

Reaction time as a measure of inter- and intraobject visual similarity: Letters of the alphabet

PETER PODGORNÝ and W. R. GARNER
Yale University, New Haven, Connecticut 06520

Interobject visual similarity is often measured by the time that subjects require to say that two objects are not identical, with long RTs indicating high similarity. In Experiment 1, using a complete set of uppercase alphabetic stimuli, we show: (1) RTs correlate strongly with direct ratings of visual similarity. (2) Stimuli that are similar as measured by RT or by direct judgment have the same profile of similarity to all other stimuli in the set as measured by correlational analysis. (3) The order in which the stimuli are compared has only a small effect on measured similarity. All of these results indicate desirable properties of any measure of similarity. *Intraobject* similarity is a concept pertinent to the relation of an object to itself. In Experiments 1 and 2 we show: (1) All letters are not equally similar to themselves, since RT for the "same" response is different for different letters. (2) The relation between RT and intraobject similarity is opposite to that for interobject similarity, with short RTs indicating high intraobject similarity. (3) Even though intraobject similarities differ, each letter is more similar to itself than to any other letter in the set. We discuss the implications of these results in terms of assumed proximity constraints underlying similarity data.

Confusability as a Measure of Similarity

There are many different ways of defining the visual similarity between a pair of objects. Psychologists concerned with developing performance-based theories of pattern perception, as well as with explaining possible misperception, have often relied on a "confusion" or "substitution" measure of similarity. Pairs of objects whose members are frequently confused with, and hence substituted for, one another would receive high values of similarity by such a measure, whereas pairs whose members are not elicited by their counterparts would receive lower values. Absolute judgment or identification of objects under degraded viewing conditions is the most commonly used technique for generating such confusions, but studies of stimulus generalization in learning and of selective adaptation to repeated stimuli—studies which can be performed readily on infrahuman and nonadult populations—also are based on a variant of the confusion-substitution paradigm.

The rationale behind a definition of similarity in terms of confusability is quite simple. In the laboratory as well as in real life, an object is most often "confused" with itself—which is to say the

object is not confused or misidentified at all. An object's similarity to itself, then, is the maximal value of similarity that the object, as a member of a pair, can possibly receive. Adopting the relation of identity as the relation of maximal similarity requires that nonmaximal values of similarity be allocated in graded fashion as a function of the frequency with which an observer misidentifies a nonexistent relation of identity. Such an assignment of nonmaximal values of similarity is reasonable given that all nonidentical objects are not equally confused with one another.

Reaction Time as a Measure of Confusability

In binary-choice tasks, a type of confusion error is committed that typically has an experimentally useful concomitant. If asked to respond whether or not two stimuli are the "same" or "different," a subject will generally answer correctly (for non-demanding discriminations) but will occasionally make an error either by calling two identical stimuli "different" or by calling two nonidentical stimuli the "same." The latter type of error is directly analogous to the confusion errors discussed above for tasks using absolute judgment or identification, since two nonidentical objects are classified as being identical. The interesting covariant of this type of error is reaction time (RT). The pairs of objects that lead to incorrect responses of "same" are usually responded to more *slowly* by the subject. For this reason, choice RT in tasks requiring visual discrim-

This research was supported by Grant MH 14229 from the National Institute of Mental Health to the second author and was conducted during the first author's tenure as a National Science Foundation Graduate Fellow. Reprint requests may be sent to either author at the Department of Psychology, Box 11-A Yale Station, New Haven, Connecticut 06520.

ination has proven to be an extremely useful and popular measure of confusability and, therefore, of similarity (see Shepard & Podgorny, 1978, for a recent review).

The fact that RT and number of errors are correlated in a binary-choice task provides protection against any extreme "speed-accuracy tradeoff" (Pachella, 1974). Moreover, RT as an experimental measure offers significant advantages over most confusion measures, since, in order for subjects to commit a sufficiently large number of confusion errors to permit analysis, experimenters must degrade the viewing conditions under which subjects operate. Typical methods of producing such degradation are by briefly exposing the stimuli, by reducing the contrast between figure and ground, and by masking the stimuli with patterned noise. Since different types of degradation are known to produce different patterns of confusions (Garner & Haun, 1978), the resulting similarity structure is going to depend on the type of degradation chosen. In addition, even when the average probability of a confusion error is fixed at some high value (say, 50%), there are still likely to be many cells in the resulting confusion matrix that have an entry of zero. In contrast, choice RT can be obtained with sets of stimuli that are not visually degraded, and every cell of the resulting RT matrix is guaranteed to contain an informative experimental observation.

Outline of this Paper

In the present paper, we investigate some consequences of adopting discriminative RT as a measure of visual similarity:

First, we attempt to validate the RT measure as a measure of visual similarity by testing for a comparable similarity structure between an RT matrix based on speeded choice judgments and an independent matrix based on subjects' direct judgments of "how visually similar" two objects are.

Second, we explore the extent to which the RT measures conform to three general "proximity constraints" for *any* similarity data (Shepard, 1962). Briefly, these constraints require that: (1) The similarity of an object to itself is greater than the similarity of that object to any other object (the identity condition). (2) The similarity between two objects does not depend on the order in which the objects are compared (the symmetry condition). (3) Two objects that are maximally similar without being identical should be approximately equal in similarity to any third object (the common profile condition).

Since the frequency with which a subject says that an object is identical to a reference object is a reasonable measure of the similarity of that object to the reference, the first proximity constraint codifies

the rationale for using measures of confusion as measures of similarity. The next two proximity constraints appear to exemplify certain of our intuitions about similarity, but these constraints are not mandated by our choice of confusion as the prototypical measure of similarity and, instead, require more elaborate assumptions about the model alleged to underlie a similarity structure (see, e.g., Beals, Krantz, & Tversky, 1968).

In fact, there is no guarantee that all pairs of identical objects will be evaluated as equally similar. From what has been said above about confusion data, we can conclude only that a given object should be more similar to itself than to any other object, but not that every object must be equally similar to itself. In terms of an n by n similarity matrix, for which the value in cell (i,j) represents the similarity of row object i to column object j , the diagonal cells (i,i) need not be equal for all i . Rather, the fundamental constraint is that for a given object i , cell (i,i) must be greater than both cell (i,j) and cell (j,i) for all j . Under the above assumption, it is theoretically possible for cell (i,i) to be less than cell (j,k) —a result whose interpretation would mean that object i is less similar to itself than object j is to object k .

Third, our ultimate objective is to analyze and to discuss the similarity structure of the particular set of stimuli we have chosen to study: the uppercase English alphabet. Such an analysis and discussion should throw light on possible modes of internally representing letters and on possible operations involved in the process of internal representation. In particular, we would like to know whether "same" RTs and "different" RTs are pointing to a common similarity structure and to common principles of operation within that structure or whether the times for these separate judgments implicate the importance of different structural principles and processes.

EXPERIMENT 1

In this experiment, we collected two independent 26 by 26 similarity matrices for the uppercase alphabet. One such matrix is derived from choice RT data and is a complete matrix. The other matrix is derived from direct ratings of visual similarity and does not contain diagonal entries.

Method

Subjects

Thirty students and staff at Yale University, 15 men and 15 women, participated as experimental subjects either for pay or for course credit.

Equipment

Each subject interacted with a Digital Equipment Corporation VT-50 video terminal that was operated from an adjoining

room by a PDP-11 computer. Each of these terminals was in a separate noise-shielded booth.

Stimuli

The uppercase English letters of the VT-50 terminal were used as stimuli. To produce each letter, a subset of cells within a 7 by 5 matrix was illuminated. The size of this matrix was 5 mm in height \times 2.7 mm in width. Every letter except "I" touched the four sides of this imaginary 7 by 5 matrix (see Figure 1).

Procedure

Three independent groups of 10 subjects were formed. We recorded the choice RTs of two of these groups, and we collected similarity ratings from the third. For all groups, a useful distinction can be made between the two letters that were compared on every trial. The *target* is the letter that occurred first (or, in the case of one-half of the subjects who rated similarity, the *top* letter). The *display* is the letter that occurred second (or the *bottom* letter). In the RT conditions, if the target and display letters matched, the subject was asked to make an "identity" or "same" response; if the letters differed, the subject was asked to give a "nonidentity" or "different" response. (In the case of similarity ratings, only nonidentical pairs of letters were presented, and subjects were told to rate the degree of similarity between members of a pair.)

RT: Fixed-target condition. This condition was characterized by the target's being kept fixed or unchanged for a long block of trials. In effect, the subject compared the fixed target letter in memory with each display letter. The 10 subjects in this condition were divided into two subgroups of 5. For one subgroup, the probability of a "same" trial was equal to that of a "different" trial (.50 vs. .50), whereas for the other subgroup a "same" trial was only half as likely as a "different" (.33 vs. .66). Aside from this shift in the probability of the two types of trials, the two subgroups were treated identically.

