A congruity effect in the discrimination of presentation frequencies: Some data and a model

DOUGLAS L. HINTZMAN and ERIC GOLD
University of Oregon, Eugene, Oregon 97403

In a mental comparison task, subjects discriminated the frequencies of words that had been presented from zero to five times each. Reaction times were recorded. Half the subjects indicated which member of a pair had occurred more often and half indicated which had occurred less often in the list. Reaction times displayed both a distance effect and a congruity effect. A simulation model, based on a theory in which the congruity effect entails a speed-accuracy tradeoff, did a good job of accounting for the data.

A congruity effect is demonstrated in a mental comparison task when subjects respond especially quickly to test pairs whose members fall toward the end of the comparative dimension named by the instructions. For example, subjects asked to choose the louder of two imagined sounds are faster at comparing loud sounds than soft sounds, but the instruction to choose the softer sound reverses the direction of the difference (Intons-Peterson, 1980). Congruity effects have been found in a variety of semantic comparison and perceptual comparison tasks. In this article, we demonstrate a congruity effect in an episodic memory task: the discrimination of presentation frequencies. We then describe a simulation model that accounts for our results and may be applicable to other mental comparison tasks, as well.

The experiment we report was an attempt to understand a puzzling phenomenon uncovered by Hintzman, Grandy, and Gold (1981). In their Experiment 2, subjects saw individual words from one to four times each and then were shown the words in pairs and were asked to choose the more frequent member of each pair. Reaction times (RTs) displayed two phenomena of interest: First, the greater the difference between the two frequencies being compared, the shorter was the RT. This "distance effect" had been predicted by Hintzman et al. on the assumption that frequency discrimination is done by retrieving information about both frequencies and comparing them. The result appears to rule out theories in which no such explicit comparison process occurs. Second, they found that the greater were the absolute frequencies of the two items, the shorter was the RT. It is this phenomenon, which we shall call the magnitude effect, with which the present article is primarily concerned.

Hintzman et al. (1981) suggested two explanations of the magnitude effect. The first, labeled the "synergistic retrieval" hypothesis, was that different memory traces of the same item tend to activate each other, so that the more such traces there are, the more quickly each of them is retrieved. An increase in absolute frequency, by this hypothesis, reduces RT by reducing retrieval time. The second hypothesis was that the decrease in RT with increasing frequency was a case of the congruity effect, that is, that the instruction to choose the more frequent alternative imparted an advantage to comparisons of high-frequency but not low-frequency pairs.

The present experiment was a direct test of the second hypothesis. Some subjects were instructed to choose the more frequent and some the less frequent members of the test pairs. The prediction was that the magnitude effect would be found with the "more frequent" instructions but would be reversed with the instruction to choose the less frequent member of each pair.

METHOD

The experimental stimuli were 180 common nouns. The words were assigned to three 60-word sets, two consisting of four-letter and one of three-letter words. Each subject saw three lists: one created from each word set. Each 162-item list was generated as follows: The first 6 and last 6 positions were occupied by filler words, and the middle 150 positions were occupied by 50 experimental words from the appropriate set, 10 occurring at each of the frequencies one to five. The remaining 10 words were assigned the frequency "zero" and did not appear in the list. Within each list, two words at each frequency were randomly paired, for testing, with two words at each of the other five frequencies. To equate recencies of the paired words, final presentations were assigned semirandomly to Positions 107-156, with words to be paired on the test appearing in adjacent positions. Earlier presentations were randomly ordered in Positions 7-106 with the restriction that a given word could not occur twice in succession. Six lists were constructed from each of the three word sets by rotating words through the frequencies zero to five. Five subjects in each instruction condition saw each of the six lists.

