Evidence for implicit scaling in comparative judgment

KEVIN M. SAILOR

Lehman College, The City University of New York, Bronx, New York

and

KIM M. PINEDA

University of Michigan, Ann Arbor, Michigan

In two experiments, we examined the effects of manipulating the density of stimuli on comparison difficulty in a comparative judgment task. In Experiment 1, subjects were slower at judging the relative size of a pair when the members were adjacent items in the linear order than when the members were separated by items of intervening magnitudes. In Experiment 2, the advantage of choosing the larger rather than the smaller of two large stimuli (e.g., the congruity effect) increased when the linear order included many small items. In contrast, the advantage of choosing the smaller of two small items increased when the linear order included many large items. The applicability of the range-frequency theory (Parducci, 1965) to these results is discussed.

One of the more interesting effects of context is that it can qualify the truth of statements about even the most basic properties of objects. For example, the statement "the moon is small" is almost certainly false in the context of statements about everyday objects, but is probably true in the context of statements about other celestial bodies in the solar system.

For some time, researchers have known that context has similar effects on relational judgments (Banks, 1977; Čech & Shoben, 1985). Relational or comparative judgments involve comparing two objects on some specified dimension. For example, elephants are larger than fleas (Čech & Shoben), lions are fiercer than sheep (Holyoak & Mah, 1981), and 2 is less than 5 (Banks, Fujii, & Kayra-Stuart, 1976). These results can be quite striking. For example, Čech and Shoben found that the size and direction of the congruity effect for selected pairs changed as a function of context. The congruity effect is the finding that it is relatively easier to discriminate two objects when the form of the question matches the magnitude of the objects. For example, it is easier to choose the larger rather than the smaller of elephant and rhino, but it is easier to choose the smaller rather than the larger of mouse and rat. Initially, their subjects compared the sizes of animals ranging from that of a flea to an elephant. After one block of trials in this full range, these comparisons were subsequently restricted to a smaller range of items:

from flea to beaver or sheep to elephant. In the full range, Čech and Shoben found that small pairs (such as rabbit-beaver) were discriminated more rapidly under the instruction to select the smaller item, and large pairs were discriminated more rapidly under the instruction to select the larger item. However, in the restricted range, the larger of the small pairs (such as rat-rabbit and rabbit-beaver) were discriminated more readily under the *larger* instruction. Similar findings were observed when the range was restricted so that only large animals were compared. This result is analogous to the effects of context on absolute judgments of size. In the case of absolute judgments, the validity of any statement about the size of an object depends on the size of the object relative to other objects. Apparently, the perceived magnitude of any pair in a comparative judgment task depends on the sizes of other objects. Thus, judgments of the absolute size of an object or the relative size of a pair of objects is not performed in isolation, but depends on the size of contextually relevant objects.

This idea has dominated much of the research on psychophysical judgments. One of the more successful articulations of this idea, the range-frequency theory (Parducci, 1965), was developed primarily as an account of contextual effects on categorical judgments but has been applied to a number of psychophysical tasks as well. According to the range-frequency theory, responses are mediated by two principles: the frequency principle and the range principle. According to the frequency principle, equal portions of the response range tend to be used with equal frequency, whereas the range principle holds that equal portions of the response range are used for equal portions of the subjective stimulus range. For example, Parducci and Haugen (1967) presented subjects with a

Preparation of this article was supported in part by PSC/CUNY Faculty Research Award 661486 to the first author. We wish to thank Claude Čech, Edward Shoben, Allen Parducci, and two anonymous reviewers for their comments and suggestions on a draft of this article. Requests for reprints should be sent to K. Sailor, Department of Psychology, Lehman College, Bronx, New York 10468-1589.

standard followed by an interpolated weight, and then asked them to judge whether or not a third weight was heavier than the original standard. They found that the probability of making a "heavier" judgment for a given weight was greater when the relative frequency with which light weights were presented was greater. This result indicates that the frequency principle impacts both comparative and categorical judgments.

The success of the range-frequency theory at explaining the effects of context in this psychophysical task suggests that this theory may also help to explain the effects of similar contextual manipulations in comparative judgment tasks using symbolic stimuli. The obvious parallel between comparing two physical objects and symbolic comparisons was quickly noted by researchers (Moyer, 1973; Paivio, 1975). Although the exact relationship between psychophysical judgment and symbolic judgment has been vigorously debated (Banks, 1977), most researchers have agreed that the relative magnitude of remembered objects is derived by retrieving and comparing information about the magnitude of these objects. The present paper explores an idea proposed by Holyoak and Mah (1982) that an implicit scaling of this magnitude information occurs when the relative magnitude of two objects is determined. It is the current proposal that the preceding contextual effects are the product of an implicit scaling that is governed by the principles outlined in the range-frequency theory. This proposal was tested by selecting a set of symbolic stimuli (e.g., remembered items rather than the actual physical items) and demonstrating that the effects of relative frequency on absolute judgments can be systematically related to analogous effects in comparative judgment tasks.

EXPERIMENT 1

The primary purpose of this manipulation was to affect the difficulty of judgment by changing the symbolic distance of pairs, which is the difference in the items' magnitudes. A large body of research (see Banks, 1977) has demonstrated that the time it takes to make a judgment about a pair of items decreases as the difference in their magnitudes increases. If the range-frequency hypothesis is correct, then it should be possible to change the symbolic distance of a pair by changing the difference in the ranks of two items. For example, if a number of filler items whose magnitudes fall between *mouse* and *rabbit* are added to a set of stimuli, the range-frequency theory claims that the difference in the assessed magnitudes of *mouse* and *rabbit* will increase. Thus, inserting filler items should increase the symbolic distance of a pair and speed relative judgments about the pair.