The subject sat in front of the video terminal and placed his or her middle and index fingers of the right hand on the "k" and "j" keys, respectively, of the terminal's keyboard. The computer announced on the screen that the next block was number x , where x ranged successively from 0 to 26. (The zeroth block was a practice block.) The computer also announced a target letter for the upcoming block. The target for a given block was chosen randomly from the complete set of alphabetic characters for Block 0, was again chosen randomly (after replacement) for Block 1, and was thereafter chosen randomly (without replacement) for all remaining blocks.

When prepared to begin a block of trials, the subject pressed the "return" key, and 4 sec later a letter appeared in the center of the screen. The subject was instructed to press the "j" key if the two letters—the target held in memory and the display on the screen—were identical, and to press the "k" key if they were nonidentical. The subject's response terminated the visual character, stopped a clock that measured the amount of time the character was displayed, and began an intertrial interval that varied randomly within the range of 1,000 to 1,500 msec. At the end of this random interval, another display character appeared, and so on until the end of the block. For an entire block, these display characters were the 25 letters of the alphabet not identical to the target (and these appeared either once or twice in the block depending upon whether the probability of a "same" trial was .50 or .33, respectively) plus 25 letters that were identical to the target. The order of these 50 (or 75) display characters was randomly permuted anew for each subject and for each block.

The subject rested $\frac{1}{2}$ min between all blocks except Blocks 13 and 14, between which blocks there was a 5-min break. Errors were monitored by the computer; and if the subject committed an error, the computer issued a "buzz" noise. Also, if the subject took more than 1,000 msec to make a response, the computer buzzed. Both of these types of trials (actual errors and long laten-

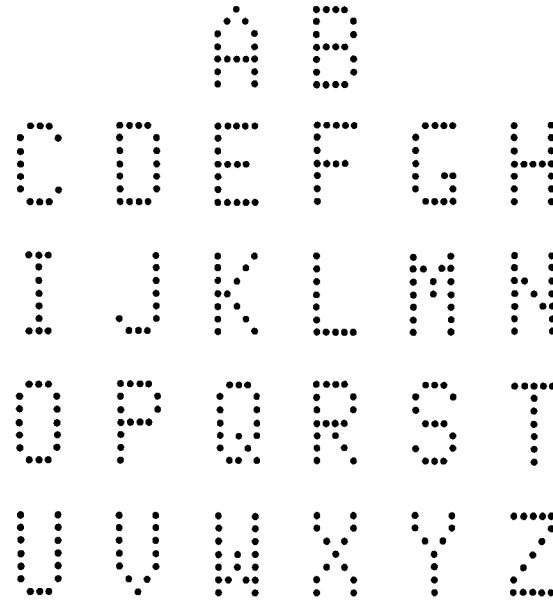


Figure 1. The 26 alphabetic stimuli that were used in the experiments. The dots appeared as illuminated cells of a CRT.

cies) were considered errors and were repeated later in the block until an error-free time was collected. Subjects were paid an amount inversely related to the number of "buzz errors" they caused, so subjects were encouraged to reach a compromise between speed and accuracy.

Subjects in the 50/50-probability subcondition completed the above design in about 1 h. They then repeated the above design in a second session of 1 h. Subjects in the 33/66-probability subcondition completed their version of the above design in about 1.5 h. These subjects did not repeat the design in a second session since, due to the change in probability, both subgroups now had equal numbers of entries in the off-diagonal cells of the resulting RT matrix.

RT: Varied-target condition. This condition was characterized by the target's being changed randomly from trial to trial within a block of trials. Unlike in the previous condition, the subjects in this condition actually saw both the target and display on every trial and did not have to remember the target for a long block of trials. This condition was very similar to the 33/66-probability subcondition just discussed, since the probability of a "same" trial was .33 in both. Only the procedural differences between this condition and the aforementioned subcondition will be listed here.

The major such difference was that instead of blocking the total set of 1,950 trials (per subject) by target letter, we instead randomly permuted these 1,950 trials, added 50 practice trials, and broke this whole series into 20 (arbitrary) blocks of 100 trials each. Since now the subject did not hold the same target in memory for a block of trials, we presented the target on each trial just before we presented the display letter. The sequence for an individual trial was as follows: The target appeared in the center of the screen for 300 msec and was succeeded by a 500-msec blank interstimulus interval. Then the display letter occurred in the center of the screen and stayed there until the subject's choice response. The subject's "same" or "different" response initiated an intertrial interval of 1,000 msec. As in the last condition, a clock timed the interval during which the display was present.

Each of the 10 subjects in this condition took a little over 2 h to complete the above design. There was a 5-min break between Blocks 10 and 11; between other blocks there was a $\frac{1}{2}$ -min rest.

Similarity-rating condition. The 10 subjects in this condition were asked to give visual-spatial similarity ratings for all possible permutations of two different alphabetic characters (i.e., 650 pairs). These subjects were divided into two subgroups according to the method by which stimuli were presented. For one subgroup, the method of presenting letters was like that used in the fixed-target condition discussed above. That is, while holding a target in memory, subjects witnessed a series of display letters that were to be judged with respect to the fixed target. For the other subgroup, the letters were also blocked by target; but in this subcondition, the fixed target actually appeared along with the display. That is, on each trial, the target was presented directly above the display, and approximately 10 mm separated the two letters in the vertical direction.

In each subcondition, there were 26 blocks with 25 trials per block. There was no practice block. For each block of trials, a particular letter of the alphabet—chosen randomly and without replacement—was deemed the target against which the 25 display letters in the block were to be compared. After the subject gave a rating to a particular target-display combination, a 1,000-msec intertrial interval began.

The target and display were always nonidentical, and the subject's task was to rate, on a 5-point scale (1 = very similar; 5 = very dissimilar), how visually-spatially similar the two letters were. The subjects were urged to use all rating categories and to confine their judgments to properties of the letters as depicted in the typography of the VT-50 terminal. Since the numeric keys 1-5 on the terminal's keyboard are at the upper-left of the keyboard, the placement of letters was changed from the center of the screen to a position above those keys being used to respond.

If subjects forgot either what the target was or how the scale worked, they could at any time in a block press the "t" key, for "target," or the "s" key, for "scale," and the computer reminded them of these facts. RT was not measured in this condition and errors were undefined, so the computer issued a "buzz" only if the subject gave a response that did not exist within the predefined scale. If a buzz occurred, the computer asked the subject to rate the current pair of letters again. The subjects took a little less than 1 h to complete the above design.

Results and Discussion

Mean correct RTs for the combined fixed- and varied-target conditions are shown in Table 1. Mean ratings of similarity for the similarity-rating condition are shown in Table 2. In both of these tables, rows represent targets and columns represent displays. The marginals for rows and columns, excluding the diagonal entries, have also been provided in the tables. Error data for the two RT conditions will be treated later in this section and in the General Discussion. Before addressing substantive questions concerning any of these data, we present preliminary analyses that justify collapsing the two RT conditions—and the two subconditions within fixed-target—into a single RT matrix, as well as collapsing the two subconditions of the rating condition.

Preliminary Condensation of Data

The two subconditions of fixed-target. These two subconditions were distinguished by a difference in the probability of same-different trials. The major effect of changing the probability of a "same" trial from .50 in one subcondition to .33 in the other was to produce a response bias for saying "same"

or "different" that influenced the absolute RT of these two responses. When "same" trials occurred with a probability of .50, the "same" response was an average of 26 msec faster than the "different" response, whereas when this probability was reduced to .33, the "different" response was 27 msec faster than the "same" response. The statistical reliability of this result was established by the following analysis of variance: For each of the 10 subjects, a mean "same" time and a mean "different" time were calculated for each of the 26 target letters. The subjects factor was then treated as a random effect and was nested within subconditions. The above interaction between subconditions and same-different response was significant [$F(1,8) = 7.40, p < .05$].

In addition, this analysis revealed that the factor of letters was a significant source of variation [$F(25,200) = 2.18, p < .01$], indicating that all letters were not equally good at playing the role of target. However, most important for present purposes is the observation that the letters factor did not interact with subconditions [$F(25,200) = 1.14$], nor was the Letters by Response by Subconditions interaction significant [$F(25,200) = .79$]. The lack of interaction is our primary justification for collapsing these two subconditions. The consistency of several correlational analyses to be reported later also supports the decision to collapse.

Fixed-target vs. varied-target. The question to be answered here is whether the varied-target condition can be considered, for all practical purposes, a "replication" of the data obtained in the fixed-target condition. An estimate of the reliability of the 325 off-diagonal cells (collapsed across the principal diagonal) from the fixed-target condition—an estimate computed by correlating the collapsed half-matrices of the two subconditions and then by correcting this correlation for the effect of averaging the two sets of numbers by using the Spearman-Brown formula (Cronbach, 1970)—is $r = .597$. This is the product-moment correlation coefficient we would expect to obtain by repeating the fixed-target condition a second time with 10 new subjects and by then correlating this new set of data (the collapsed half-matrix) with the old set. If, instead, we assume that the varied-target condition is the new set of data to be considered for purposes of replication, and if we then correlate this set with the set from the fixed-target condition, we obtain $r = .605$, a value that corresponds nicely to the value predicted by straight replication.