Sixty University of Oregon students and members of the Eugene community were tested individually. Half were assigned to the "more" instruction condition and half to the "less" condition. Each subject was seated before a video screen controlled by a PDP-15 computer, and each heard a description of the procedure. The subject was told to try to respond quickly while remaining as accurate as possible. After a 5-sec warning, the first list was shown, one word at a time, with a 3,000-msec exposure and a 500-msec ISI. The subject then initiated presentation of the 30 test pairs. The ordering of the pairs and right-left word assignments were randomly determined for each subject. Subjects indicated their choices by pressing either a left- or a right-hand response key, as appropriate. A response terminated the display, and, after a 5,000-msec ISI, the next test pair was presented. Lists 2 and 3 were presented and tested in the same way as List 1.

RESULTS AND DISCUSSION

Data were collapsed over the three lists. Mean correct RTs are shown in the top two panels of Figure 1. The parameter of the curves represents the larger and the baseline, the smaller of the two frequencies. Since general trends in the data are most easily summarized by correlational statistics, in that which follows conclusions will be based upon correlations between independent variables and the data points of Figure 1.

The RTs of both groups showed a distance effect. That is, RT decreased as the difference between the larger and smaller frequencies increased. The correlation
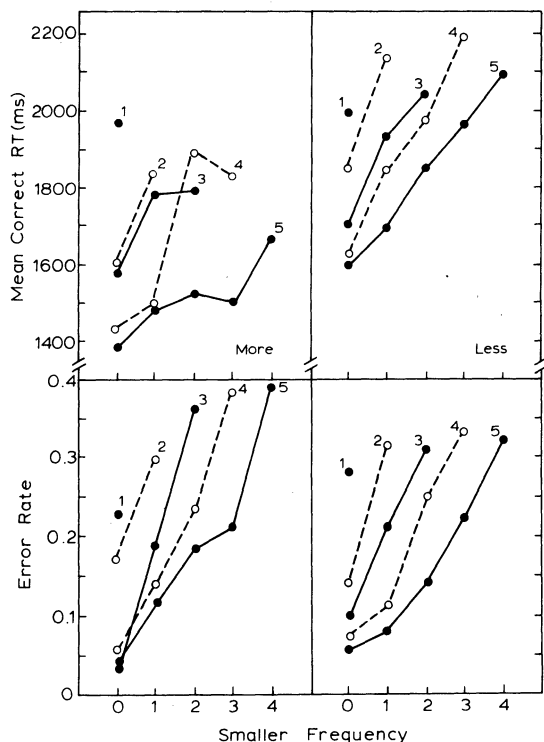

Figure 1. RT and error data for subjects instructed to choose the more frequent (left panels) and the less frequent (right panels) item. The parameter of the curves is the larger frequency.

of mean correct RT with distance was −.82 for the more condition and −.94 for the less condition. The magnitude effect can best be assessed with partial correlations. The correlation of RT with the larger frequency, with distance factored out, was −.45 for the more condition and .63 for the less condition. That the magnitude effect was reversed by the instructions is indicated by the reversal of the sign. To see the distance effect in Figure 1, notice, in both of the top panels, that RT decreased with the larger frequency and increased with the smaller frequency. To see the effect of instructions, notice the relatively large vertical separation of the curves in the top left panel and the relatively steep slopes of those in the right.

A planned-comparisons analysis of variance confirmed the differential effect of instructions. The overall between-group difference was not significant [$F(1,58) = 2.78$, $p > .05$]. Linear trends on the larger and the smaller frequencies were significant [$F(1,58) = 28.41$ and 16.22, respectively; both ps < .001]. Both of these trends interacted with groups [$F(1,58) = 4.09$ and 4.92, respectively; both ps < .05]. The general form of this interaction is that of a congruity effect. That is, quite apart from the effect of distance, subjects in the more group tended to respond more quickly to high frequencies than to low, whereas those in the less group responded more quickly to low frequencies than to high.

Mean error rates are shown in the bottom two panels of Figure 1. Both groups' error rates displayed a distance effect: Correlations with distance were −.87 and −.94 for the more and less groups, respectively. The partial correlation with the larger frequency, with distance factored out, was .74 for the more and .49 for the less condition. Since neither value was negative, we can conclude that a magnitude effect did not show up in error rates. Indeed, the partial correlation was more positive for the more condition than for the less condition, which is the opposite of the effect of instructions on RTs. This suggests that the instructional manipulation may have produced a complex speed-accuracy tradeoff.

