

Semantic congruity in symbolic comparisons: Evidence against an expectancy hypothesis

KEITH J. HOLYOAK and WESLEY A. MAH
University of Michigan, Ann Arbor, Michigan 48104

A previously proposed expectancy hypothesis predicts that the semantic congruity effect typically observed in symbolic comparison tasks will be eliminated when the items to be compared are presented prior to the comparative. Previous studies testing this hypothesis have been inconclusive. The present experiments demonstrate that a semantic congruity effect can be obtained when the comparative follows the pair after a 1-sec (Experiment 1) or 4-sec (Experiment 2) delay. Lexical markedness effects were also obtained. The key to producing a "comparative-after" congruity effect is to intermix questions about several different dimensions, so that when the pair is presented, subjects will be unable to anticipate the question. The results disconfirm the expectancy hypothesis.

Studies of mental magnitude comparisons have yielded two robust empirical findings. First, reaction time (RT) to choose the greater or lesser of two stimuli decreases as the magnitude difference between the stimuli increases. For example, subjects can choose the larger item more quickly for the pair "horse-mouse" than for the pair "horse-deer." This general result is termed the "symbolic distance effect." Second, time to compare items nearer one or the other extreme of the magnitude continuum varies with the form of the comparative adjective. RT is relatively fast with the "lesser" comparative for low-magnitude items, whereas it is relatively fast with the "greater" comparative for high-magnitude items. For example, subjects can more quickly select the smaller of "mouse" and "flea," whereas they can more quickly select the larger of "hippo" and "moose." This interaction between the form of the comparative and the scale position of the items is termed the "semantic congruity effect." Distance and congruity effects have been obtained in numerous studies (e.g., Banks, Fujii, & Kayra-Stuart, 1976; Holyoak & Walker, 1976; Jamieson & Petrusic, 1975; Kosslyn, Murphy, Bemdeserfer, & Feinstein, 1977; for reviews, see Banks, 1977; Moyer & Dumais, 1978).

A number of alternative models have been proposed to account for the process of symbolic magnitude comparison. One general view is that symbolic magnitude is coded in memory by values with interval-scale properties and that these analog magnitude values are

compared to make speeded relative judgments (Moyer & Landauer, 1967). The more discriminable the internal values, the faster a decision will be made, producing a distance effect. The magnitude representation has been variously characterized as a form of mental imagery (Paivio, 1975, 1978) and as a more abstract analog code (Buckley & Gillman, 1974; Holyoak, 1977). But in any case, the semantic congruity effect poses a challenge to analog models. Presumably, the magnitude codes stored in semantic memory should be independent of the form of the question asked the subject; yet the congruity effect involves a change in the pattern of RTs across pairs as a function of the particular comparative used. The semantic congruity effect has in fact been used as the central evidence for a model of the comparison process based on discrete linguistic codes (Banks et al., 1976).

It is possible, however, to account for the congruity effect while preserving the basic assumption that analog magnitude values are compared. One general approach is to assume that the comparative modifies the decision process by altering an implicit reference point (Holyoak, 1978; Jamieson & Petrusic, 1975) or the magnitude range within which the subject is maximally sensitive to differences (Kosslyn et al., 1977). An alternative explanation of the congruity effect, which will be the focus of the present paper, involves an expectancy hypothesis.

The expectancy hypothesis has been considered and rejected by Banks and Flora (1977) and later elaborated and supported by Marschark and Paivio (1979). Marschark and Paivio state the hypothesis as follows: "The congruity effect is assumed to depend simply upon the fact that, in the usual symbolic comparison task, the comparative is presented prior to the stimulus pair, creating a situation wherein the subject may prepare in some way for the range of stimuli (e.g., small or large items) that will be presented" (1979,

This research was supported by NIMH Grant 5-R01-MH33278-02 to Keith Holyoak. We thank Stephen Kosslyn for providing us with the critical idea of intermixing questions about multiple dimensions. We also thank Roger Dominowski for a helpful review of an earlier draft. Shari Roll assisted in testing subjects. Reprint requests should be sent to K. Holyoak at the University of Michigan, Human Performance Center, 330 Packard Road, Ann Arbor, Michigan 48104.

p. 176). Decisions will be facilitated if the expectancy proves correct (i.e., when the comparative "primes" the magnitude range in which the stimuli actually lie) and hindered if the expectancy proves incorrect. The result over a series of trials will be a congruity effect.

As Marschark and Paivio (1979) acknowledge, a variety of alternative mechanisms could produce what might be characterized as effects of the question on subjects' expectancies about the items. For example, an influence of the question on an internal reference point might be described in terms of changes in expectancies. However, the expectancy hypothesis advanced by Marschark and Paivio is made concrete by an explicit prediction: "Regardless of the nature of the hypothesized preparation state, the result is that the comparative creates an expectancy that would not be induced if the presentation order were reversed so that subjects viewed the stimuli prior to the comparative or viewed both simultaneously" (1979, p. 176). For the purpose of the present paper, we will treat the above prediction as an operational definition of the "expectancy hypothesis." The theoretical claim appears to be that the effect of the comparative is limited to the initial encoding of the stimuli. Once the stimuli have been encoded, the comparative will not alter the subsequent comparison process. For example, the comparative "smaller" might differentially facilitate initial lexical access for names of relatively small objects (cf. Meyer & Schvaneveldt, 1971, 1976). But if the names are read and comprehended prior to presentation of the comparative, the congruity effect should be eliminated.