In Experiment 1, the rank of stimuli was manipulated by adding interpolated filler items at either end of the stimulus range. In the small-context condition, small filler items were added to the set of stimuli. This addition should make the judgment of small experimental items in this con-

dition easier by increasing differences in the rank of the smaller experimental stimuli. Similarly, the addition of large filler items in the large-context condition should make the judgment of large experimental pairs easier. Thus, small pairs should be easier to judge in the small-context condition, and large pairs should be easier to judge in the large-context condition.

This experiment differs from previous studies (Banks, 1977; Čech & Shoben, 1985)¹ in that earlier confounds between the manipulation of rank and stimulus range were removed. For example, Banks compared identical pairs across a full-range context and two restricted-range contexts. In addition to a critical increase in the ordinal separation of these test pairs, there was also a decrease in the range of stimulus values in the restricted-range contexts relative to the full-range context. Although it is tempting to conclude that the advantage of the test pairs in the restricted range was produced by the increase in the ordinal separation of these items, the decrease in the stimulus range may also have been responsible for the advantage of these pairs in the restricted-range contexts. According to the range principle, a decrease in the stimulus range should enhance discrimination because each response or magnitude category must cover a smaller range of stimuli. It is quite possible that discrimination of test pairs may have been enhanced in the restricted-range contexts because two stimuli that occupied a similar magnitude category in the full-range context were assigned to different magnitude categories in the restricted-range contexts. Thus, previous manipulations have not made it possible to determine the extent to which the distribution of stimulus values has an effect on comparative judgment.

Method

Subjects. Of the 27 subjects who participated in the large-context condition, 4 subjects failed to make an error criterion of 48 or fewer errors. In the small-context condition, 3 of the 27 subjects who participated failed to meet this error criterion. All of the subjects were right-handed undergraduates of the University of Illinois who participated for course credit.

Materials. The materials were selected from the Shoben, Čech, Schwanenflugel, and Sailor (1989) norms, which contain size judgments for 596 common items. These items range from extremely small objects (e.g., ant, snowflake) to extremely large objects (e.g., destroyer, dinosaur). Items were rated on a scale of -9 for *very small* to +9 for *very large*. The large range and number of items made it possible to create two contexts with very different distributions of values across the dimension.

To ascertain whether the experimental manipulation had the predicted effect on the judgment of the absolute magnitudes of the experimental items, separate groups of subjects rated the materials in each context. One version of each list was constructed by randomly assigning items to a given position in the list. This ordering was reversed to form a second version of each list. Each list was presented to a separate group of approximately 40 subjects. The subjects were instructed to rate the size of each item on the list on a 20-point scale. They were to assign small numbers to small objects, large numbers to large objects, and appropriate intervening numbers for objects of intervening magnitudes. Two examples, whose rated sizes in the Shoben et al. (1989) norms just bracketed the sizes of the objects in the two lists, were provided to anchor

the scales. These ratings will be discussed along with the reaction time (RT) results in the Results section.

For the comparative judgment task, specific comparisons were restricted to pairs formed exclusively from experimental items or exclusively from filler items. Within each set of items, specific comparisons were constructed by pairing each item with another item from each of the next two serial positions. This process produced 26 experimental pairs, and 35 pairs in each of the two filler sets.

Design. The design of the experiment can be conceptualized as a separate linear order of experimental items that was embedded in a larger context. This linear order of experimental items was defined by equating magnitude positions in the Shoben et al. (1989) norms with serial position or rank. The advantage of equating rank with magnitude is that the effects of manipulating rank can easily be examined for a relatively large number of items. The Shoben et al. norms provided a common metric prior to the frequency manipulation. There were eight serial positions that were separated by approximately two scalar units (-7.5 , -5.5 , -3.5 , -1.5 , 1.5 , 3.5 , 5.5 , 7.5). Two items were selected at each magnitude position.

The actual rank of these experimental items was manipulated by adding different filler items to the two contexts. These items were chosen because they had magnitude positions that fell between two of the magnitude positions in the experimental set of items. In the small context, seven items were added at each of the following magnitude positions: -6.5 , -4.5 , -2.5 , and 0.0 . In the large context, seven items were added at each of the following magnitude positions: 0.0 , 2.5 , 4.5 , and 6.5 .

To eliminate the possibility that the presence of different filler items might bias the subjects to respond to the experimental items differently in each context, the filler items were never paired with the experimental items. This practice maintained the same response probabilities for experimental items in both conditions.

Procedure. In the small-context condition, the subjects were presented with the 26 experimental pairs and the 35 small filler pairs. In the large-context condition, they were presented with the 26 experimental pairs and the 35 large filler pairs. Each pair appeared twice with the instruction to select the smaller item and twice with the instruction to select the larger item. Position of the item (left or right) was counterbalanced with instruction. Thus, each of the 61 pairs appeared four times in a block. The order of trials was randomized within each block. The subjects were presented with two blocks of trials, for a total of 488 trials.

The stimulus pairs were presented on IBM XTs equipped with EGA monitors. Each trial began with a ready signal (the word "READY"), which remained on the screen until the subject pressed the space bar, at which point the comparative appeared on the screen. The comparative remained on the screen for 1.8 sec and was followed immediately by the test pair. RT was measured from the onset of the stimulus pair.