A planned-comparisons analysis of variance was performed on the arcsine-transformed error proportions. It revealed no main effect of instructional group [$F(1,58) = .06$], but it did show significant linear trends on both the larger and the smaller frequency [$F(1,58) = 16.71$ and 219.6; both ps < .001]. Group did not interact reliably with either the larger or the smaller frequency [$F(1,58) = 1.69$ and 3.69], although the latter value was nearly significant (.05 < p < .10).

These results replicate and extend the findings of Hintzman et al. (1981) in several ways. They provide further evidence for a distance effect on RTs in the frequency discrimination task. This outcome indicates that subjects retrieve and evaluate the frequencies of both alternatives in making their decisions, rather than searching memory for either alternative and responding

with the first one that is found. A model assuming the latter process was proposed by Estes (1976; Whitlow & Estes, 1979). The data also reaffirm that the probability of correctly choosing A over B can exceed the frequency ratio f(A)/[f(A) + f(B)], which is the upper bound predicted by the Estes model. While performance was not as high as that in Hintzman et al. (1981), it exceeded the model's prediction for every pairing not including a frequency of zero.

That the magnitude effect obtained by Hintzman et al. (1981) is replicable is shown by the performance of the "more" group, and the fact that RTs of the "less" group did not display the magnitude effect is as predicted by the hypothesis that the phenomenon represents a congruity effect. Apparently, the alternative notion of synergistic retrieval, outlined earlier, is superfluous; congruity with the instructions is sufficient to account for the magnitude effect.

It is particularly noteworthy that the congruity effect in RTs may have entailed a speed-accuracy tradeoff, even though the trend in error rates fell short of significance, especially since some theories of congruity effects predict the opposite result. In particular, there are a number of models that assume that the to-be-compared items are located on an analog continuum and that the instructions set an expected range of magnitudes (Marschark & Paivio, 1979), a region of maximal sensitivity (Kosslyn, Murphy, Bemesderfer, & Feinstein, 1977), or a reference point (e.g., Holyoak, 1978; Marks, 1972), making nearby locations more discriminable than more remote points on the continuum. Presumably, comparisons in the sensitive region should take less time, but they should also be more accurate. Thus, RTs that are shortened by the instructional manipulation should be associated with low error rates and those that are lengthened, with high error rates. Given the present pattern of errors, it seems unlikely that such models are appropriate for the frequency discrimination task.

SIMULATION MODEL

In attempting to understand the process that may underlie our results, we have developed a simple random-walk simulation model. The model was designed to capture two intuitions we have about performance in mental comparisons tasks. The first is that the mental comparison paradigm involves not only the comparison of two values, but also an affirmative response toward one of them as having a value generally consistent with the instructions (e.g., that the value is relatively low or relatively high on the scale). The second intuition is that what is accepted as consistent with the instructions may be dependent on an "adaptation level" established by the experimental context.

A brief description of the model follows: On a test trial, during each cycle or instant of time, evidence is sampled at random regarding the values of both A and B on the comparative dimension. There are six overlapping sampling distributions, one for each of the frequencies zero to five. Specifically, V(A), the sampled value of A, is generated by a random process that produces integers approximately normally distributed with a mean of .5f(A) and a variance of 3.3 + .25f(A), where f(A) is the item's frequency.[1] The sampled values contribute to two simultaneous random walks, one accumulating evidence favoring

Alternative A and one accumulating evidence favoring B. Whichever process is first to reach a criterion value, c, determines the choice, and the number of sampling cycles taken to reach c determines RT.