The expectancy hypothesis espoused by Marschark and Paivio (1979) thus predicts that a congruity effect will be obtained when the comparative is presented prior to the stimulus pair, but not when it is presented simultaneously with or after the pair. A second prediction can also be derived from the hypothesis. For many pairs of relative adjectives, such as "small" and "large," one member (e.g., small) is "marked," whereas the other (e.g., large) is "unmarked" (Clark, 1969). In judgment tasks involving such adjective pairs, RT is often faster overall for the unmarked adjective (Clark, 1969; Clark, Carpenter, & Just, 1973). While the precise mechanism underlying the markedness effect is yet to be determined, it seems that in a general sense marked adjectives simply require more time to be comprehended. It therefore follows, as Marschark and Paivio point out, that the markedness effect should be eliminated if the comparative is presented before the stimulus pair, so that it can be comprehended prior to the start of RT measurement. In the context of the expectancy hypothesis, this implies that the markedness and congruity effects should be mutually exclusive: A markedness effect will be obtained only when the comparative is presented simultaneously with or after the pair, whereas a congruity effect will be obtained only when the comparative is presented prior to the pair.

The literature reviewed by Marschark and Paivio (1979), as well as their own experiments, supports the predicted pattern of occurrence for the markedness

effect. The more critical prediction for the expectancy hypothesis concerns the conditions necessary for obtaining a congruity effect. In most comparative judgment experiments, trials have been blocked by form of the comparative, creating a "comparative-first" situation. As predicted, congruity effects are reliably obtained in such experiments. The controversial issue, however, is whether or not a congruity effect can be obtained with simultaneous or "comparative-second" presentation, and here the evidence is less clear. Marschark and Paivio report two experiments involving "smaller" and "larger" judgments for pairs of animal and object names, in which the comparative and the items were presented simultaneously. In one experiment (Marschark & Paivio, 1979, Experiment 3), the comparative was presented either above or below the stimulus pair. A nonsignificant 27-msec congruity effect was obtained.¹ Another experiment (Experiment 4) required subjects to make true-false decisions for sentences (e.g., A goat is larger/smaller than a dog). The overall congruity effect in the latter experiment was a nonsignificant 18 msec.

One should be cautious, of course, in accepting the null hypothesis in the face of trends toward a congruity effect. In fact, an unpublished experiment by Clark, Glass, and Holyoak (described by Holyoak, 1976, Chapter 6) obtained a significant congruity effect in a true-false sentence verification task. Their item set consisted of 192 sentences and included pairs of comparative adjectives for eight different dimensions. Half of the pairs for each dimension were high in absolute magnitude (e.g., Oceans are deeper/shallower than lakes), and half were low in magnitude (e.g., Ponds are deeper/shallower than puddles). Both a significant markedness effect (217 msec in magnitude) and a significant congruity effect (115 msec in magnitude) were obtained. Furthermore, both effects could be statistically generalized across both subjects and dimensions (Clark, 1973). It appears that the pairs used by Clark et al. (cited in Holyoak, 1976) tended to be more extreme in magnitude than those used in Marschark and Paivio's (1979) Experiment 4, which may account for the greater magnitude and reliability of the congruity effect obtained in the former study.

However, as Marschark and Paivio (1979) point out, a significant congruity effect obtained with simultaneous presentation need not lead to unequivocal rejection of the expectancy hypothesis. It is possible that subjects may adopt a strategy of processing the comparative first or at least that encoding of the items may not be completed prior to processing of the comparative. It is therefore conceivable that an expectancy produced by the comparative could influence encoding of the pair, even though presentation of the comparative and presentation of the pair are nominally simultaneous. The critical prediction of the expectancy hypothesis, then, is that the congruity effect will be eliminated if subjects are given time to encode the pair prior to presentation of the comparative.

Banks and Flora (1977) reported three experiments

in which the comparative was presented after the stimulus pair (following either a 500-msec or a 2-sec delay). While congruity effects were found in each experiment, they were quite small in magnitude (typically about 30 msec); furthermore, Marschark and Paivio (1979) have argued that their statistical reliability is not clear from the reported analyses. The largest "comparative-after" congruity effect obtained in the Banks and Flora study was produced when the stimuli were pictures and subjects were asked "to avoid deciding which picture denoted the larger or smaller object while waiting for the instructions" (1977, p. 286). Marschark and Paivio suggest that such "refrain" instructions may have led subjects to avoid fully encoding the items until after the comparative was presented. Under such circumstances, a congruity effect is not inconsistent with the expectancy hypothesis. Marschark and Paivio report an experiment (Experiment 2) in which the comparative ("larger" or "smaller") followed presentation of the pair (either words or pictures) after a 3-sec delay, without "refrain" instructions. They obtained a nonsignificant reversal (-11 msec) of the congruity effect.