The subjects were informed that they would see pairs of items, preceded by a comparative ("smaller?" or "larger?"). They were to determine the correct answer and respond by pressing the "5" key if the correct answer appeared on the left side of the screen, and the "9" key if it appeared on the right side of the screen. If the subject made an incorrect response, the message "WRONG ANSWER" appeared for 4 sec.

The subjects were given detailed feedback about their performance at the end of each block and were informed about their accuracy and their mean response latency. They were strongly encouraged to respond accurately on at least 95% of their trials and to respond as quickly as possible while maintaining this level of accuracy.

Prior to the beginning of the experimental trials, the subjects received eight practice pairs that were not among the test stimuli but were selected from the same pool of items in the original Shoben et al. (1989) norms. A 20-sec break was provided halfway through each block.

Results

There are two major issues to consider in evaluating the results from this experiment. The first is whether the effects of manipulating rank on rated magnitude were consistent with the predictions of the range-frequency theory. A failure to observe an effect of stimulus rank on rated size would be problematic, because the implicit scaling hypothesis claims that similar factors govern the scaling of stimuli in both absolute and comparative judgment tasks. The second and primary issue is whether the results of the comparative judgment task are consistent with the range-frequency theory.

The mean size ratings and the rank for experimental items in each of these two contexts are presented in Table 1. The fact that the rank order of experimental items was greater in the small context than in the large context indicates that the distribution of stimulus values was positively skewed in the small context and negatively skewed in the large context. According to the range-frequency theory, this contextual manipulation should have had two main effects on rated size. First, the positive skew of stimulus magnitude in the small context in contrast to the negative skew of stimulus magnitude in the large context should increase the rated size of items in the small context relative to the large context. An inspection of Table 1 reveals that rated size was greater when small filler items were included ($x = 8.11$) than when large filler items were included ($x = 7.11$) [$F(1,77) = 13.90, p < .001$].

More important, for the predictions concerning the comparative judgment task, increasing the density of stimuli between any two stimuli should increase the difference in their rated sizes. Average differences in the rated size of pairs at each serial position are presented in Table 2. These data were analyzed by averaging the differences between the four smallest serial positions (e.g., 1-2,

Table 1
Rated Size, Rank, and Assigned Serial Position of the
Experimental Stimuli in Each Context for Experiment 1

Serial Position	Stimulus	Context			
		Small		Large	
		Rating	Rank	Rating	Rank
1	Raisin	1.30	3	1.13	1
	Vitamin	1.23	2	1.23	2
2	Mouse	3.53	14	2.85	4
	Whistle	2.50	10	2.51	3
3	Jar	4.38	20	3.59	6
	Photograph	3.22	11	3.15	5
4	Boot	5.82	28	4.51	7
	Telephone	5.93	29	4.56	8
5	Vulture	8.23	38	6.38	16
	Rake	7.63	34	6.08	15
6	Goat	10.73	40	8.21	22
	Tuba	10.57	39	7.85	19
7	Igloo	15.45	42	14.13	38
	Zebra	13.07	41	11.62	33
8	Truck	17.63	43	17.92	43
	Cabin	18.58	44	18.15	44

Table 2
Average Differences in Rated Size of Adjacent Items in Experiment 1

Serial Position	Context	
	Small	Large
1-2	1.75	1.50
2-3	.79	.69
3-4	2.07	1.17
5-6	2.72	1.80
6-7	3.61	4.84
7-8	3.85	5.16

2-3, 3-4) and the four largest serial positions (e.g., 5-6, 6-7, 7-8). For the three smallest serial positions, the average difference between adjacent items was greater in the small context ($x = 1.54$) than in the large context ($x = 1.12$), but for the three largest serial positions the average difference between adjacent items was smaller in the small context ($x = 3.39$) than it was in the large context ($x = 3.93$) [$F(1,77) = 16.74, p < .001$]. This interaction is consistent with the prediction that differences in rated size increase with the density of stimuli in a particular region of the linear order because differences in rated size were greatest in the densest region of the linear order for each context.

According to the implicit scaling hypothesis, an analogous scaling of stimuli occurs as part of the comparative judgment decision. Given that previous research has demonstrated that comparison difficulty is fairly sensitive to differences in rated size, these differences should produce observable effects on performance in a comparative judgment task. Therefore, a failure to observe an effect of stimulus density on comparative judgment cannot be attributed to a general insensitivity to the effects of stimulus density on assessed magnitude for these stimuli.

In addition to matching the qualitative predictions for Experiment 1, these data fit more quantitative predictions of the range-frequency theory quite closely. According to the range-frequency theory (Birnbaum, 1974), categorical ratings are described by Equation 1 when stimulus and response ranges are held constant:

$$C_{ik} = a s_i + b F_k(Q_i) + c, \quad (1)$$

where C_{ik} is the categorical rating of stimulus i in context k , s_i is the subjective magnitude of stimulus i , $F_k(Q_i)$ is the cumulative proportion of stimuli less than Q_i , and a , b , and c are constants. It follows from this equation that the difference in ratings in the two contexts is a function of differences in the cumulative proportion of stimuli smaller than a given stimulus in each context. In other words,

$$C_{\text{small},i} - C_{\text{large},i} = b[F_{\text{small}}(Q_i) - F_{\text{large}}(Q_i)]. \quad (2)$$

When differences in the rated size of stimuli across the two contexts were correlated with differences in cumulative proportion of stimuli less than this stimulus, the resulting correlation was $r = .78$ [$F(1,14) = 21.9, p < .05$]. Thus, the rating data match the specific predictions that were outlined for this experiment in a manner that is quite consistent with the range-frequency theory.