Moreover, if we compare the similarity structures of these respective data sets by using nonmetric multidimensional scaling as our structure-finding technique (Kruskal, 1964; Shepard, 1962), we achieve good correspondence. The method used to make this comparison (and many additional comparisons to be

Table 1
Mean Reaction Times (in Milliseconds) for Same-Different Judgments, Combined Fixed- and Varied-Target Conditions, of Experiment 1

Tar- get	Display																										RM	
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z		
A	444	460	445	462	440	441	440	449	444	444	444	426	435	455	442	438	453	455	453	442	429	443	459	453	442	441	434	445
B	454	444	441	473	438	440	450	439	431	432	426	427	446	422	454	472	476	482	456	442	438	428	435	441	438	415	443	
C	445	461	454	452	446	415	502	461	436	474	437	447	449	439	472	443	487	449	448	445	440	443	427	444	446	420	449	
D	489	465	454	457	456	449	441	446	430	432	408	446	431	439	503	467	499	448	449	458	442	448	441	413	441	446	449	
E	460	480	445	448	453	489	461	451	434	447	443	461	450	467	441	457	459	447	442	459	436	451	459	467	425	428	452	
F	446	431	426	441	468	455	457	467	478	483	467	468	478	451	439	452	434	462	428	465	457	445	451	483	451	455	455	
G	476	455	492	475	443	454	470	457	449	478	430	427	435	443	473	470	477	458	454	455	446	466	455	453	453	459	457	
H	452	461	444	458	451	450	486	467	462	459	457	419	502	483	439	432	442	451	452	458	471	464	527	439	454	452	458	
I	427	439	433	486	442	440	431	455	447	458	433	471	449	436	443	442	443	462	434	549	464	453	450	437	434	422	449	
J	456	431	442	433	425	464	422	453	472	457	457	465	447	440	430	445	453	458	443	456	442	462	465	458	454	446	448	
K	452	453	445	441	483	458	482	457	488	463	470	464	481	432	451	450	490	433	453	457	467	453	512	479	458	462	452	
L	455	456	437	436	469	454	464	447	482	503	459	444	444	456	436	480	459	447	443	465	457	464	453	453	448	453	456	
M	446	450	437	442	466	465	435	506	443	446	446	447	473	540	473	441	469	453	461	443	470	454	578	468	423	439	461	
N	439	440	441	445	444	433	454	493	465	481	469	455	586	467	454	423	441	473	442	454	474	475	527	449	448	458	462	
O	489	441	457	509	423	431	458	443	430	446	453	428	426	420	448	461	555	440	445	455	484	452	448	428	456	428	452	
P	437	458	435	441	432	468	447	445	419	424	419	424	430	419	416	450	451	447	460	439	441	447	461	430	419	435	437	
Q	471	456	471	476	435	433	478	444	437	469	431	449	456	456	532	460	460	455	459	438	470	467	464	439	445	451	457	
R	469	460	421	434	443	431	436	455	430	457	434	425	436	418	443	477	446	459	440	448	423	442	422	443	439	428	440	
S	447	461	447	443	452	442	462	439	436	439	447	441	429	437	449	435	451	444	447	438	454	446	445	449	439	448	444	
T	464	461	441	438	437	451	459	458	484	459	461	458	451	428	424	436	438	428	446	443	438	434	459	454	473	443	448	
U	460	439	464	441	452	443	453	454	453	448	440	454	460	483	482	443	463	453	439	461	456	595	454	439	465	431	458	
V	479	443	468	424	425	449	461	461	458	478	462	473	465	442	451	437	454	493	419	456	563	487	476	448	477	455	460	
W	431	446	447	429	464	433	444	504	463	447	452	438	566	476	453	453	456	443	456	439	451	453	482	474	474	468	458	
X	461	436	424	446	412	436	447	445	443	464	488	444	440	468	424	438	445	447	433	448	431	457	455	467	479	443	446	
Y	453	441	472	448	452	430	451	471	448	468	475	445	495	464	434	437	434	453	469	505	455	468	438	459	467	453	456	
Z	459	439	428	448	442	462	441	430	485	476	485	466	434	456	443	449	447	464	451	436	460	472	450	454	451	463	453	
CM	456	450	446	450	443	447	453	458	450	460	448	447	460	452	451	450	459	456	445	455	456	460	461	451	450	442		

Note—RM = row marginals, CM = column marginals; these are mean values computed without respect to diagonal entries.

Table 2
Mean Ratings of Visual Similarity for the Similarity-Rating Condition of Experiment 1

Tar- get	Display																										
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	RM
A		30	40	35	37	35	37	13	40	43	37	42	35	35	34	31	39	30	39	41	36	32	32	37	35	37	36
B	31		34	13	13	25	29	23	38	45	26	32	43	41	32	12	35	16	26	38	35	40	45	38	41	40	33
C	39	28		19	29	38	14	43	37	31	34	32	46	42	10	36	20	35	30	45	23	29	46	44	42	40	34
D	35	16	13		32	37	28	37	37	40	37	31	40	40	10	21	22	28	40	42	21	28	41	48	41	50	34
E	32	19	25	28		10	30	28	30	46	26	16	38	38	42	27	45	32	37	27	46	42	42	41	44	36	34
F	32	22	36	36	10		33	24	25	43	31	26	40	38	46	21	44	25	45	24	44	39	40	41	36	37	34
G	36	29	10	25	38	42		40	43	40	44	37	47	43	20	38	16	37	35	44	29	41	47	48	47	46	38
H	18	22	45	39	26	25	40		27	45	29	30	19	17	46	35	47	33	45	34	34	37	28	33	41	37	34
I	41	41	42	45	27	23	43	31		21	32	15	36	35	46	36	50	39	47	10	39	41	42	43	27	38	37
J	41	38	24	33	44	42	35	41	19		42	19	47	43	25	35	32	41	35	32	16	26	46	45	38	42	36
K	33	29	41	38	27	31	41	23	32	46		32	26	23	46	38	47	17	43	33	43	38	35	17	33	32	35
L	40	36	33	28	19	26	34	30	11	20	30		37	36	47	37	47	37	45	20	34	40	40	46	39	32	35
M	33	40	46	45	36	37	42	18	35	46	30	36		10	47	41	47	37	48	40	35	32	10	33	36	37	37
N	34	40	44	40	35	40	45	18	35	44	24	36	15		43	42	44	28	39	41	37	36	17	24	38	21	35
O	30	26	10	10	43	43	21	41	44	27	46	41	42	40		28	10	32	30	47	14	19	43	46	43	50	34
P	30	13	30	17	29	12	33	35	28	42	31	35	44	40	25		35	10	31	29	37	40	45	44	27	44	32
Q	35	28	13	17	45	44	17	41	45	30	42	38	47	44	10	31		32	35	45	21	31	46	46	44	46	36
R	25	15	38	29	31	24	35	32	39	44	20	41	35	34	40	10	40		38	41	43	43	43	36	43	43	35
S	39	25	24	35	40	38	27	42	46	36	41	42	45	41	28	33	36	32		44	40	41	44	31	42	31	38
T	40	40	40	41	22	19	40	32	11	28	35	17	39	35	50	29	48	37	47		43	38	39	42	21	33	36
U	26	36	24	24	43	45	32	32	43	15	45	27	40	41	16	40	25	40	38	44		10	25	39	32	46	34
V	27	41	32	41	44	44	43	33	37	29	31	38	35	27	30	41	39	46	41	41	10		16	26	21	42	35
W	33	40	44	46	34	37	46	22	38	46	29	39	12	17	45	43	47	40	44	41	19	18		30	26	35	36
X	36	38	46	44	37	35	44	24	40	45	15	36	28	25	47	42	42	33	41	38	40	26	28		18	24	36
Y	36	43	44	44	42	37	45	31	27	35	25	34	28	33	46	35	48	40	50	22	26	21	30	18		39	36
Z	42	40	43	46	32	38	42	33	42	38	26	36	30	24	44	44	48	39	33	35	45	38	31	24	40		38
CM	35	32	34	34	34	34	36	32	35	38	33	33	37	35	36	36	39	34	40	37	33	34	37	38	37	39	

Note—Mean ratings multiplied by ten. Hence, 10 = very similar and 50 = very dissimilar. RM = row marginals, CM = column marginals.

discussed shortly) can be described as follows. The collapsed half-matrix from the fixed-target condition and the analogous half-matrix from the varied-target condition were each separately scaled by the nonmetric scaling program KYST2. Nonmetric multidimensional scaling seeks to represent the similarities between objects as distances in a low-dimensional metric space. The regression function relating similarity to distance is constrained only to be monotonically decreasing, and therefore only the nonmetric or ordinal properties of the original similarity data are used in fitting this function. The “stress” of a scaling solution is a measure of how poorly the regression function fits the similarity data. Hence, “low stress” means “good fit.” For all the scaling applications reported in this paper, we have selected as options in the KYST2 program (1) the Euclidean distance formula, (2) stress formula 1, and (3) Kruskal’s “primary approach” to ties.