While there have thus usually been trends toward congruity effects in comparative-after experiments, these trends have been small (smaller than congruity effects obtained in comparative-first experiments) and statistically weak. The critical prediction of the expectancy hypothesis, while not unambiguously supported, has certainly not been unambiguously disconfirmed. However, the comparative-after studies of both Banks and Flora (1977) and Marschark and Paivio (1979) suffer from a serious flaw. As soon as the pair is presented in such experiments, subjects may compare the relative magnitude of the two items and prepare the correct response for both potential questions. The subject can then simply read the presented comparative and make the appropriate response. This strategy is particularly likely to be used when the delay between the pair and the comparative is long. As noted above, Marschark and Paivio used a 3-sec delay. Not surprisingly, the strategy outlined above was the most frequent one reported by their subjects in a postexperimental questionnaire. In addition, Banks and Flora (1977, Experiment 5) found that the distance effect was greatly attenuated in their comparative-after task, supporting the possibility that subjects often completed the comparison process prior to presentation of the comparative.

If subjects have already completed the comparison process before the comparative is presented and the locus of the congruity effect is in the comparison process, that effect will necessarily be eliminated. The residual trends obtained in the studies reviewed above may simply indicate that on some proportion of the trials, the subjects were unable to complete the comparison process during the delay. Note, however, that elimination of the congruity effect as a result of the above strategy does not constitute support for the

expectancy hypothesis. That hypothesis suggested that the comparative influences initial encoding of the items, but not the actual comparison process. This claim is not being tested if subjects are allowed to complete the comparison process before the comparative is presented.

What is needed, then, is a comparative-after paradigm in which subjects are allowed to encode the pair during the delay but are nonetheless prevented from making a relative judgment until the comparative is presented. The two experiments reported below provide such a paradigm. The basic idea is to increase subjects' uncertainty about the question, so that they will be unable to effectively prepare answers to all possible questions during the delay interval. This was accomplished by intermixing possible comparatives, producing eight potential questions that a subject might be asked on a given trial. A delay long enough for the subject to encode the pair would therefore not likely be sufficient to allow completion of all possible relative judgments. This paradigm thus allows a critical test of the expectancy hypothesis. If a congruity effect can now be obtained in a comparative-after situation, the hypothesis can be rejected.

EXPERIMENT 1

Subjects were timed as they judged the relative magnitudes of pairs of animal names, which were displayed visually for 1 sec prior to presentation of the comparative. The first step in selecting word pairs was to collect systematic magnitude norms for individual words along four different dimensions.

Method

Size norms. Eighty animal names were selected for inclusion in the item pool. Four sets of booklets were constructed. The booklets in each set listed all 80 words in a different random order. The instructions for each booklet set asked subjects to rate the subjective magnitude of typical instances of each animal on a 9-point scale, with a rating of 9 indicating maximum magnitude. A different dimension of magnitude was specified for each form of the booklets. The dimensions used were fierceness, intelligence, size, and speed. Twenty-five University of Michigan undergraduates served as paid subjects. Each subject was tested individually and completed all four sets of ratings. The order in which the booklets were presented was counterbalanced across subjects.

Magnitude norms for each dimension were then derived by successive interval scaling (Bock & Jones, 1968). This method provides a simultaneous normalization of the distribution of responses to each item across the nine response categories. Parameters for the boundaries between response categories and for item means on the resulting scale were estimated using a least-squares technique. The correlation between the successive interval values and the mean ratings for each item was $r = .99$ for each of the four dimensions. Table 1 presents the successive interval values for the 80 animal names on all four dimensions. The values for each dimension are normalized to range from 0 to 10.

Pair selection. A set of 128 pairs of animal names were selected. Thirty-two pairs were used to test each of the four dimensions. For each dimension, 24 pairs were critical test