The results of the comparative judgment task appear in Table 3. Mean RTs and error rates are given for adjacent pairs at each serial position. Across the 61 experimental and filler stimulus pairs, error rates were moderately low (6%) and were correlated with RT ($r = .70$). For the experimental items at the three smallest and three largest serial positions, error rates were slightly higher (8%) and were correlated with RT (.54). RTs greater than three standard deviations from the mean were excluded from further analysis.

The result of primary interest is whether the RTs for the small pairs were fastest in the small condition and whether the RTs for the large pairs were fastest in the large condition. The data were analyzed by averaging the RTs for experimental items at the three smallest serial positions and by averaging the RTs for experimental items at the three largest serial positions. The subjects in the small-context condition were faster at judging pairs from the three smallest serial positions ($x = 1,648$ msec) than they were at judging pairs from the three largest serial positions ($x = 1,748$ msec), but those in the large-context condition were slightly faster at judging the largest pairs ($x = 1,709$ msec) than they were at judging the smallest pairs ($x = 1,736$ msec). This interaction between context and serial position was reliable [$F(1,45) = 5.18, p < .05$] and does not appear to be associated with any general differences in comparison difficulty in the two conditions. RT collapsed over all experimental pairs was quite comparable in the small context ($x = 1,698$ msec) and the large context ($x = 1,722$ msec) ($F < 1$). These results are important because they indicate that increasing the density of stimuli between any two stimuli facilitates comparisons made about these stimuli.

In addition to the interaction between context and serial position, there was a significant interaction between serial position and the form of the question [$F(1,45) = 18.50, p < .001$]. This interaction occurred because the subjects were faster at choosing the smaller than the larger of two small items, but were faster at choosing the larger than the smaller of two large items. This finding is a replication of a very robust phenomenon in the literature that is referred to as *the congruity effect*. No other difference in RT was significant.

Table 3
Reaction Times (RT, in Milliseconds) and Error Rates as a Function of Context and Stimulus Pair Magnitude in Experiment 1

Question Type	Low Magnitude		High Magnitude	
	RT	Error Rate	RT	Error Rate
Small-Context Condition				
Smaller	1,635	.04	1,805	.10
Larger	1,662	.05	1,691	.09
Difference	27		114	
Large-Context Condition				
Smaller	1,711	.08	1,743	.08
Larger	1,762	.09	1,674	.07
Difference	52		69	

The error data provide strong support for the claim that relative frequency affects comparison difficulty. The same interaction between serial position and context that was observed in the RT data is apparent in the error data. The subjects in the small-context condition made fewer errors on Serial Positions 1–3 (.04) than they did on the three largest serial positions (.09). In the large-context condition, they made slightly more errors on small items (.09) than they did on large items (.08). This interaction between context and serial position was very reliable [$F(1,45) = 16.80, p < .001$]. Finally, the error rate for small pairs (.06) was slightly less than the error rate for large pairs (.09) [$F(1,45) = 7.12, p < .05$].

Discussion

Theoretically, these crossover interactions in the RT and error data are important because they support the idea that manipulating relative frequency has its effect on an initial assessment of magnitude. This support stems from the fact that crossover interactions provide a great deal of constraint on any formalization of the judgment process. If one assumes that comparative judgments involve an initial assessment of the magnitudes of the stimuli followed by some comparison operation, then the effects of symbolic distance on the judgment process can be captured by Equation 3:

$$R_{ijk} = G_k[M_k(s_i) - M_k(s_j)]. \quad (3)$$

Equation 3 closely parallels a model of stimulus comparison proposed by Birnbaum (1982). In Birnbaum's formulation, R_{ijk} is the judged ratio of the subjective magnitudes of stimulus i to stimulus j . G_k is the comparison or judgment function for context k , M_k is the magnitude-assessment function for context k , and s_i and s_j are the subjective magnitudes of stimuli i and j , respectively. Although the exact form of judgment function G_k depends on the nature of the required response, it is assumed that the function either monotonically increases or decreases over the entire range of its arguments. In the case of "ratio" judgments, Birnbaum argued that G_k is approximately exponential, but for "difference" judgments the response, D_{ijk} , is an approximately linear function of the difference in the subjective magnitudes of the stimuli. In the current analysis, R_{ijk} refers to RT, and it is assumed that G_k is a monotonically decreasing function of the difference in the assessed magnitudes of stimuli i and j .

The application of Equation 3 is not intended to provide a complete model of the comparative judgment process because RT is not always a monotonic function of symbolic distance. In addition to differences in symbolic distance, RT is affected by the form of the question (e.g., semantic congruity) and the serial position of the items. For a given symbolic distance, RT will be greater when the magnitude of the items is incongruent with the form of the question than when it is congruent with the form of the question, and when the items are of moderate rather than extreme magnitude. Although Equation 3 does not account for these influences of semantic congruity and

serial position (Holyoak & Mah, 1982), virtually every model of the judgment process assumes that this relationship between RT and symbolic distance should hold when the effects of these other two factors can be excluded.