We obtained scaling solutions in four, three, and two dimensions for the fixed- and varied-target conditions. The two-dimensional solutions were then rotated into maximal congruence by a Procrustean rotation program, CONGRU, developed for this purpose by Olivier (Note 1). As a measure of how well the two rotated solutions overlapped, we chose the product-moment correlation coefficient that can

be computed between the coordinates of the corresponding letters in the two solutions. This coefficient for these 52 pairs of numbers—two coordinates for each of the 26 letters in each of the two solutions—was $r = .860$. This high correlation coefficient means that corresponding letters in the two sets of data occupied relatively similar spatial positions as points in the scaling solutions.

Finally, if we consider the “same” responses or diagonal entries, we obtain a product-moment correlation coefficient between the 26 letters in the two conditions of $r = .636$. This correlation, then, suggests that the diagonal entries—as well as the off-diagonal entries tested above—were fairly comparable.

The two subconditions of the similarity-rating condition. These subconditions were distinguished by the target’s either being held in memory or being visually present on every trial. The comparability of the two subconditions is easily established by noting an extremely high product-moment correlation between the collapsed half-matrix of one subcondition and that of the other subcondition: $r = .899$. Hence, a reliability estimate for the combined half-matrices is $r = .947$. In effect, we can conclude not only that these subconditions are comparable, but that the rating data are more reliable than the RT data. Many more RT observations per

cell would need to be collected in order to bring their reliability up to that of the rating data. Whether the two types of data—RTs and similarity ratings—are providing the same information will be treated next.

Validation of Off-Diagonal RT as a Measure of Interobject Similarity

Does discriminative RT, as assumed in the introduction, provide a measure of interobject similarity or confusability? The answer to this question is contained in a comparison of the data in Table 1 with those in Table 2. If the similarity structures—the pattern of high and low numbers in the matrices—are comparable for RT and for direct similarity ratings, then we will answer affirmatively. One way of comparing these two matrices is to correlate them directly. A product-moment correlation between the off-diagonal collapsed half-matrix of Table 1 and the analogous half-matrix of Table 2 gives a coefficient of $r = -.588$. The negative sign indicates that high similarity ratings (low scale values) are related to long RTs.

A better way of comparing the similarity structures in these two half-matrices is to use the technique described earlier in which multidimensional scaling solutions are obtained for each collapsed half-matrix separately and are then rotated into approximate, but “best,” congruence. If we perform these operations, we get the results portrayed in Figure 2. The vector connecting identical letters is directed toward the point whose position was established by the similarity-rating matrix (stress = .213) and is directed away from the point whose position was determined by the RT matrix (stress = .265). The product-moment correlation coefficient between the coordinates of corresponding points in the two solutions is $r = .785$.

The rather close spatial correspondence between analogous points in Figure 2 permits us to conclude that the similarity structures inherent in Tables 1 and 2 have a strong common component that we can attribute to the visual similarity of pairs of nonidentical letters. However, the two sets of data, strictly speaking, are not telling us exactly the same thing. That is, one set of data is not completely correlated with, or redundant to, the other set. One way of making this point is to remove from the RT data that part of their variance that can be explained by the rating data. Under the assumption of complete redundancy, the RT residuals should be insignificant and explainable in terms of error in measurement. In fact, if we compute these residuals for the collapsed half-matrix obtained from the fixed-target condition and for the collapsed half-matrix obtained from the varied-target condition, and if we then correlate these two sets of residuals, we obtain a signifi-

cant positive correlation ($r = .454$, $p < .01$). In other words, the residuals are reliable.

Off-Diagonal RT and the Proximity Constraints

It may seem peculiar or backward to have used multidimensional scaling as a structure-finding data-analytic technique when (1) this technique is based on a metric version of the three proximity constraints outlined in the introduction, and (2) a specific objective of this paper is to ascertain whether the RT data, in particular, conform to the proximity constraints. However, we as yet have offered no interpretation of the scaled results and, instead, have used the scaling technique only to *confirm* the presence of a common pattern or similarity structure in two sets of data. We now wish to show that, indeed, the RT data (and the similarity ratings) appear to conform to at least two of the three proximity constraints—Constraints 2 and 3—if we are willing to grant the reasonableness of Constraint 1, which states that an object is most similar to itself. Of the three constraints, the first is the most readily justified in terms of our definition of confusability as the prototypical similarity relation.

Two objects that are very similar have approximately equal similarity to any third object. It is easiest to think of this constraint in terms of two extreme conditions in which two objects are either maximally similar (without being identical) or maximally dissimilar. For the two extremely similar objects, the above constraint says that knowing how similar one of the two objects is to all other objects should provide much information about how similar the second object is to all other objects. That is, if we define the “similarity profile” of an object as the list of RTs or ratings for that object in com-

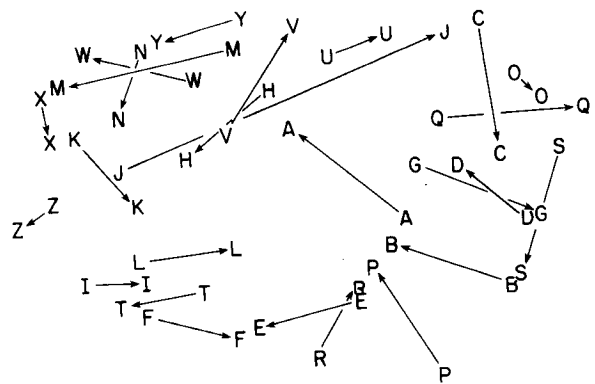


Figure 2. Approximate, but “best,” spatial superposition of two Euclidean, two-dimensional scaling solutions. The vector connecting corresponding points is directed away from the point whose spatial location was derived from the RT data of Table 1 and is directed toward the point whose position was derived from the rating data of Table 2.

parison with every other object in the set, then the similarity profiles of two similar objects should be similar. For the two extremely dissimilar objects, the above proximity constraint suggests that it is not possible to find a common third object that is highly similar to both objects; and, further, that if separate similar objects are found for each of the two original objects, these separate objects cannot themselves be highly similar. That is, except for certain degenerate cases, the similarity profiles of dissimilar objects should themselves be dissimilar.

One way of formulating this proximity constraint in a testable fashion is to ask whether a particular similarity structure, such as that derived by multi-dimensional scaling, based on either the raw RTs or raw similarity ratings of Tables 1 and 2 is comparable to an analogous similarity structure based on some derived measure of similarity that takes into account the profile similarity of an object's RTs or ratings over all objects. If common structures were found, we could conclude that the raw similarity measures conform to the above constraint, since they yield the same structure as do the derived measures that were designed to embody the properties of this constraint. One possible derived measure of similarity that takes into account an object's similarity profile is a correlation coefficient calculated between every pair of rows (or columns) of a matrix. This operation produces a new derived similarity matrix in which cell (i,j) contains the correlation coefficient (the "correlated row or column profile") between row i and row j (or column i and column j).

If we attempt to perform this operation on Tables 1 and 2, we are confronted with a problem concerning diagonal entries. Table 2 has no such entries, and Table 1 has entries that clearly violate the first proximity constraint. If we interpret diagonal RTs in the same way as off-diagonal RTs (i.e., high RT = high similarity), then the diagonal entries of Table 1 should be among the highest in that table. Instead, these entries are rather average.

We momentarily leave to one side the question of why the diagonal entries in Table 1 are as they are, and instead we assume that the "true" diagonal entries conform to the first proximity constraint. Thus, each true diagonal entry is larger than its largest row (or column) entry for the RT matrix and is smaller than its smallest row (or column) entry for the rating matrix. We next rank order separately each row (or column) so that each cell contains a value from 1 to 26, where the diagonal entry is always 26 for the RT matrix and is always 1 for the rating matrix. We then correlate all pairs of rows (or columns) to obtain a derived similarity matrix.

To give a concrete example of this derived measure of similarity, consider the letters W and M as targets in Table 1. Target W is assumed to be most

like display W, and this pair is given a rank of 26. Then, from the data, we see that display M is next most similar (has the longest RT), and that pair is given a rank of 25; display H is next most similar, with a rank of 24, N with rank 23, and so on to F, A, and D, which are given the lowest ranks because they have the fastest RTs to Target W. Target M is likewise given the rank of 26 for itself, and then, from the data, display W is next most similar and is given the rank of 25, followed by N with rank 24, H with rank 23, and so on to C, G, and Y, which have the fastest RTs to target M. These similarity profiles for these two letters are then correlated (actual value = .461), and this correlation becomes the derived measure of similarity. In this instance, M and W are not only rated as highly similar, and have the slowest RT for any pair for either letter, but also have similarity profiles that are similar, as indicated by the high positive correlation between their two similarity profiles. Two complete matrices of these correlations as derived measures of similarity were generated, one from the RT data of Table 1 and one from the similarity rating data of Table 2.

Having reached this step, we are again in the position of wanting to compare matrices for a common similarity structure. The scale-and-rotate procedure explained above can be adopted for this purpose. Specifically, for both RTs and ratings, the derived similarity half-matrices, based on product-moment correlations between pairs of rows after each row has been ranked, were submitted separately to KYST2. The resulting two-dimensional scaling solutions were each rotated into congruence with the appropriate set of points (obtained from either RTs or ratings) shown in Figure 2. For the RT data, the goodness-of-fit measure for these two solutions was $r = .880$. For the rating data, this measure was $r = .918$.