Table 1
Magnitude Scales for 80 Animal Names on Four Dimensions

Animal	Dimension				Animal	Dimension			
	Fierceness	Intelligence	Size	Speed		Fierceness	Intelligence	Size	Speed
Alligator	8.88	3.67	5.46	5.03	Horse	4.85	5.22	6.66	7.97
Anteater	4.93	3.86	4.46	4.55	Human	7.93	10.00	5.55	5.55
Antelope	4.39	4.42	5.77	9.36	Hummingbird	2.37	3.08	2.03	7.36
Armadillo	4.62	3.17	3.91	4.00	Kangaroo	5.41	4.70	5.89	7.58
Badger	6.93	4.39	4.04	4.99	Leopard	9.45	5.67	5.74	9.40
Bat	5.38	3.30	2.63	7.05	Lion	9.16	5.72	6.10	7.92
Beaver	5.24	5.07	3.80	5.14	Lobster	4.47	1.50	2.88	3.13
Boar	8.38	4.32	5.11	6.18	Mole	3.83	3.08	3.03	4.52
Bobcat	8.17	5.11	4.63	8.09	Moose	5.98	3.96	7.04	6.29
Buffalo	6.59	3.86	7.14	6.28	Mouse	3.08	3.34	2.41	5.02
Bull	7.76	3.80	6.95	6.37	Octopus	5.53	3.58	4.66	4.99
Camel	5.37	4.31	6.96	6.12	Ostrich	4.81	2.97	5.69	7.21
Canary	2.24	3.49	2.33	6.38	Otter	4.84	5.37	4.07	6.58
Cat	5.16	5.85	3.60	6.22	Owl	5.09	4.39	3.45	5.88
Chicken	3.86	2.82	3.51	4.06	Ox	5.87	3.49	6.84	4.74
Chimpanzee	5.28	7.63	4.67	6.05	Panda	5.11	4.51	5.77	4.82
Chipmunk	3.62	3.72	2.62	5.88	Penguin	3.75	3.85	4.01	4.10
Cow	3.95	3.35	6.52	4.59	Pig	5.21	4.39	4.58	4.96
Coyote	7.52	5.72	4.90	7.11	Polar Bear	8.69	5.01	6.95	6.16
Crab	4.39	1.98	2.74	3.41	Rabbit	2.79	3.49	3.48	7.06
Deer	3.91	4.61	5.73	7.82	Raccoon	5.41	4.82	3.70	5.52
Dinosaur	6.66	2.28	9.31	4.57	Rat	5.43	3.85	2.96	5.16
Dog	5.78	6.12	4.41	6.40	Rhinoceros	8.45	3.69	7.67	6.24
Dolphin	4.11	7.48	5.74	7.66	Seagull	3.77	3.62	3.45	7.04
Donkey	5.20	4.80	5.98	5.87	Seal	4.58	5.53	5.23	6.23
Dove	2.33	3.45	2.82	6.37	Shark	10.00	4.02	6.64	7.93
Duck	3.46	3.42	3.52	5.41	Sheep	3.33	3.26	4.76	5.20
Eagle	7.13	4.84	3.89	8.25	Snail	1.29	.24	1.47	0.00
Elephant	6.75	4.92	8.42	5.37	Sparrow	2.35	2.96	2.52	7.04
Flea	2.52	.24	0.00	3.65	Squirrel	4.15	4.10	3.16	6.37
Fly	3.00	.62	.79	5.70	Tarantula	6.26	1.80	2.11	3.65
Fox	6.14	5.71	4.05	6.79	Tiger	9.57	5.57	6.05	8.46
Frog	2.05	1.85	2.28	4.14	Toad	2.54	1.83	2.18	3.74
Gazelle	4.08	4.49	5.77	10.00	Turkey	4.20	2.30	3.79	4.50
Giraffe	4.43	4.17	7.70	7.55	Walrus	5.89	4.00	6.59	4.59
Goat	5.03	4.52	4.77	5.61	Weasel	5.65	4.93	3.69	6.56
Goldfish	1.35	1.45	1.91	4.18	Whale	5.18	4.98	10.00	6.58
Gopher	3.98	4.17	3.07	5.79	Wolf	8.05	6.12	5.00	7.18
Gorilla	7.74	6.39	6.61	5.42	Worm	0.00	0.00	1.31	.90
Hippopotamus	6.02	3.54	7.55	4.43	Zebra	5.00	4.61	6.12	7.75

items and the remaining 8 were filler items. Half of the 24 critical pairs consisted of low-magnitude words referring to animals (i.e., animal names with magnitude values less than approximately 5.0), and half consisted of high-magnitude words. Magnitude difference was not manipulated systematically; in general, pairs were selected to be relatively close in magnitude, but sufficiently separated so that subjects could be expected to reach an unambiguous decision. Table 2 presents examples and a description of the critical test items for each condition.

Each of the 79 animal names used to construct test items was used in only one pair for any particular dimension; however, 77 of the words were included in pairs for at least two dimensions ("human" was the only word in Table 1 not used at all). Pairs were constructed so that across different dimensions, most words (56 of the 79) occurred as both the "greater" and the "lesser" member in different pairs. The eight filler pairs constructed for each dimension were used to help achieve the above distribution of words; the fillers also included some pairs containing items drawn from the middle range of the dimensions.

Relative judgment task. The stimuli were presented on TV screens controlled by an IBM 1800 computer. The sequence of events on each trial was as follows. A fixation cross was displayed on the center of the screen for 500 msec. It was then replaced by

the two words to be compared, which appeared on each side of the position previously occupied by the fixation cross. After a delay of 1 sec, one of eight possible comparatives (fiercer-meeker, smarter-dumber, larger-smaller, faster-slower) appeared centered below the word pair, initiating RT measurement. The pair and the comparative then remained on until the subject responded by pressing either a left or a right decision key to indicate which word was the correct response. If 4 sec elapsed without a response, an error RT of 4 sec was recorded. If the subject made an error, the word ERROR was displayed for 1 sec. The next trial then began automatically.

The 128 test trials were presented in a different random order to each subject. Each subject saw one of four versions of the test items. Which of the two possible questions was given for each pair and whether the left or right word was the correct response were counterbalanced across the different versions. Each subject thus saw any one pair only once, but across subjects, each pair was tested with both possible comparatives.

