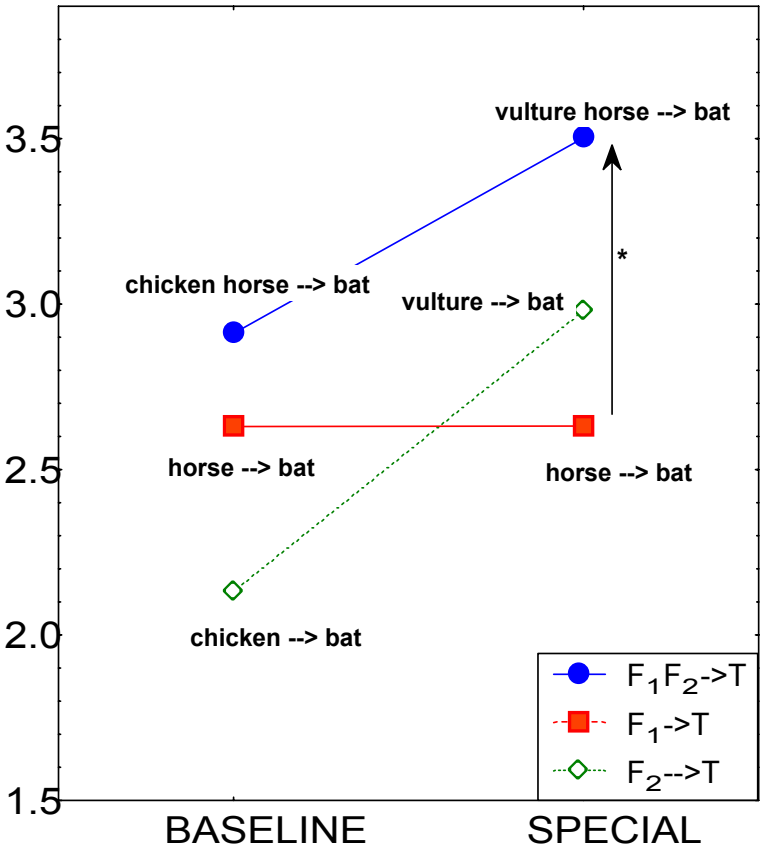


Figure 9 Similarity and improper category



Author Note

Please address correspondence to David Leiser, Dept. of Behavioral Sciences, Ben-Gurion University, PO Box 653, Beer Sheva 84105, Israel. dleiser@bgumail.bgu.ac.il. We gratefully acknowledge the contributions of Jacques Lautrey to the design of Experiment 2, those of Ravid Bogger, Dganit Barnea, Idit Lev and Ravit Benvenisti, and the helpful comments by Evan Heit and Michael Masson.

Footnotes

[1] Categorical arguments like: (*Hyenas have P) AND (Cows have P), HENCE (Wolves have P)* will be notated in the format (Hyenas Cows → Wolves). We will not specify the property involved, for reasons that will become clear below. Further, we will use the same format to refer to *questions* based on those arguments, as for instance: "*If Hyenas have P and Cows have P, how likely is it that Wolves have P too*?", and rely on context to disambiguate.

[2] The formulation is idiomatic in the original Hebrew.

[3] The predicates are not altogether meaningless, since they all contain a meaningful biological general term (e.g., tissue, system) in addition to some proper name. We felt that this would be more natural than asking inferences with totally meaningless terms, such as "*If Horses have SPI, how likely is it that Zebras do too*". Still, the properties, while blank, do imply anatomical or histological features, and it is known that this promotes certain inferences, different from those promoted by predicates that suggest a behavioral trait (Heit and Rubinstein, 1994).

[4] A last case, ABC (the two premises and the conclusion belong to three different higher order categories) was included, but was combined with the AAA case, since there was no theoretical reason to treat it differently.

[5] It is noteworthy that the size of F_1-F_2 component of the regression equation was larger when the pace was slower, and larger yet in Experiment 3, when subjects answered a questionnaire leisurely.