In conclusion, the similarity structure extracted from the raw measures of similarity and that extracted from the derived measures of similarity are comparable for both RTs and ratings. The fact that the derived similarities were chosen so as to embody the properties of the third proximity constraint suggests that the raw similarity measures inherently obeyed this constraint.

The similarity between two objects does not depend on the order of their comparison. The procedure outlined in the above subsection also permits a possible test of the second proximity constraint. This constraint requires that the similarity matrix should be roughly symmetric about its principal diagonal. One way of expressing such symmetry is in terms of a definition of derived similarity based on the correlation between similarity profiles, since it makes a difference whether we correlate pairs of

rows or pairs of columns of a matrix *only if* the matrix is not perfectly symmetric. That is, in the case of a symmetric matrix, an object's similarity profile across all row objects is identical to its profile across all column objects. Our interest is not in perfect symmetry but in approximate or "functional" symmetry, so we can phrase the following testable question: In terms of the resulting similarity structure, does it matter whether we correlate pairs of rows or pairs of columns in creating our derived similarity matrix?

If we repeat the analyses of the last subsection on pairs of *columns* rather than on pairs of rows, we can compare goodness-of-fit measures between various pairs of two-dimensional scaling solutions. These measures are reported in Table 3. Note that for both the RT data and the rating data, (1) the raw measures of similarity, (2) the derived measures of similarity based on row profiles, and (3) the derived measures of similarity based on column profiles, all point to a common similarity structure. For this reason, we conclude that, to a first-order approximation, the original similarity matrices conform to the second proximity constraint, which states that a similarity matrix should be symmetric.

Diagonal RT and an Object's Self-Similarity

The problem of comparing the absolute values of measures generated by different responses. The first proximity constraint, which concerns an object's similarity to itself, is the one we assumed in the foregoing analyses of the other two constraints. We should answer the question why we had to assume this constraint rather than finding the constraint in the data themselves. For the rating data, we did not ask subjects to compare the similarity of a letter to itself. We did not want to confuse subjects by asking them to make what might be a bizarre comparison. For the RT data, the only potential self-similarity data are contained in the diagonal entries, or the "same" responses; and, as mentioned earlier, these numbers are not large, but average.

However, there are many reasons why the absolute value of the diagonal "same" responses would be different from that of the off-diagonal "different" responses. A major reason, and one empirically demonstrated in this paper, is response bias. By changing the response probability of a "same" trial in the two subconditions of the fixed-target condition, we completely reversed the direction of the difference between average "same" and "different" RT. A second reason for expecting a difference in the absolute amounts of time is the fact that a given "same" pair, as a *pair*, gets repeated many more times than does a given "different" pair. Such repetition could easily facilitate the processing of "same" pairs.

Table 3
Correlation Coefficients Computed Between Coordinates
of Corresponding Points in Pairs of Two-Dimensional
Scaling Solutions

	Reaction Times			Similarity Ratings		
	Raw	CRP	CCP	Raw	CRP	CCP
Reaction Times						
Raw						
CRP	.880					
CCP	.902	.833				
Similarity Ratings						
Raw	.785	.796	.712			
CRP	.733	.758	.711	.918		
CCP	.772	.788	.764	.937	.930	

Note—CRP = correlated row profile, CCP = correlated column profile. Each coefficient is based on 52 pairs of numbers: two coordinates for each of 26 letters in each of two scaling solutions.

Differences between measures generated by the same response are interpretable. The lesson, then, is not to worry about absolute time differences between two different types of response, but to concern ourselves with relative time differences *within* a type of response. Essentially, the previously reported scaling solutions were all based on the off-diagonal "different" responses and, thus, conformed to this advice. Applying the same logic to the diagonal "same" responses, we immediately see that all diagonal entries are not nearly equal. In fact, the diagonal entries vary as much as or more than the off-diagonal row and column marginals (see Table 1). Two questions arise: (1) Is there any way in which the variance of these diagonal entries can be providing information about self-similarity? (2) Is there any order or pattern to the variance in the diagonal entries?

To attempt to answer the second question first, we have included information about "same" RTs in a multidimensional scaling solution derived from the "different" RTs. In Figure 3 the log area of each circle is directly proportional to the "same" RT for the encircled letter, and each circle is centered about the letter's independently defined position in a two-dimensional scaling solution.

The particular scaling solution used in this analysis varies slightly from that previously reported in Figure 2 for the RT data. That previous analysis was based on the collapsed half-matrix formed from the raw RTs in Table 1. The scaling solution in Figure 3 is based on a *ranking* of RTs. Specifically, (1) each row of the full matrix in Table 1 was rank ordered (without respect to the diagonal entry); (2) each column of the same original matrix was rank ordered (without diagonal); (3) these two matrices based on ranks were added together to form an average ranked matrix; and (4) this average matrix was collapsed into a half-matrix and submitted to

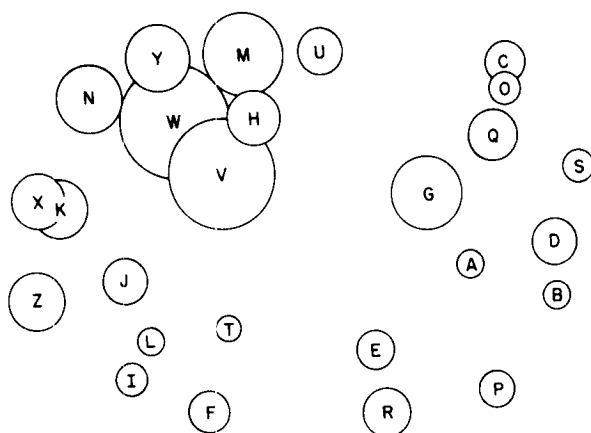


Figure 3. A representation that combines in a single picture information obtained both from "different" and from "same" RTs of Experiment 1. A letter's position in this picture was determined by its coordinates in a two-dimensional scaling solution derived from ranked-RT data of Table 1 (see text). The log area of the concentric circle encompassing a letter is proportional to the mean "same" RT for that letter (see the diagonal of the matrix in Table 1).

KYST2. The resulting two-dimensional scaling solution (stress = .281) is shown in Figure 3. In terms of our previous criterion of goodness of fit or spatial congruence, the scaling solution in Figure 3 agrees strongly with that for RTs in Figure 2 ($r = .989$). This agreement indicates that the common similarity structure evident in Figures 2 and 3 does not depend very much on differences in the absolute amount of time that subjects take to respond to different letters. Later we will isolate a subtle difference between these two solutions that can be explained in terms of the effect of scaling ranks vs. that of scaling raw RTs (even though we use a *nonmetric* scaling technique).

From this plot of circle size in relation to spatial location, it is obvious that there is a correlation between the size of circles and their position in two-dimensional space. In general, the larger circles can be found in or toward the upper-left corner of the scaling solution. Letters occupying this spatial region generally contain diagonal lines and generally do not contain curved lines (J and U are possible exceptions) or right-angled lines (H is an exception). The visual property of diagonality, or angularity, then, appears to be associated with a slow "same" response.

The mapping between self-similarity and RT. The next experiment directly addresses the question of whether the variance in the "same" RTs can be interpreted in terms of self-similarity. Note here that if we adopt the same direction of mapping between RT and similarity that we used for off-diagonal entries (i.e., high RT = high similarity), then we would

expect that letters in the upper-left corner of the similarity space shown in Figure 3—the letters with higher "same" RTs—should have greater self-similarity. However, it is very possible that the effect of similarity is different for "same" and "different" responses. Given that two objects are actually non-identical, then it should take *more time* to discriminate between them if they are highly similar; but given that two objects are actually identical, then it might well take *less time* to match them if they are highly similar (see, e.g., Posner, 1964).

Error Data

Error rates were generally low for subjects in the two RT conditions (recall that errors were undefined in the rating condition) and ranged from .1% to 6.5% for individual subjects, with a two-condition average of 2.1%. Forty-six percent of the 756 total errors occurred on trials in which subjects gave "same" responses incorrectly. The pattern of these errors and their relation to RT will be covered in the General Discussion.

EXPERIMENT 2

As mentioned previously, there is a problem with comparing the diagonal and off-diagonal entries of the matrix in Table 1 since each set of numbers was produced by a separate response. If the diagonal entries can, in fact, be interpreted as measures of self-similarity, we currently have no independent evidence with which to establish the direction of mapping between RT and self-similarity. Does a high "same" RT mean that an object is very similar or very dissimilar to itself? In the present experiment, we study a classification task in which the diagonal entries and the immediately surrounding off-diagonal entries are measures produced by the same type of response. This task permits us to evaluate directly (1) whether the similarity of an object to itself is greater than its similarity to any other object, and (2) whether all measures of intraobject similarity are necessarily greater than any measure of interobject similarity.