The session began with 24 practice items, using pairs that were not among the 128 test items. Subjects were told to base their judgments on typical examples of the animals in each pair; they did not receive any more specific instructions about how to perform the task. They were told to respond as quickly as

Table 2
Description and Examples of Word Pairs Used in Experiments 1 and 2

Mag- ni- tude	Dimension											
	Fierceness (Fiercer-Meeker)			Intelligence (Smarter-Dumber)			Size (Larger-Smaller)			Speed (Faster-Slower)		
	Example	M	MD	Example	M	MD	Example	M	MD	Example	M	MD
Low	Sheep-Goldfish	3.35	1.92	Crab-Worm	2.45	1.97	Squirrel-Canary	2.98	1.28	Goat-Toad	4.32	1.75
High	Leopard-Elephant	7.44	2.40	Gorilla-Zebra	5.35	1.55	Elephant-Hippo	6.11	1.25	Bobcat-Moose	7.45	1.77

Note—M = mean magnitude; MD = mean magnitude difference.

possible but to avoid errors. The entire session lasted about 30 min. Afterward, subjects were asked to what extent they felt they had been able to anticipate the question during the delay.

Twelve University of Michigan undergraduates served as paid subjects.

Results and Discussion

Separate analyses of variance on the RT data were performed, treating subjects and items (i.e., word pairs) as random effects, and minimum quasi-F ratios (min F's) were calculated (Clark, 1973). For this purpose, the data were collapsed over the four counterbalancing conditions and left-right pair order. Since different subjects provided the means for any particular pair-comparative combination, item means were based on standard scores calculated for each subject. This procedure avoids allowing subject variance to inflate estimates of the variance associated with item interactions. Analyses were performed using either all RTs or RTs for correct decisions only. Essentially the same pattern of results was obtained in both sets of analyses; only the results for correct responses will be reported here.

The expectancy hypothesis predicts that a markedness effect will be obtained when the comparative follows the word pair, but that a congruity effect will not be observed. Contrary to this prediction, both of these effects were obtained in Experiment 1. Judgments involving unmarked comparatives (fiercer, smarter, bigger, faster) were made 104 msec more quickly than were judgments involving marked comparatives (meeker, dumber, smaller, slower) [$\min F(1,59) = 5.36, p < .025$]. In addition, subjects responded 116 msec more quickly when the magnitude of the pair was congruent with the comparative, rather than incongruent [$\min F(1,33) = 6.08, p < .025$].² Neither of these effects varied significantly across the four dimensions used in the experiment. As Figure 1 illustrates, the overall markedness effect combined with the congruity effect to create a funnel-shaped crossover interaction, with a significant difference between the two comparatives for the high-magnitude pairs.

The only other significant effect involved variations in the overall RT to the four dimensions [$\min F(3,101) = 2.93, p < .05$]. Overall RT and error rates were ordered in the same way across the dimensions: size, 1,331 msec (9.4%); fierceness, 1,449 msec (14.9%); intelligence, 1,507 msec (20.1%); speed, 1,528 msec

(20.5%). The import of these dimension differences is not entirely clear, but it seems likely that people have more precise knowledge of magnitude for some dimensions, such as size, than for others. Postexperimental questioning of the subjects suggested that the rather high overall error rate in part reflected disagreements as to the correct response. No subject reported being able to anticipate the question on more than a small fraction of the trials.

EXPERIMENT 2

The results of Experiment 1 indicated that, contrary to the prediction of the expectancy hypothesis, a congruity effect can be obtained when the comparative follows the question. Experiment 2 was performed to provide a replication of this critical result and also to determine whether instructions to use imagery might influence the comparison process.

Method

The materials and procedure were identical to those used in Experiment 1, with two exceptions. First, half the subjects were instructed to use imagery to perform the task, and half (like

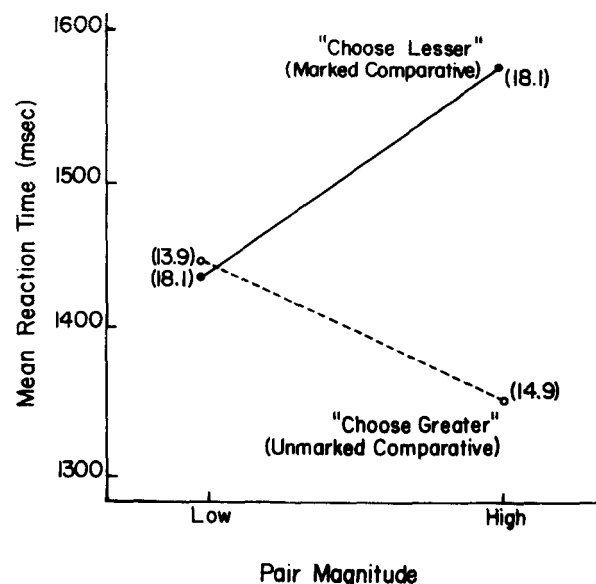


Figure 1. Reaction time to choose the greater or lesser of two animal names as a function of magnitude of the pair (Experiment 1). Error rates are given in parentheses.

subjects in Experiment 1) received no special instructions. Subjects receiving imagery instructions were told that when the word pair appeared, they were to form images of typical instances of each animal, side by side. Then, when the question appeared, they were to use whatever information they could extract from the images to help make a decision. Examples were given to illustrate potentially relevant information in images. Subjects were told that relative size could be simply "read off" the images and that various visual properties might predict the correct response for the other dimensions (e.g., sharp teeth might indicate ferocity, long legs might indicate speed, and certain facial features might indicate intelligence). Subjects were told, however, to make the response they considered correct even if their images suggested the opposite response. One could argue, of course, that it is rather unnatural to base decisions about an abstract dimension such as intelligence on information in visual images. However, all subjects indicated that they understood the instructions, and none expressed any doubts about being able to follow them.