In the current circumstances, there is no reason to believe that the basic finding depends on either semantic congruity or serial position. First, the advantage of the small items in the small context and the large items in the large context did not differ greatly for each form of the question ($F < 1$). This indicates that the basic finding does not depend on semantic congruity. Second, the effects of serial position should have produced an interaction in a direction opposite to the one that was obtained. Instead of being faster in the small context than in the large context, the small pairs should have been slowest in the small context because the serial position of these pairs was more moderate in the small context than in the large context. Similarly, large pairs should have had an advantage in the small context rather than the observed advantage in the large context because the serial position of large pairs was more moderate in the large context than in the small context. This means that the observed advantage of small pairs in the small context and large pairs in the large context cannot be attributed to an advantageous change in their serial position across contexts. Finally, the application of Equation 3 provides a parsimonious account of the rating data in this experiment. Differences in the rated sizes of items have been widely utilized as an independent measure of symbolic distance because of their documented impact on RT in comparative judgment tasks (Banks & Flora, 1977; Paivio, 1975; Shoben et al., 1989). If it was assumed that the sole influence of the contextual manipulations was on a component of the decision process not identified in Equation 3, then a much more complicated account of the relationship between rated size and RT would have to be developed to account for the effects of contexts on the two tasks in this experiment. Thus, the application of Equation 3 simply models a basic relationship between RT and symbolic distance in a manner that is consistent with both the empirical literature and the predictions of virtually all existing models of comparative judgment.

An important implication of this equation is that the locus of the contextual effect will have a profound impact on the pattern of RTs. If contextual variation affects only the comparison operation, then Equation 3 simplifies to Equation 4 (Birnbaum, 1982):

$$R_{ijk} = G_k(s_i - s_j). \quad (4)$$

In this simplified equation, differences in the subjective magnitudes of stimuli, $s_i - s_j$, remain constant across contexts. Provided that G_k is a monotonically decreasing function of $s_i - s_j$, RT for given pairs will vary with context but the rank order of these RTs will remain constant across contexts. On the other hand, if contextual variation influences the initial assessment of the stimuli (e.g., Equation 3), then differences in the magnitudes will not remain constant across context and one might obtain a crossover interaction. The fact that the relative difficulty

and speed of judgments about small and large pairs reverses with context strongly supports the idea that manipulating stimulus rank influences an initial assessment of the magnitudes of the two stimuli.²

EXPERIMENT 2

The aim of Experiment 1 was to show how models that assume that the symbolic distance effect arises from differences in the discriminability of stimuli must assume that stimulus rank influences how magnitude is assessed. Although all current models of comparative judgment assume that the symbolic distance effect arises from the relative discriminability of stimuli, the case for applying range-frequency principles to comparative judgment should be greatly strengthened by further tests that do not involve the symbolic distance effect. It is for this reason that we examined the effect of stimulus rank on a very different phenomenon, the semantic congruity effect, in Experiment 2.

The congruity effect refers to the relative ease with which individuals can select either the greater of two objects that are high on a dimension or the lesser of two objects that are low on a dimension. Generally, the congruity effect occurs in the form of one of two types of interaction between the form of the question and the magnitudes of the stimuli. When this advantage of the congruent form of the question occurs for both forms of the question, the congruity effect appears in the form of a crossover interaction. When there is a general advantage of one form of the question that is smaller for incongruent judgments, the interaction is commonly referred to as a *funnel effect* interaction because the plotted data form a funnel (Audley & Wallis, 1964; Banks, Clark, & Lucy, 1975). Both types of interaction are commonly observed (Banks, 1977).

According to the range-frequency theory, the rated magnitude of an object decreases in contexts in which there are many larger objects and increases in contexts in which there are many smaller objects. If the range-frequency theory can be applied to contextual manipulations in comparative judgment tasks, then it should be possible to manipulate the locus of the congruity effect by skewing the distribution of stimuli.

Consider the pairs rat-book and lion-horse for a moment. In most contexts, these stimuli would be among the smallest and largest objects present. In these contexts, it is relatively easier to choose the smaller of rat-book and the larger of lion-horse, because *rat* and *book* are more congruent with the comparative "choose smaller" and *lion* and *horse* are more congruent with the comparative "choose larger." However, the range-frequency theory predicts that it should be possible to change this congruency by skewing the distribution of stimuli. Adding stimuli that are somewhat smaller than *rat* and *book* should cause the assessed magnitudes of these two objects to increase and become less congruent with the comparative "choose smaller." In contrast, skewing the distribution

by adding a number of stimuli that are larger than *lion* and *horse* should decrease the assessed magnitude of these objects and cause them to become less congruent with the comparative "choose larger." Thus, the advantage of the comparative "choose smaller" should be more pronounced for the pair rat-book in the presence of many large stimuli, but the advantage of the comparative "choose larger" for the pair lion-horse should be more pronounced in the presence of many small stimuli.

Evaluating this prediction is complicated by the fact that the magnitude of the congruity effect depends on symbolic distance. More specifically, the semantic congruity effect is either greatly reduced or disappears as the differences in the magnitudes of two objects increases (Petrusic & Baranski, 1989). For this reason, the design of Experiment 2 required that filler items be added in a manner that did not increase differences in the rank of certain critical experimental pairs. The rank of stimuli was manipulated by adding filler items whose magnitudes did not intervene between these critical pairs. For example, if *ant* were the smallest item in the experiment, then the ranks of *rat* and *book* would be increased by adding a term like *mouse*. Adding this term would increase the rank of both *rat* and *book* by one, but it would not change the difference in their ranks. Furthermore, adding *mouse* to the stimulus set would not change the stimulus range, because a mouse is larger than an ant. This general manipulation was the basis of Experiment 2.