Choice RT is still the dependent measure of interest; but instead of discriminating and matching pairs of letters, the subject is now instructed to classify the two letters on the basis of their visual similarity. We make this a practicable task by judiciously choosing the subset of letters whose members the subject must classify. Referring to Figure 3, we see that the letters HMNW form a relatively compact cluster in the upper-left portion of the space whereas the letters CDOQ form a cluster in the upper-right portion of the space. Although intracluster distance is relatively small for both subsets, intercluster distance is quite large. Therefore, sub-

jects should find it reasonable to classify any pair of objects from the *same* cluster as similar and any pair from *different* clusters as dissimilar. We can then look at the amount of time it takes a subject to say that two identical letters are similar vs. the time it takes the subject to say that two non-identical letters are similar. If the first proximity constraint is, in fact, correct and if high visual similarity facilitates classification based on similarity, then the former times should be less than the latter.

Method

Subjects

Ten students at Yale, five men and five women, participated in this experiment for course credit and were included in the data analyses to be reported later. An 11th subject's data were not included due to this subject's high error rate—a rate twice that of the next most errorful subject.

Equipment

The same equipment used in Experiment 1 was used in this experiment.

Stimuli

The letters used as stimuli were created in the same way as the letters used in the first experiment. However, instead of studying the entire uppercase alphabet, we chose to study the subset of letters HMNWCDOQ which can itself be broken, on the basis of similarity, into the subsets HMNW and CDOQ.

Procedure

The subject sat in front of the video terminal with his or her middle and index fingers of the right hand placed on the "k" and "j" keys, respectively, of the terminal's keyboard. The computer announced on the screen that the next block was number *x*, where *x* ranged successively from 0 to 8. (The zeroth block was a practice block.) When ready to begin, the subject pressed the "return" key, and 4 sec later the first trial began.

The structure of an individual trial was as follows. A target letter appeared in the center of the screen and remained there for 300 msec. A blank interstimulus interval of 750 msec followed and was itself succeeded by the occurrence of a second letter, the display, in the same spatial location as the target. The onset of the display was simultaneous with the clock's start, and the display remained in view until the subject responded, which response also stopped the clock. The next trial began 2,000 msec after the subject's response.

The subject pressed the "j" key if the two letters were from the same cluster and pressed the "k" key if they were from different clusters. Subjects were told to make their judgments on the basis of visual similarity, and similarity was defined for them in terms of intracluster membership. Errors were monitored by the computer and were indicated to the subject by a "buzz" sound, but error trials were not repeated later in the session. As in Experiment 1, trials with latencies greater than 1,000 msec were considered errors.

Within a single block, there were 64 trials, and these trials represented the 64 possible permutations of pairs (including identical pairs) of the eight letters. The sequence of the 64 trials was randomly permuted anew for each subject and for each block. Between each of the nine blocks, there was a 1-min break, and the entire session took approximately 1 h for each subject.

Results and Discussion

Error trials were excluded from analysis. For each subject, a mean correct RT was calculated across blocks for each of the 64 permutations of letter

Table 4
Mean Reaction Times (in Milliseconds) for Similar-Dissimilar Judgments of Experiment 2

Display										
Mean		C	D	O	Q	H	M	N	W	Mean
II					I					
495**	C	497	546	537	550	554	561	557	561	568*
	D	538	490	509	590	581	574	577	568	
531†	O	502	526	493	518	584	542	571	554	
	Q	512	540	508	501	573	579	574	576	
III					IV					
594*	H	581	625	588	620	563	587	583	580	545**
	M	582	603	577	594	603	526	536	586	
	N	591	587	563	625	589	573	532	593	582†
	W	575	604	586	605	590	574	589	560	

Note—Summary numbers for each quadrant are reported adjacent to the quadrant. *Mean of all entries in quadrant.

**Mean of all diagonal entries in quadrant. †Mean of all off-diagonal entries in quadrant.

pairs. Next, for each of these 64 cells, a mean RT was computed across subjects. These latter means are reported in matrix form in Table 4. Note that the upper-left and lower-right quadrants (II and IV, respectively) contain the mean times for the "similar" response whereas the lower-left and upper-right quadrants (III and I, respectively) contain the mean times for the "dissimilar" response. In this table, rows represent targets and columns represent displays, as in Experiment 1.

Is an Object Most Similar to Itself?

We are primarily interested in the "similar" response, since diagonal and off-diagonal entries within this type of response can be compared directly (unlike the comparable entries from the RT matrix in Experiment 1 which contained "same" and "different" times, respectively). If we confine our attention to quadrants II and IV, we see that each of the eight diagonal entries is the smallest number in its respective row and column of its respective quadrant. An analysis of variance confirmed the reliability of this result across subjects: A mean diagonal RT and a mean off-diagonal RT were calculated for each subject and for each quadrant (i.e., II and IV). The F test showed that diagonal RT was less than off-diagonal RT [$F(1,9) = 16.31$, $p < .01$], but that this factor did not interact with quadrant [$F(1,9) < 1$]. In addition, the fact that the numbers along the diagonal are the smallest rather than the largest suggests that in this experiment the direction of mapping between RT and similarity is negative: Low RT means high similarity.

Another way of testing this assumption of an inverse mapping between RT and similarity in the present experiment is to correlate each subset of off-diagonal "similar" RTs with the corresponding "different" RTs from Experiment 1. Recall that in

Experiment 1 the off-diagonal RTs were shown to *increase* with increasing visual similarity. Therefore, if the RTs in the present experiment in fact *decrease* with increasing similarity, a correlation between analogous cells in the two experiments should be negative. The correlations for each of the two subsets, although not strong, are indeed negative ($r = -.430$ for CDOQ and $r = -.714$ for HMNW, based on the respective collapsed half-matrices for the corresponding subsets in the two experiments). With the inverse mapping between RT and similarity established, it is clear that an object is more similar to itself than to any other object, and the first proximity constraint is validated.

Is Every Object Equally Similar to Itself?

In this classification task, "similar" RT decreases with increasing visual similarity. Therefore, the results of Table 4 indicate that the self-similarities in the CDOQ subset are greater than those in the HMNW subset, since the mean RTs for the diagonal entries of quadrant II are less than those of quadrant IV [$t(9) = 2.32$, $p < .05$]. This result, in turn, may throw light on the "same" RTs or diagonal entries from Experiment 1, since in that experiment the four mean "same" times for the subset CDOQ were all *less* than the four mean times for the subset HMNW, just as they are in the present experiment. We tentatively conclude, then, that in Experiment 1 "same" RT also *decreased* with increasing visual similarity (just as "similar" RT did in the present experiment) whereas "different" RT *increased* with increasing visual similarity.

The fact that the first proximity constraint requires only that each diagonal entry in Table 4 be less than every entry in its row and column of its quadrant, and does *not* require that all diagonal entries be equal, provides the theoretical rationale for the following empirical finding: In Table 4, there are many off-diagonal entries in quadrant II that are less than the diagonal entries in quadrant IV. This result confirms a prediction that was considered in the introduction, namely, that two different objects may be more similar to each other than a particular single object is to itself (e.g., C is more similar to O than W is to itself). Furthermore, this result is troublesome for the metric version of the first proximity constraint that techniques of multidimensional scaling assume, since the distance between an object and itself would not be less than the distance between certain pairs of nonidentical objects (i.e., the distance between an object and itself could not be represented as a single point; see, also, Tversky, 1977).

A Problem Concerning the Purity of the Measures

However, earlier we mentioned the presence of a small but interpretable difference in the scaling solu-

tions for RTs in Figures 2 and 3 that may bear on this issue. Recall that the scaled points for RTs in Figure 2 were based on raw RTs, whereas the points in Figure 3 were based on RTs ranked by row and by column. Even though we used nonmetric multidimensional scaling in both cases, the two solutions were not completely identical. We adopted the procedure of first ranking by rows and by columns in an attempt to reduce mean row and column differences among the 26 letters (see the row and column marginals of Table 1). The necessity or even the desirability of such a correction depends on the nature of the information contained in the marginals of the raw RT matrix. Is the marginal entry for a given row or column providing information about inter- or intraobject similarity?

For example, if a particular row marginal of Table 1 appears to be large relative to other row marginals, this fact could be interpreted in one of two ways. Either the letter represented by that row could be quite similar to many of the 26 display letters (i.e., the row marginal correctly reflects that letter's high interobject similarity to other letters) or that letter could be an especially difficult letter to process *in general* (e.g., difficult to encode or to compare) irrespective of what other letter is paired with it. In the latter case, the large row marginal is not providing information about the letter's interobject relationship to other letters, but is, instead, providing information about the specific letter itself; and this latter type of information is related to what we have been calling intraobject similarity. The procedure of ranking by rows and by columns reduces the differences among the marginals of the resulting matrix and thereby controls for the possibility that intraobject variables will falsely manifest themselves as interobject effects.

The relevance of the above analysis is supported by the observation, to be discussed in the General Discussion, that both the row and column marginals of Table 1 are positively correlated with their respective diagonal entries. This same type of correlation can be seen in the present experiment in that the average RT in quadrant II, for both diagonal and off-diagonal entries, is considerably less than the average RT in quadrant IV. The issue we are raising here concerns the relationship between inter- and intraobject similarity in the choice task and how these two types of similarity influence the single response we measure: choice RT. In the General Discussion, we consider a recent multidimensional scaling approach to this problem.