The second change introduced in Experiment 2 was to increase the delay between presentation of the pair and presentation of the comparative to 4 sec. This increase was intended to allow subjects given imagery instructions sufficient time to construct images of the two animals. A 4-sec delay is longer than that used in any previous comparative-after experiment. If subjects can use the extra time to anticipate the question, the result could be an attenuation of the congruity effect. At the end of the experimental session, subjects were asked whether they felt they had been able to anticipate the question on some trials. They were also asked whether and how they had used imagery in reaching decisions.

Forty University of Michigan undergraduates served as paid subjects.

Results and Discussion

Analyses of variance were performed in the same manner as for Experiment 1. The results of the earlier experiment were entirely replicated. Decisions were made 80 msec more quickly when the comparative was unmarked than when it was marked [$\min F'(1,125) = 15.5, p < .001$]. In addition, subjects responded 68 msec more quickly when the pair and the comparative were congruent than when they were incongruent [$\min F'(1,118) = 5.92, p < .025$]. Neither the markedness nor the congruity effect varied significantly as a function of either the particular dimension or the instruction condition. As Figure 2 illustrates, the pattern of results was quite similar to that obtained in Experiment 1.

Imagery instructions had no apparent influence on either RTs or error rates. The instruction variable did not interact with any other factor, and overall mean RTs for the two instruction conditions differed by only 2 msec [$\min F' < 1$]. The latter finding suggests that the imagery-instructed subjects may not have used images to make their decisions, since previous work on mental size comparisons indicates that the conscious use of imagery greatly increases RT relative to that of control subjects who do not receive imagery instructions (Holyoak, 1977). Postexperimental interviews indicated that all imagery-instructed subjects claimed to have formed images of the two animals while waiting for the question; however, many said they often did not find the images useful once the question appeared. Some indicated that images helped for the size dimension and, to a lesser extent, for fierceness. Most subjects claimed images were

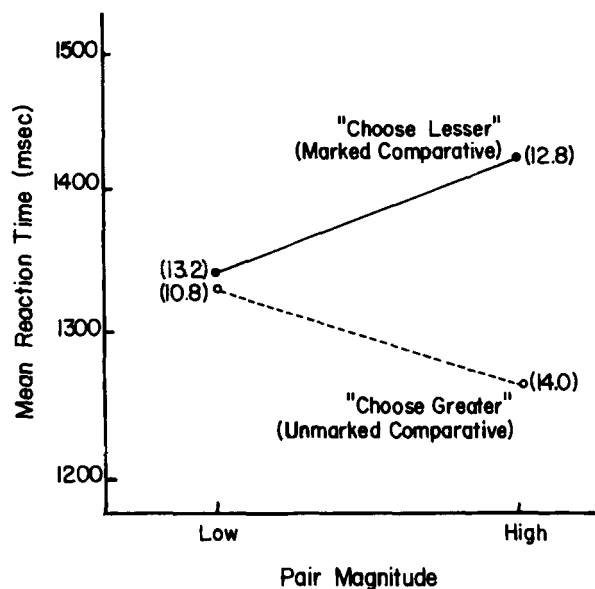


Figure 2. Reaction time to choose the greater or lesser of two animal names as a function of magnitude of the pair (Experiment 2). Error rates are given in parentheses.

least helpful for the intelligence dimension. It should be emphasized, however, that mean RT did not vary significantly as a function of instructions for any of the four dimensions. It is therefore questionable whether our subjects had any real retrospective insight into the role that imagery may have played in their decision making (cf. Nisbett & Wilson, 1977). A plausible interpretation of these data is that imagery-instructed subjects formed images during the delay interval but then based their subsequent decisions largely on nonimaginal information. One could argue, of course, that the null results indicate simply that our uninstructed subjects spontaneously used imagery to perform the task. However, the control subjects reported using imagery either on a relatively small proportion of the trials (usually for the size dimension) or else not at all.

As in Experiment 1, overall RT varied significantly across the four dimensions [$\min F'(3,165) = 10.6, p < .001$]. Furthermore, dimensions were ordered in difficulty precisely as they were in the earlier experiment, both in RT and in error rate: size, 1,165 msec (7.3%); fierceness, 1,364 msec (12.3%); intelligence, 1,398 msec (14.3%); and speed, 1,417 msec (17.0%). These dimension differences should be interpreted cautiously, since the scale units associated with the norms for the four dimensions are not comparable. Nevertheless, it is interesting that the two "easiest" dimensions, size and fierceness, were also those for which subjects most often claimed to sometimes use imagery to make decisions. It is possible that visualization aided the judgment process for these dimensions; but if so, the benefit was independent of whether subjects were instructed to use imagery. The present results are clearly inconclusive with respect to the possible role of imagery in mental comparisons.