Two kinds of evidence contribute to either random walk. The first is the signed difference between the two sampled values. Since a difference favorable to one alternative is negative with regard to the other, the amounts added to the two accumulators sum to zero. It is through this negative yoking of the two processes that the frequency of the rejected alternative influences RT. The other kind of evidence contributing to either random walk arises from independent comparisons of the two sampled values with a standard, a kind of adaptation level falling somewhere within the range of recently sampled values. If a sampled value is between the standard and the end of the dimension named by the instructions, then the absolute difference between the value and the standard is added to the appropriate accumulator (otherwise, nothing is added). Although strictly speaking only the first kind of evidence is directly relevant to the subject's task of evaluating A vs. B, note that the second kind of evidence is correlated with the first. Thus, the higher is the frequency of A compared to the average, the more likely it is to be higher than the frequency of B. The model simply makes use of this partially redundant information, and this is what causes the congruity effect.

Consider, first, Version 1 of the program, which decides which alternative is more frequent. Let $V_t(A)$ and $V_t(B)$ be the values sampled for Alternatives A and B on Sampling Cycle t, and let S be the standard. Then, at the end of Cycle T, the accumulator for A contains:

$$Q_T(A) = j \times \sum_{t=1}^{T} [V_t(A) - V_t(B)] + k \times \sum_{V(A)>S} [V_t(A) - S], \quad (1)$$

where j and k are weights given the two kinds of information. $Q_T(B)$, the sum in the B accumulator, is obtained by exchanging the roles of A and B in Equation 1. Processing is terminated when either $Q_T(A) > c$ or $Q_T(B) > c$ [unless $Q_T(A) = Q_T(B)$]. The value of T at this point determines RT.

Version 2 of the program, which decides which alternative is less frequent, is similar except for the accumulation rules. The sum in the accumulator for A is:

$$Q_T(A) = j \times \sum_{t=1}^{T} [V_t(B) - V_t(A)] + k \times \sum_{V(A)<S} [S - V_t(A)], \quad (2)$$

and $Q_T(B)$ is obtained by switching A and B in Equation 2.

After preliminary exploration of parameter values, simulation runs were conducted using j = 3, k = 1, and c = 20. The value of S was set at 1 for Version 1 of the program and at 0 for Version 2. Each pairwise comparison among the frequencies zero to five was replicated 1,500 times with Version 1 and 1,400 times with Version 2. The results are shown in Figure 2.

As is revealed by the top panels of Figures 1 and 2, the model produced a congruity effect in RTs similar to that shown by our human subjects. In particular, the larger frequency affected RT more in Version 1 (the more condition), and the smaller frequency affected RT more in Version 2 (the less condition). The similarity between the simulation and the human data establishes the plausibility of our conjecture that the congruity effect we found may reflect a parallel comparison process in which the alternatives are compared not only with each other, but also, independently, with a kind of adaptation level or average.

Examination of the error rates, in the bottom panels of Figures 1 and 2, is also revealing. The parallel comparison process that produces the congruity effect in the simulated RTs is one that necessarily entails a speed-accuracy tradeoff. Thus, evidence that the frequency of A is high relative to S is largely redundant with evidence that the frequency of A is higher than
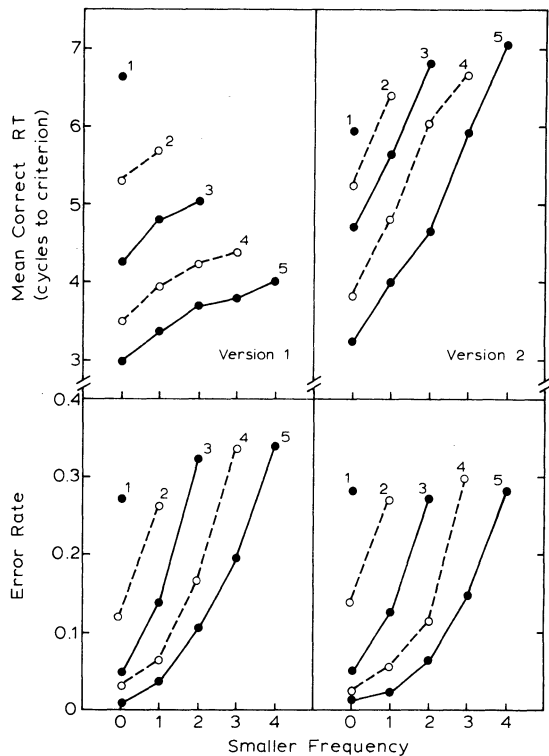
Figure 2. Data from Versions 1 and 2 of the simulation model.

ure 2 probably would not be statistically significant. Moreover, in showing that a substantial increase in speed can be achieved at a very low cost in errors, the simulation suggests that a person would be inefficient not to engage in the sort of processing assumed by the model if he or she could do it.