For present purposes, the above analysis leads us to prefer the scaling solution in Figure 3 over the scaling solution for RTs in Figure 2. Since the scaling solution in Figure 2 is based on an analysis of *raw* RTs, the letters with long "same" responses and

with long RT marginals are relatively closer together than are letters with short "same" responses and with short RT marginals. Specifically, in Figure 2, the cluster HMNW (for the RT data) is more compact than that same cluster in Figure 3, whereas the CDOQ cluster in Figure 3 is more compact than that same cluster in Figure 2. Our claim is that in Figure 2 the closeness of letters in the HMNW cluster and the relative lack of closeness of the letters in the CDOQ cluster is determined in part by the fact that, in general, subjects took more time to process letters from the HMNW cluster than from the CDOQ cluster. In turn, this same argument helps to explain why the RTs in quadrant II of Table 4 are less than the RTs in quadrant IV.

GENERAL DISCUSSION

So far in this paper, we have concerned ourselves with properties of measurements and with whether certain RT measurements can be considered valid measures of similarity, where similarity is defined in terms of how subjects directly rate objects as similar and, also, in terms of certain general proximity constraints whose operation confers on data the ability to represent similarity relations. In this final section, we want (1) to return to the question of "similarity as confusability" by discussing the errors of Experiment 1; (2) to specify some of the properties of letters that appear to cause either high inter- or intraobject similarity; and (3) to discuss these results concerning visual similarity in terms of general models of choice RT tasks.

RT and Confusability

In the introduction, we postulated a link between confusability and discriminative RT. The argument was as follows: Given that confusability is an operationally (and ecologically!) reasonable indicant of what we mean by "similarity," and given that "confusion" errors in which subjects incorrectly respond "same" are positively correlated with discriminative RTs, it follows that the RT measure itself can be used as a measure of similarity and, indeed, *should* be used since it has several desirable properties that error measures frequently do not have. Following that opening argument, we provided other types of evidence that help to establish the legitimacy of using RT as a measure of visual similarity. We return now to an empirical test of our initial argument by analyzing the errors from Experiment 1.

As expected, the number of errors on trials in which subjects incorrectly said "same"—excluding errors due to long latencies—was positively correlated with correct "different" RT for a pair of letters. This relationship can be clearly seen by ex-

amining those off-diagonal cells of the cumulative error matrix that have the largest entries. For the following analysis, symmetric cells were combined, and all cells with an entry greater than four are reported here. The rank order of these cells from highest number of errors to lowest number of errors is: VU, MH, QO, WM, WN, WH, OD, TI, NM, GC, QC, NH. (Note that the order of letters in a pair is irrelevant in this ranking because we collapsed across the diagonal.)

However, the cells with the greatest number of confusions are also the cells that have the longest RTs. In fact, if we collapse the RT matrix of Table 1 about its principal diagonal and then isolate all cells that have a mean RT equal to or greater than 500 msec, we are left with the following letter-pair combinations (again in rank-order from highest to lowest): VU, WM, NM, QO, TI, WH, OD, MH, WN, XK. The overlap between this set of letter pairs and the set obtained with the error measure is remarkably clear. In addition, the present error data show one of the previously mentioned weaknesses of confusion data: Of the 325 cells of the error half-matrix, 171 contain an entry of zero; 91 contain an entry of one; 40 contain an entry of two; and only 23 contain an entry of three or more.

However, the errors along the diagonal—which represent those trials in which subjects incorrectly responded "different" when the letters were in fact the same—are somewhat more useful, since each of the 410 errors of this type belongs to one of only 26 cells, whereas the 346 errors we considered previously were distributed among 325 off-diagonal cells. If we correlate the diagonal entries of the cumulative error matrix (which contains numbers ranging from 6 for B to 24 for M) with the corresponding diagonal entries of the RT matrix in Table 1, we obtain a correlation coefficient of .539. This high positive correlation is plainly consistent with our previous interpretation of the diagonal "same" RTs as measures related inversely to intraobject similarity. Stated probabilistically, our interpretation is that subjects are more likely to call two identical objects "different" if these objects have low self-similarity.

In summarizing the results from the error data, we can say that whenever a sufficient number of errors occur, the pattern of errors confirms what the RTs have told us.

Properties of Letters and Letter-Pairs that Produce High Visual Similarity

What can we say about the properties of visual form that cause two nonidentical letters to be seen as similar? Figure 4 contains a hierarchical clustering analysis of the collapsed RT half-matrix from Table 1 based on the "maximum" or "compactness" method of incorporating objects into clusters (see

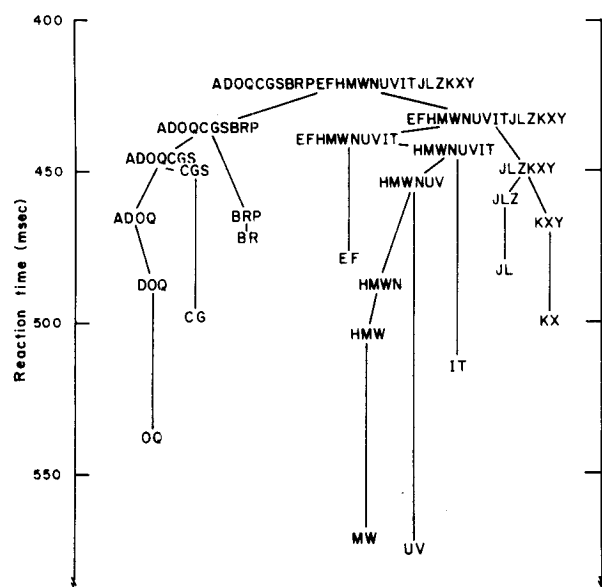


Figure 4. Hierarchical clustering solution, based on the maximum or compactness method of incorporating objects into clusters, for the RT data of Table 1. The RT scale on the left can be used to determine, for a given cluster, the minimum RT that was obtained between any pair of letters in that cluster.

Johnson, 1967). The physical—as opposed to topological—depth of a cluster in the tree structure is determined by the minimum RT (i.e., maximum distance) among all letter pairs in that cluster. For a two-letter cluster, the branch depth is equal to the “different” RT for that letter pair. If we scan along the lowest branches of the tree and isolate the two-letter clusters, we are immediately struck by a “theoretical confound”: Those pairs of letters that are most visually similar differ only by a distinctive feature *or* by a small discrepancy in physical overlap. That is, the most similar pairs of letters support indiscriminately either a feature-analytic or template-matching approach to a theory of representation.

Several pairs of letters make this point quite clearly: O and Q are rated as very similar and also give a slow RT, but the difference between them can be described equally well as one of an additional feature in the Q as contrasted with the O, or as one of high overlap between templates. C and G, also seen as highly similar, have the same dual type of difference, as do the letter pairs E and F, I and T, and even K and X. From such pairs, it is indeed impossible to differentiate a feature and a template theory of letter discrimination.

Strictly speaking, the similarity between the letters in pair JL and pair MW can be explained by a template theory only after “preprocessing” or “normalizing” operations have been applied to the letters to bring them into better registration (Neisser, 1967). Recently, however, the probable existence of just this

sort of operation has been effectively demonstrated for rotational transformations (Cooper & Shepard, 1973; Shepard & Metzler, 1971) and for size-scaling transformations (Bundesen & Larsen, 1975; Larsen & Bundesen, 1978). In light of these recent findings, we recommend that renewed attention be given to template-like theories of pattern perception.

It may be only a Pythagorean coincidence (cf. Miller, 1956), but a template-theory of similarity that asserts that letters in a pair (1) are compared by internally transforming one into the orientation, size, and what-have-you of the other, and then (2) are judged with respect to identity (or similarity) in terms of physical overlap, could easily explain why the RT and rating data of the present experiments appear to conform to the three proximity constraints. For, it is certainly true that a letter will overlap maximally with itself, will overlap equally with another letter irrespective of which is “on top,” and will necessarily overlap somewhat with a second letter if each of these overlaps moderately with a third.

In terms of general visual properties of letters that might confer high similarity on pairs, the most important such attribute in the present data appears to involve the “curved”-“noncurved” distinction, where the “noncurved” property can perhaps be further divided into “diagonality” vs. “rectilinearity.” Both the multidimensional scaling solution of Figure 3 and the hierarchical clustering solution of Figure 4 provide support for these distinctions. In the scaling solution, curved letters generally fall toward the right of Figure 3 and noncurved letters fall toward the left; and rectilinear letters typically occur toward the bottom left and diagonal letters occur toward the top left. In the clustering solution, the first major break between letters (at approximately 420 msec) occurs between a predominantly curved and noncurved set.

Other investigators have attempted more elaborate classifications of the “features” that combine to form alphabetic stimuli (Geyer & DeWald, 1973; Gibson, 1969; Kuennapas & Janson, 1969); and one of us (Garner, Note 2) has considered the effects of context and of stimulus variables, such as symmetry, on letter comparison processes. We currently remain uncommitted to any descriptive system for characterizing those properties of alphabetic stimuli that lead to confusion errors and to long RTs.