It should be noted that in terms of both RT and error

rate, performance in Experiment 2 was superior overall to that in Experiment 1. This may be due to the extra preparation time that subjects received in Experiment 2, prior to presentation of the question. However, subjects reported being able to anticipate the question on at most a few of the trials, and the observed congruity effect clearly indicates that the judgment process was influenced by the comparative despite the delay.

GENERAL DISCUSSION

The present study indicates that both markedness and semantic congruity effects can be obtained in a symbolic comparison task, even when the pair is presented for as long as 4 sec prior to the comparative. The key to producing a robust comparative-after congruity effect is simply to intermix questions about several different dimensions, so that subjects are unable to anticipate the question and, therefore, cannot complete the comparison process before the comparative is presented. These results disconfirm the expectancy hypothesis outlined by Marschark and Paivio (1979). Contrary to that hypothesis, the congruity effect need not be eliminated when the comparative is presented after the items to be compared, nor are the congruity and markedness effects mutually exclusive.

How do the present results relate to possible explanations of the congruity effect? It is useful to view the mental comparison task as being composed of three stages: (1) encoding the items, (2) comparing their relative magnitudes, and (3) executing a response. The present results would seem to disconfirm any model that claims that the locus of the congruity effect is entirely in the initial encoding stage, as the expectancy hypothesis of Marschark and Paivio (1979) apparently does, since in the present experiments, encoding was presumably completed before RT measurement was initiated. However, rather than accepting the above conclusion, a supporter of the expectancy hypothesis might suggest that item encoding may not always be completed prior to presentation of the comparative, even with a delay of 4 sec. As a result, the comparative could still have influenced encoding of the pair in our experiments. The plausibility of this possible attempt at rebuttal clearly hinges on the definition of the term "encoding." If encoding is taken to mean complete activation of all the information associated with a concept, including its values on all possible attribute dimensions (or even the four that might be tested), 4 sec might well be insufficient. Indeed, the consistent overall RT differences between the four dimensions might be considered evidence that not all of the dimensional values were equally activated prior to presentation of the comparative (although this finding clearly has alternative interpretations).

However, such a broad definition of encoding does not appear to be a useful one. In fact, there may well be no amount of time sufficient to guarantee complete and

simultaneous activation of all the attribute information stored with a concept. The present delay paradigm is similar to other semantic verification studies in which one concept is presented in advance of another. These studies have demonstrated that variables that influence verification time (most notably, measures of association strength, such as production frequency) continue to have an effect even after a delay (Holyoak, Glass, & Mah, 1976; Loftus, 1973). Such evidence argues that not all attribute information is retrieved during a delay. Indeed, it may be an oversimplification to suppose that dimensional values are simply retrieved and then compared. Many theorists have suggested that the decision process in mental comparison tasks involves retrieving and comparing magnitude values in an iterative fashion, until a decision criterion is satisfied (Buckley & Gillman, 1974; Holyoak, 1978; Moyer & Dumais, 1978). Under this conception, the process of accessing magnitude values is properly viewed as a component of the comparison process, rather than of initial encoding.

It is certainly the case that the processes of initial perception and lexical look-up, which are typically embraced by the term "encoding," require much less time than the 4 sec allowed in Experiment 2. For example, dual-word lexical decisions (i.e., word vs. nonword judgments with word pairs) are typically completed in approximately 1 sec (Glass, Holyoak, & Kiger, 1979; Meyer & Schvaneveldt, 1971). For that matter, 4 sec is more than ample time to perform the entire process of comparative judgment, if both the pair and the comparative are presented. It therefore seems reasonable to assume that initial encoding of the pairs, in its usual sense, was completed during the 4-sec delay interval. It should also be noted that the elimination of the congruity effect in the comparative-after paradigm is the only prediction that has so far been derived from Marschark and Paivio's (1979) expectancy hypothesis. If this prediction is somehow avoided, it is unclear how the hypothesis can be tested.

In contrast, numerous alternative models, which ascribe the congruity effect to the comparison or response stages, are compatible with the present findings. For example, the comparative might influence a process of sampling magnitude values of the stimuli (Kosslyn et al., 1977), or it might set an internal reference value with respect to which the items are compared (Holyoak, 1978; Jamieson & Petrusic, 1975). As we noted earlier, the latter models might be characterized as involving variations in expectancies, but in a broader sense than was apparently intended by Marschark and Paivio (1979). In terms of the semantic coding model of Banks et al. (1976), the semantic code for the comparative may match or mismatch the codes derived for the stimuli. The above models implicate the comparison stage as the locus of the congruity effect. Yet a further alternative is that conflicting codes for the comparative and stimuli may produce Stroop-like interference, perhaps affecting response execution (Banks & Root,

1979; John, 1975). Holyoak and Mah (Note 1) review a variety of alternative accounts of the congruity effect.