Second, one can ask whether congruity effects that have been observed in other mental comparison tasks might be explained in essentially the same way. One finding, from a serial order judgment task, that seems to support our approach is that adding a new end item (and thereby shifting the average) decreases the size of the congruity effect on items near that end (Banks, White, & Mermelstein, 1980). However, the prediction of a speed-accuracy tradeoff seems to be what primarily differentiates our model from other congruity effect models in the literature, and unfortunately, as we have seen, the predicted tradeoff in errors is not large. It could easily be mistaken for a null or even the opposite effect, particularly if the error sample is small, as it is in most mental comparison experiments. It has not been possible, in the articles we have examined, to determine whether a tradeoff of the kind predicted by the model may have occurred. Typically, the reporting of error rates is perfunctory, and errors may have been too infrequent to be revealing, in any case. In order to adequately rule out a speed-accuracy tradeoff explanation of congruity effects, it appears that mental comparison experiments with both large numbers of observations and relatively high error rates will have to be done.

REFERENCES

BANKS, W. P., WHITE, H., & MERMELSTEIN, R. Position effects in comparative judgments of serial order: List structure vs. differential strength. *Memory & Cognition*, 1980, 8, 623-630.

ESTES, W. K. The cognitive side of probability learning. *Psychological Review*, 1976, 83, 37-64.

HINTZMAN, D. L., GRANDY, C. A., & GOLD, E. Memory for frequency: A comparison of two multiple trace theories. *Journal of Experimental Psychology: Human Learning and Memory*, 1981, 7, 231-240.

HOLYOAK, K. J. Comparative judgments with numerical reference points. *Cognitive Psychology*, 1978, 10, 203-243.

INTONS-PETERSON, M. J. The role of loudness in auditory imagery. *Memory & Cognition*, 1980, 8, 385-393.

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.

MARKS, D. F. Relative judgment: A phenomenon and a theory. *Perception & Psychophysics*, 1972, 11, 156-160.

MARSCHARK, M., & PAIVIO, A. Semantic congruity and lexical marking in symbolic comparisons: An expectancy hypothesis. *Memory & Cognition*, 1979, 7, 175-184.

WHITLOW, J. W., & ESTES, W. K. Judgments of relative frequency in relation to shifts of event frequencies: Evidence for a limited-capacity model. *Journal of Experimental Psychology: Human Learning and Memory*, 1979, 5, 395-408.

NOTE

1. Details are available from the first author.

that of B, and processing can be speeded by capitalizing on that redundancy. But the increase in speed comes at some cost in errors, since the frequency of A could be fairly high and still be less than that of B. What the simulation shows is that the cost of capitalizing on this redundancy is quite low. Indeed, at first glance, the two lower panels of Figure 2 appear to be nearly identical. Nevertheless, the tradeoff exists, and it is best summarized by partial correlations. For Version 1 of the model, the correlation of error rate with the larger frequency, with distance factored out, was .55; the corresponding figure for Version 2 was .07.

The simulation model can be evaluated with regard to two questions. First, does it provide an adequate account of our experimental results? The purpose of the simulation exercise was not to provide a quantitative fit to the data, but only to determine at a qualitative level whether our account of the obtained congruity effect was reasonable. In this respect, the model was clearly successful. The simulation demonstrates that a congruity effect in RTs can result from the parallel accumulation of two partially redundant kinds of information, as we have assumed. The model is consistent with the apparent speed-accuracy tradeoff in our experimental data, and it helps explain why the tradeoff was significant in RTs but not in error rates. Without a very large number of observations per data point, a tradeoff as small as the one exhibited in the error data of Fig-