Method

Subjects. Of the 26 subjects who participated in the large-context condition, 2 subjects failed to make an error criterion of 45 or fewer errors. In the small-context condition, 3 of the 27 subjects who participated failed to meet this error criterion. All of the subjects were undergraduates at Lehman College and participated for course credit.

Materials. The items were selected from the Shoben et al. (1989) norms. Again, these norms guided the assignment of stimuli to serial positions. Specifically, the experimental items were selected by assigning two items to each of the seven serial positions noted in the Design section. Filler items in the two contexts were selected by assigning four items to each of the two respective magnitude levels, specified below.

Materials for judging the absolute magnitude of items in each context were constructed in the same manner as they were in Experiment 1.

For the comparative judgment task, specific comparisons were restricted to pairs formed exclusively from experimental items or exclusively from filler items. The formation of experimental pairs was complicated by the requirement that each item appear equally often. At Serial Positions 1, 2, 5, and 6, adjacent pairs were formed by pairing each item with both items at the next higher serial position. At Serial Positions 3 and 4, each item was paired with a single item at the next higher serial position. Nonadjacent pairs were restricted to those consisting of items separated by one serial position (e.g., Serial Positions 1 and 3). Each item was paired with only one item from the larger serial position. These restrictions led to the formation of 20 unique adjacent pairs and 10 unique split 2 pairs. Filler pairs were formed by pairing each filler item with every filler item at the other magnitude level. This process formed 16 unique pairs.

Design. Two contexts were created by adding either small or large filler items to a list of experimental items. Again, serial position

or rank was established by equating serial position with rated magnitude in the Shoben et al. (1989) norms. Experimental items were selected at each of seven levels of magnitude or serial position (−7.9, −4.0, −2.06, .05, 1.93, 3.93, 8.15).

The confounding effects of symbolic distance were avoided by adding filler items whose magnitudes did not intervene between certain critical pairs. In the small context, filler items whose magnitudes fell between the magnitudes of the experimental items in Serial Positions 1 and 2 were added to the set of experimental items (−7.0, −5.1). In the large context, filler items whose magnitudes fell between the magnitudes of the experimental items in Serial Positions 6 and 7 (5.1, 7.1) were added. These additions increased the rank of stimuli in Serial Positions 2–6 in the small context, but decreased the rank of these stimuli in the large context without affecting the distances among these stimuli in either context. Thus, pairs formed from Serial Positions 2 and 3, and 5 and 6 provided the critical pairs necessary to determine whether skewing the distribution of stimuli affects the semantic congruity effect.

Procedure. The procedure of Experiment 2 was identical to that of Experiment 1, except that the subjects rated the absolute sizes of the items prior to the comparative judgment task. Thus, each of the 46 pairs appeared four times in each of two blocks, for a total of 368 trials.

Results

The results of the comparative judgment task appear in Table 4. Mean RTs and error rates are given for adjacent pairs at each serial position. Across all 46 experimental and filler pairs, error rates were moderately low (5%) and were correlated with RT ($r = .72$). Error rates on the critical test pairs (11%) were somewhat greater. RTs greater than three standard deviations from individual subjects' means were excluded from further analysis.

The major question that is addressed by this experiment is whether the size and direction of the congruity effect can be influenced by the distribution of stimuli. As was expected, there was an overall congruity effect in the two contexts that appeared as a funnel-effect interaction [$F(1,45) = 12.20, p < .01$]. For the question "which is larger?" the subjects were 205 msec faster at judging large pairs than judging small pairs, but for the question "which is smaller?" they were only 48 msec faster at judging large pairs than small pairs.

To determine whether the locus of this congruity effect varied with condition, the advantage of the comparative "choose smaller" for small pairs (e.g., pairs formed from Serial Positions 2 and 3) and the advantage of the

comparative "choose larger" for large pairs (e.g., pairs formed from Serial Positions 5 and 6) was evaluated in each condition. The advantage of each comparative was calculated by subtracting the congruent comparative (e.g., faster) from the incongruent comparative (e.g., slower). In other words, for small pairs, each subject's RT for "choose smaller" was subtracted from their RT for "choose larger," and for large pairs, their RT for "choose larger" was subtracted from their RT for "choose smaller." This subtraction yielded two difference scores, which increased as the advantage of the congruent comparative increased. If adding small filler items makes other items seem larger and adding large items makes other items seem smaller, then the congruency of small pairs in the small context and large pairs in the large context should be greatly attenuated. In concrete terms, difference scores for small pairs should be much smaller in the small context and difference scores for large pairs should be much smaller in the large context. The data bear out this prediction. The advantage of "choose smaller" for small pairs in the large context (148 msec) was 141 msec greater than it was for these pairs in the small context (7 msec). In contrast, the advantage of "choose larger" for large pairs in the large context (41 msec) was 76 msec less than it was for these same large pairs in the small context (117 msec). This interaction between magnitude of the pair and context was significant [$F(1,45) = 5.73, p < .05$].

In addition to this shift in the congruity effect, there was a main effect of serial position or magnitude. In both contexts, the subjects were faster at judging the larger items ($x = 2,006$ msec) than they were at judging the smaller items ($x = 2,132$ msec) [$F(1,45) = 14.81, p < .001$].