All of the above models, and no doubt others, assume the congruity effect arises in a stage that follows encoding of the items to be compared (comparison or response execution); hence, all are compatible with the comparative-after congruity effect observed in the present study. While these results unambiguously rule out a particular version of an expectancy hypothesis, the actual mechanism or mechanisms underlying the semantic congruity effect are yet to be determined.

REFERENCE NOTES

1. Holyoak, K. J., & Mah, W. A. *Cognitive reference points in judgments of symbolic magnitude*. Manuscript in preparation, 1981.
2. Marschark, M. Personal communication, September 1980.
3. Banks, W. P. Personal communication, April 1979.

REFERENCES

- BANKS, W. P. Encoding and processing of symbolic information in comparative judgments. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 11). New York: Academic Press, 1977.
- BANKS, W. P., & FLORA, J. Semantic and perceptual processes in symbolic comparisons. *Journal of Experimental Psychology: Human Perception and Performance*, 1977, 3, 278-290.
- BANKS, W. P., FUJII, M., & KAYRA-STUART, F. Semantic congruity effects in comparative judgments of magnitudes of digits. *Journal of Experimental Psychology: Human Perception and Performance*, 1976, 2, 435-447.
- BANKS, W. P., & ROOT, M. Semantic congruity effects in judgments of loudness. *Perception & Psychophysics*, 1979, 26, 133-142.
- BOCK, R. D., & JONES, L. V. *The measurement and prediction of judgment and choice*. San Francisco: Holden-Day, 1968.
- BUCKLEY, P. B., & GILLMAN, C. B. Comparisons of digits and dot patterns. *Journal of Experimental Psychology*, 1974, 103, 1131-1136.
- CLARK, H. H. Linguistic processes in deductive reasoning. *Psychological Review*, 1969, 76, 387-404.
- CLARK, H. H. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 1973, 12, 335-359.
- CLARK, H. H., CARPENTER, P. A., & JUST, M. A. On the meeting of semantics and perception. In W. G. Chase (Ed.), *Visual information processing*. New York: Academic Press, 1973.
- GLASS, A. L., HOLYOAK, K. J., & KIGER, J. I. Role of antonymy relations in semantic judgments. *Journal of Experimental Psychology: Human Learning and Memory*, 1979, 5, 598-606.
- HOLYOAK, K. J. *Symbolic processes in mental comparisons*. Unpublished doctoral dissertation, Stanford University, 1976.
- HOLYOAK, K. J. The form of analog size information in memory. *Cognitive Psychology*, 1977, 9, 31-51.
- HOLYOAK, K. J. Comparative judgments with numerical reference points. *Cognitive Psychology*, 1978, 10, 203-243.
- HOLYOAK, K. J., GLASS, A. L., & MAH, W. A. Morphological structure and semantic retrieval. *Journal of Verbal Learning and Verbal Behavior*, 1976, 15, 235-247.
- HOLYOAK, K. J., & WALKER, J. H. Subjective magnitude information in semantic ordering. *Journal of Verbal Learning and Verbal Behavior*, 1976, 15, 287-299.
- JAMIESON, D. G., & PETRUSIC, W. M. Relational judgments with remembered stimuli. *Perception & Psychophysics*, 1975, 18, 373-378.
- JOHN, I. D. A common mechanism mediating the time-order error and the crossover effect in comparative judgments of loudness. *Australian Journal of Psychology*, 1975, 27, 51-60.
- KOSSLYN, S. M., MURPHY, G. L., BEMESDERFER, M. E., & FEINSTEIN, K. J. Category and continuum in mental comparisons. *Journal of Experimental Psychology: General*, 1977, 106, 341-375.
- LOFTUS, E. F. Category dominance, instance dominance, and categorization time. *Journal of Experimental Psychology*, 1973, 97, 70-74.
- MARSCHARK, M., & PAIVIO, A. Semantic congruity and lexical marking in symbolic comparisons: An expectancy hypothesis. *Memory & Cognition*, 1979, 7, 175-184.
- MEYER, D. E., & SCHVANEVELDT, R. W. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 1971, 90, 227-334.
- MEYER, D. E., & SCHVANEVELDT, R. W. Meaning, memory structure, and mental processes. *Science*, 1976, 192, 27-33.
- MOYER, R. S., & DUMAIS, S. T. Mental comparison. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 12). New York: Academic Press, 1978.
- MOYER, R. S., & LANDAUER, T. K. Time required for judgments of numerical inequality. *Nature*, 1967, 215, 1519-1520.
- NISBETT, R. E., & WILSON, T. D. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 1977, 84, 231-259.
- PAIVIO, A. Perceptual comparisons through the mind's eye. *Memory & Cognition*, 1975, 3, 635-647.
- PAIVIO, A. Mental comparisons involving abstract attributes. *Memory & Cognition*, 1978, 6, 199-208.

NOTES

1. This value and others from Marschark and Paivio (1979) were reported erroneously in the original paper. Marschark (Note 2) provided us with the correct values.
2. Similar results have recently been obtained by W. P. Banks (Note 3).

(Received for publication June 20, 1980;
revision accepted September 9, 1980.)