Directionally Reversed Mappings of Similarity in the Choice Task

A common distinction among various models of the binary-choice task is whether a model is a dual-process or single-process model. A dual-process model asserts that the “same” and “different” judgments are each handled by a separate processor, whereas a single-process model claims that one processor can produce both judgments (see Krueger,

1978; Nickerson, 1972). Aside from evidence based on differences in the absolute amount of time of the two responses, a major piece of evidence that has been used to support dual-process models is that there exist experimental variables that selectively influence one type of judgment and not the other type of judgment (e.g., Egeth & Blecker, 1971). The evidence from the present experiments of the reversed direction of mapping between similarity and RT for the two types of response might also be taken as evidence for a dual-process theory.

However, Krumhansl (1978) has recently shown how it may be possible to obtain this altered mapping by taking into account *both* the distance between *and* the density surrounding points in a multi-dimensional space (see, also, Hutchinson & Lockhead, 1977; Monahan & Lockhead, 1977). In effect, Krumhansl attempts to predict both "same" and "different" judgments from a representational medium based solely on psychological distance between nonidentical objects, and thus this model—although not intended as a process model—seems consistent with a single-process approach based on a concept of "discriminability." Referring to our Figure 3, we can state Krumhansl's predictions in terms of covariation between circle size and spatial location: The large circles (long "same" RT or low self-similarity) should reside in regions of the space where many letters are close together and confusable. In fact, the largest circles do occur in one of the densest regions of the space; but note that the dense set "FILT" comprises some of the smallest circles. Another possible test of average spatial density on the speed of the "same" response is provided by correlating both the mean "different" responses by rows (row marginals) and the mean "different" responses by columns (column marginals) with the mean diagonal "same" responses. These correlations are both positive ($r = .617$ and $r = .442$, respectively) and therefore suggest that a letter's location in a densely lettered area might influence the speed of the "same" response (but see our Discussion following Experiment 2). Finally, in Figure 3, there appears to be a trend for small circles to occupy peripheral locations on the spatial solution's boundary where, necessarily, there are fewer close points.

An alternative way of conceptualizing the observed covariation between circle size and spatial location in Figure 3 is by again noting the observation that the largest circles predominantly surround letters that contain diagonal lines. Therefore, a theory that states that letters containing diagonal lines—or other disadvantageous properties—are processed slowly could produce a result in which large circles reside in a dense region of the similarity space (assuming that letters containing diagonal lines are also seen as

fairly similar). Indeed, there is much behavioral and even neurophysiological evidence (Mansfield, 1974) that suggests that diagonal lines are not processed as efficiently as vertical or horizontal lines.

These two ways of describing or explaining a covariation between inter- and intraobject similarity are quite different conceptually. According to the *discriminability hypothesis*, a low value of self-similarity for a letter (high "same" RT) is caused by that letter's being close to and confusable with other letters. Intraobject similarity, then, is a by-product of interobject similarity. According to the alternative *specific-property hypothesis*, low self-similarity is caused by a letter's containing certain properties that are not processed as efficiently as other properties in a letter-pair comparison. By this latter view, intraobject similarity is largely determined by processing consequences relating to the object itself and is not determined by the interobject context in which the object appears. Indeed, the latter hypothesis predicts that intraobject similarity can adulterate measures of interobject similarity by raising and lowering these measures in a manner that depends on the difficulty or ease with which the letters in a pair are processed as individual letters.

CONCLUSIONS

With respect to interobject similarity, little doubt remains that "different" RTs can be interpreted as measures that are related in positive fashion to visual similarity. The following evidence supports this conclusion:

(1) RTs define a similarity structure that has an unmistakable component in common with the analogous structure defined by subjects' direct ratings of *visual* similarity.

(2) RTs, by conforming roughly to the two proximity constraints that pertain to nonidentical objects, show an internal consistency as *measures of similarity*, irrespective of the nature of the properties or dimensions (visual, in our case) that determine differential similarity.

(3) The highest RTs occur for pairs of letters that also lead to the greatest number of confusion errors of the type in which subjects incorrectly respond "same," and we assume that confusion with respect to an object's identity is the prototypical indicant of what one means by "highly similar."

With respect to intraobject similarity, we have provided the following evidence:

(1) All "same" RTs are not identical. Hence, if we hope to use these numbers as measures of intraobject similarity, we must accept the conclusion that all objects are not equally similar to themselves.

(2) When subjects judge letters as similar-dissimilar rather than as same-different, RT to say "similar"

to identical objects does conform to the first proximity constraint, which requires that an object be most similar to itself. Still, as found in (1) directly above, such intraobject similarities are not all equal.

(3) Our results indicate that "same" RT varies inversely with intraobject similarity: High RT means low self-similarity.

(4) When subjects make an error by calling two identical objects "different," they do so more frequently for objects that have low self-similarity.

What remains to be established is the processing connection between the two similarity constructs. People have forceful intuitions about what constitutes visual similarity between pairs of nonidentical objects, but intuition breaks down when pairs of identical objects are considered. A general model of same-different processing must accommodate parsimoniously both types of object similarity.

REFERENCE NOTES

1. Olivier, D. *Metrics for comparison of multidimensional scalings*. Draft, August 7, 1970.
2. Garner, W. R. *Letter discrimination and identification*. To appear in a forthcoming book, edited by Anne D. Pick, in honor of Eleanor J. Gibson.

REFERENCES

- BEALS, R., KRANTZ, D. H., & TVERSKY, A. Foundations of multidimensional scaling. *Psychological Review*, 1968, **75**, 127-142.
- BUNDESEN, C., & LARSEN, A. Visual transformation of size. *Journal of Experimental Psychology: Human Perception and Performance*, 1975, **1**, 214-220.
- COOPER, L. A., & SHEPARD, R. N. Chronometric studies of the rotation of mental images. In W. G. Chase (Ed.), *Visual information processing*. New York: Academic Press, 1973.
- CRONBACH, L. J. *Essentials of psychological testing* (3rd ed.). New York: Harper and Row, 1970.
- EGETH, H., & BLECKER, D. Differential effects of familiarity on judgments of sameness and difference. *Perception & Psychophysics*, 1971, **9**, 321-326.
- GARNER, W. R., & HAUN, F. Letter identification as a function of type of perceptual limitation and type of attribute. *Journal of Experimental Psychology: Human Perception and Performance*, 1978, **4**, 199-209.
- GEYER, L. H., & DEWALD, C. G. Feature lists and confusion matrices. *Perception & Psychophysics*, 1973, **14**, 471-482.
- GIBSON, E. J. *Principles of perceptual learning and development*. New York: Appleton-Century-Crofts, 1969.
- HUTCHINSON, J. W., & LOCKHEAD, G. R. Similarity as distance: A structural principle for semantic memory. *Journal of Experimental Psychology: Human Learning and Memory*, 1977, **3**, 660-678.
- JOHNSON, S. C. Hierarchical clustering schemes. *Psychometrika*, 1967, **32**, 241-254.
- KRUEGER, L. E. A theory of perceptual matching. *Psychological Review*, 1978, **85**, 278-304.
- KRUMHANS, C. L. Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, 1978, **85**, 445-463.
- KRUSKAL, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 1964, **29**, 1-28.
- KUENNAPAS, T., & JANSON, A.-J. Multidimensional similarity of letters. *Perceptual and Motor Skills*, 1969, **28**, 3-12.
- LARSEN, A., & BUNDESEN, C. Size scaling in visual pattern recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 1978, **4**, 1-20.
- MANSFIELD, R. J. W. Neural basis of orientation perception in primate vision. *Science*, 1974, **186**, 1133-1135.
- MILLER, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 1956, **63**, 81-97.
- MONAHAN, J. S., & LOCKHEAD, G. R. Identification of integral stimuli. *Journal of Experimental Psychology: General*, 1977, **106**, 94-110.
- NEISSER, U. *Cognitive psychology*. New York: Appleton-Century-Crofts, 1967.
- NICKERSON, R. S. Binary-classification reaction time: A review of some studies of human information-processing capabilities. *Psychonomic Monograph Supplements*, 1972, **4**(Whole No. 65), 275-317.
- PACHELLA, R. G. The interpretation of reaction time in information-processing research. In B. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition*. Hillsdale, N.J.: Erlbaum, 1974.
- POSNER, M. I. Information reduction in the analysis of sequential tasks. *Psychological Review*, 1964, **71**, 491-504.
- SHEPARD, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. I & II. *Psychometrika*, 1962, **27**, 125-140, 219-246.
- SHEPARD, R. N., & METZLER, J. Mental rotation of three-dimensional objects. *Science*, 1971, **171**, 701-703.
- SHEPARD, R. N., & PODGORNÝ, P. Cognitive processes that resemble perceptual processes. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes* (Vol. 5) *Human information processing*. Hillsdale, N.J.: Erlbaum, 1978.
- TVERSKY, A. Features of similarity. *Psychological Review*, 1977, **84**, 327-352.

(Received for publication February 12, 1979;
revision accepted April 30, 1979.)