GENERAL DISCUSSION

Although most investigators who research comparative judgment have assumed that relative magnitude is computed from absolute magnitude information, there have been few attempts at integrating research on judgments of relative magnitude with research on judgments of absolute magnitude. The results of the present experiments indicate that comparative judgment is influenced by some of the same contextual factors that influence absolute or categorical judgments. Experiment 1 demonstrated that increasing differences in the ranks of items speeded judgment in a comparative judgment task and increased differences in the rated sizes of the same stimuli in an absolute judgment task. These results indicate that increasing differences in the ranks of stimuli increases the symbolic distance of a pair. Furthermore, the fact that the relative difficulty of certain pairs reversed with context provides support for the idea that this change in symbolic distance occurs in an initial operation that scales magnitude information according to the current context.

Experiment 2 demonstrated that the influence of relative frequency is not limited to the symbolic distance effect. When the rank of stimuli was increased so that they

Table 4
Reaction Times (RT, in Milliseconds) and Error Rates as a Function of Context and the Magnitude of the Congruity Effect in Experiment 2

Question Type	Low Magnitude		High Magnitude	
	RT	Error Rate	RT	Error Rate
Small-Context Condition				
Smaller	2,135	.07	2,039	.10
Larger	2,142	.07	1,922	.12
Difference	7		117	
Large-Context Condition				
Smaller	2,053	.09	2,053	.15
Larger	2,201	.10	2,012	.14
Difference	148		41	

were among the largest stimuli, the subjects were much faster at choosing the larger rather than the smaller of these stimuli. When the rank of stimuli were decreased so that they were among the smallest stimuli, the subjects were much faster at choosing the smaller rather than the larger of these stimuli. This change in the congruity of items provides further evidence that the assessed magnitude of items depends in part on their rank in the current context.

Implications for Current Models

Without modification, current models of comparative judgment cannot account for these results. Most models (Banks, 1977; Holyoak & Patterson, 1981; Jamieson & Petrusic, 1975; Marks, 1972) fail to account for these results because they do not have an explicit mechanism for determining the effects of context. For example, Banks's discrete-code model proposes that discrete magnitude codes are generated and compared. These codes are assumed to be assigned with a probability that is a function of the magnitudes of the objects—small objects have a greater probability of being coded as “small” and large objects have a greater probability of being coded as “large.” Once two different codes are generated, the object whose code matches the instruction to “choose larger” or “choose smaller” is selected. Although the criterion that determines which code is generated for an object is assumed to vary, the mechanism that determines its setting is not specified. The current results indicate that an explicit mechanism must be specified before the discrete-code model could provide an adequate account.

An important exception to these models is Holyoak's reference-point model (Holyoak, 1978; Holyoak & Mah, 1982), which explicitly proposes that the magnitude of items is scaled as a function of the current context. According to the reference-point model, the difference between the magnitude of each object and a common reference point is determined. These differences are equated with distance along the dimension in question. The object with the shorter distance to the reference point is selected. The model is described by Equation 5:

$$R_{ijk} = J_k[f(s_i - s_k) - f(s_j - s_k)], \quad (5)$$

where R_{ijk} is the response to stimuli i and j ; J_k is a monotonic judgment function for context k ; s_i , s_j , and s_k are the subject scale values for stimuli i and j and the implicit reference point k ; and f is an increasing, negatively accelerated function.

An important feature of this model is that f is a compressive function that scales the distances between each object and the reference point. It makes the difference between an item that is far from the reference point and one that is close to the reference point more similar than the original differences in their untransformed scale values. As a result, two objects will be more difficult to compare when they are far from the reference point than when they are close to it. Thus, the reference-point model scales

stimuli in terms of how far they are from a given reference point.

Normally, there are two reference points that are placed at each end of the dimension; their exact placement depends on the current context. It is assumed that the low-magnitude reference point's value is slightly less than the least extreme stimulus and that the high-magnitude reference point's value is slightly greater than the most extreme stimulus. The form of the question determines which of these two reference points is used when two objects are compared. For the question “which is larger?” the high magnitude or “large” reference point is selected, and for the question “which is smaller?” the low magnitude or “small” reference point is selected.

Most of the inability of the reference-point model to account for the current results stems from the fact that the influence of context is governed entirely by the scaled distance from the reference point specified by the comparative. For example, the answer to “which is larger, a dog or a rabbit?” is determined solely on the basis of the distance of each object from the large reference point. Given that the addition of stimuli whose values fall between *dog* and *rabbit* do not change their scale values, the distance of either stimulus from the reference point should be unchanged. This feature leaves the model unable to account for the fact that judgments were easier when more stimuli intervened between a pair of stimuli, even when the range of stimuli was held constant.

In addition, previous research indicates that the reference-point model has difficulty accounting for certain manipulations of the stimulus range. Čech and Shoben (1985) found that judgments under both comparatives were affected when either the top half or the bottom half of the stimulus range was removed. The model cannot explain this effect on both forms of the question because this manipulation should have influenced the placement of one reference point and left the other reference point in place. Taken together, these results indicate that the basic reference-point model cannot account for the current results.

The reference-point model can be modified to account for these results if one assumes that, in addition to distance from the reference point, stimulus range and density affect comparison difficulty as well. In fact, Holyoak and Mah (1982) proposed that range-frequency effects can be integrated with the basic reference-point model. Although they do not specify how the basic mathematical description of the reference-point model is to be modified, this integration can be expressed by Equation 6:

$$R_{ijk} = J_k\{f[M_k(s_i) - M_k(s_k)] - f[M_k(s_j) - M_k(s_k)]\}, \quad (6)$$

where M is a function that scales magnitude information according to the principles of the range-frequency theory. This modified reference-point model explains the density effects in Experiment 1 by assuming that increas-

ing the density caused the difference between the further stimulus and the reference point, $M_k(s_i) - M_k(s_k)$, to increase. The change in the congruity effect that was observed in Experiment 2 can be explained if it is assumed that both $M_k(s_i)$ and $M_k(s_j)$ increased relative to the reference point $M_k(s_k)$. Thus, the reference-point model can account for the basic results of the current experiments if it is modified to incorporate an implicit scaling process.

Conclusions

The preceding discussion was not intended as an endorsement of either the discrete-code or reference-point models. Even with modification, each model has difficulty accounting for other important findings in the literature. For example, the reference-point model predicts that the serial position function should be a monotonically increasing function, which is inconsistent with the finding that the serial position function is frequently "bowed" (Shoben et al., 1989).

Instead, the preceding discussion was intended to demonstrate that the problems that each model encounters with the current results are not peculiar to particular processing or representational assumptions. The fact that these experiments are problematic for two models with very different processing and representational assumptions supports the notion that models of comparative judgment must include detailed specifications about how stimulus range and density affect judgment. It is the position of this paper that the best account of this specification is one that assumes that an initial scaling of the stimuli occurs prior to their comparison.

REFERENCES

- AUDLEY, R. J., & WALLIS, C. P. (1964). Response instructions and the speed of relative judgments: 1. Some experiments on brightness discrimination. *British Journal of Psychology*, 55, 59-73.
- BANKS, W. P. (1977). Encoding and processing of symbolic information in comparative judgments. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 11, pp. 101-159). New York: Academic Press.
- BANKS, W. P., CLARK, H. H., & LUCY, P. (1975). The locus of the semantic congruity effect in comparative judgments. *Journal of Experimental Psychology: Human Perception & Performance*, 1, 35-47.
- BANKS, W. P., & FLORA, J. (1977). Semantic and perceptual processes in symbolic comparisons. *Journal of Experimental Psychology: Human Perception & Performance*, 3, 278-290.
- BANKS, W. P., FUJII, M., & KAYRA-STUART, F. (1976). Semantic congruity effects in comparative judgments of magnitude of digits. *Journal of Experimental Psychology: Human Perception & Performance*, 2, 435-447.
- BIRNBAUM, M. H. (1974). Using contextual effects to derive psychophysical scales. *Perception & Psychophysics*, 15, 89-96.
- BIRNBAUM, M. H. (1982). Controversies in psychological measurement. In B. Wedener (Ed.), *Social attitudes and psychophysical measurement* (pp. 401-485). Hillsdale, NJ: Erlbaum.
- ČECH, C. G., & SHOBN, E. J. (1985). Context effects in symbolic magnitude comparisons. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 11, 299-315.
- HOLYOAK, K. (1978). Comparative judgments with numerical reference points. *Cognitive Psychology*, 10, 203-243.
- HOLYOAK, K. J., & MAH, W. A. (1981). Semantic congruity in symbolic comparisons: Evidence against an expectancy hypothesis. *Memory & Cognition*, 9, 197-204.
- HOLYOAK, K., & MAH, W. A. (1982). Cognitive reference points in judgments of symbolic magnitude. *Cognitive Psychology*, 14, 328-352.
- HOLYOAK, K. J., & PATTERSON, K. K. (1981). A positional discriminability model of linear order judgments. *Journal of Experimental Psychology: Human Perception & Performance*, 7, 1283-1302.
- JAMIESON, D. G., & PETRUSIC, W. M. (1975). Relational judgments with remembered stimuli. *Perception & Psychophysics*, 18, 373-378.
- MARKS, D. F. (1972). Relative judgment: A phenomenon and a theory. *Perception & Psychophysics*, 11, 156-160.
- MELLERS, B. A., & BIRNBAUM, M. H. (1982). Loci of contextual effects in judgment. *Journal of Experimental Psychology: Human Perception & Performance*, 8, 582-601.
- MOYER, R. S. (1973). Comparing objects in memory: Evidence suggesting an internal psychophysics. *Perception & Psychophysics*, 13, 180-184.
- PAIVIO, A. (1975). Perceptual comparisons through the mind's eye. *Memory & Cognition*, 3, 635-647.
- PARDUCCI, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, 72, 407-418.
- PARDUCCI, A., & HAUGEN, R. (1967). The frequency principle for comparative judgments. *Perception & Psychophysics*, 2, 81-82.
- PETRUSIC, W. M., & BARANSKI, J. V. (1989). Semantic congruity effects in perceptual comparisons. *Perception & Psychophysics*, 45, 439-452.
- SHOBN, E. J., ČECH, C. G., SCHWANENFLUGEL, P. J., & SAILOR, K. M. (1989). Serial position effects in comparative judgments. *Journal of Experimental Psychology: Human Perception & Performance*, 15, 273-286.

NOTES

- Experiments 4 and 5 may constitute exceptions to this claim. The middle six items in a 12-term linear order were removed during the experiments and subjects were subsequently tested on the three largest and smallest items. However, ordinal separation does not change, and any change in the relative frequency of adjacent-item changes is very small with this manipulation.
- It should be noted that in the case of within-modality judgments of perceptual stimuli, Mellers and Birnbaum (1982) obtained support for the claim that the locus of contextual effects was in the judgment function G_k , but in the case of cross-modality comparisons they obtained support for the claim that the locus of contextual effects was in a scaling or magnitude-assessment function, M_k .

(Manuscript received March 9, 1992;
revision accepted for publication December 31, 1992.)