Perception & Psychophysics
1975, Vol. 17 (6), 532-536

A comparison of methods for measuring the interletter similarity between capital letters

MORRIS B. HOLBROOK
Columbia University, New York, New York 10027

Measures of interletter similarity are often required in perception experiments. The most reliable and valid of the available measures appears to be Townsend's (1971) set of similarity parameters based on the Luce choice model. A simple mechanical measure offered a fairly strong prediction of the Luce choice-model similarity measure, as did a subjective rating measure based on the 10-point visual similarity ratings of eight subjects. By comparison, Gibson et al.'s (1963) matching-confusion matrix faired poorly, as did Gibson's (1969) distinctive feature analysis based on a letter pair's number of shared features. Distinctive feature analysis was significantly improved by substituting the feature set proposed by Geyer and DeWald (1973) or by weighting the features optimally via regression analysis. Such analyses suggested that figural curvature may be a particularly important perceptual feature, but in no case did these feature-analytic models predict the Luce measure as well as the mechanical or subjective rating measures.

Experimental work on the perception of printed verbal material often requires some measure of interletter similarity. Holbrook (1975), for example, used such measures as control variables in testing for the effects of verbal uncertainty on the recognition of intraword letter substitutions. As an indication of the importance of interletter similarity measures, several have been proposed from a variety of methodological perspectives. Even among the restricted set of measures that have been applied to all 325 letter pairs, there is a wide range to choose from. For lowercase letters, Kuennapas and Janson (1969) employed a subjective rating measure wherein subjects rated each letter pair for visual similarity on a scale from 0 to 100. By contrast, Dunn-Rankin, Leton, and Shelton (1968) devised a mechanical measure based upon the common surface area shared by two letters (relative to their combined remaining areas). Holbrook (1973) suggested that the validity of Dunn-Rankin's mechanical measure might be inferred from its correlation with Kuennapas and Janson's subjective rating measure ($r = .77$).

Most researchers have confined their attention to the similarity of uppercase letters, perhaps because capitals vary less than small letters in type style. In one of the earliest reported similarity measures, Gibson, Osser, Schiff, and Smith (Note 1) obtained two matching-confusion matrices for capital letters in tasks requiring 4-year-old subjects to give time-limited or timed multiple-choice matching responses. These investigators combined the cells for each letter pair within these matrices to obtain an overall matching-confusion matrix. In accordance with Gibson's (1969) perceptual theory of distinctive features, Gibson et al. (Note 1) also constructed interletter similarity measures based on the number of distinctive features shared by each pair of letters (relative to the total number within the pair). When such distinctive feature measures were compared with a mechanical measure based on the surface-area overlap, the authors concluded that feature analysis performed somewhat better than the mechanical measure in predicting matching confusion and interpreted this result as providing support for a feature-analytic (as opposed to a template-matching) model of letter discrimination. Their Table 3 showed, however, that the mechanical measure actually outperformed the preferred feature analysis for 12 of the 26 letters, a comparison which suggests that the advantages of the latter measure are far from overwhelming. Indeed, Gibson et al. admitted the weakness of their results and suggested that feature analysis might be improved by differentially weighting the features. But, though Gibson (1969) reiterated this suggestion, it does not appear to have been tested empirically. One of the purposes of the present paper, therefore, is to compare the results for Gibson's distinctive feature measure with those for an "optimal" version which uses regression analysis to develop a best least-squares fit to the letter-similarity criterion.

More recently, a valuable body of letter-similarity data has become available through the work of Townsend (1971a, b), who presented a tachistoscopic confusion matrix derived from the letter-recognition performance of six subjects. These data were then used to estimate the pairwise letter-similarity parameters for two mathematical models which assume that tachistoscopic confusion depends upon both interletter similarity and a response-bias factor. The most promising of these similarity parameters appeared to come from the Luce choice model.

Townsend's (1971a) Table 6 showed that the Luce choice-model similarity measures were remarkably consistent between two experimental conditions, one with and the other without noise ($r = .97$). Moreover, Townsend (1971b) found that separate choice-model estimates derived from the tachistoscopic confusion matrices of two individual subjects were correlated at $r = .99$ with the original group choice-model values and with each other. In addition, the three Luce choice-model measures were correlated at $r = .70$ or above with a mechanical measure based upon the physical overlap of each letter pair. This indication of the validity of the mechanical measure compares acceptably with that reported by Holbrook (1973) for lowercase letters ($r = .77$). It would appear, then, that any measure of interletter similarity that is claimed to be more valid than the mechanical measure should be more strongly related to the Luce choice-model measure than $r = .70$. The present study compares various letter-similarity measures in terms of this predictive criterion.

Geyer and DeWald (1973) attempted to establish a predictive relationship between distinctive features and Townsend's tachistoscopic confusion matrix. They compared the performance of several models using various feature lists and found that a set of features developed by Geyer himself appeared to perform better than Gibson's. Because the authors validated their model against the confusion matrix and not against the Luce choice model, however, it is difficult to determine how much its excellent fit depended upon (a) the response bias that is confounded with letter-similarity in the confusion matrix and/or (b) the fact that the goodness of fit reflected the model's ability to predict the diagonals (correct responses) as well as the interletter confusions (Geyer & DeWald, 1973, p. 479). Another purpose of the present study, therefore, is to compare Geyer's feature list with that of Gibson in their ability to predict various letter-similarity measures.

METHOD

The procedures for obtaining the measures of interletter similarity discussed above are presented in the following paragraphs. In each case, a triangular matrix representing each of the 325 pairs of capital letters was generated. The various similarity measures were then compared using simple and multiple regression analyses on a sample size of $N = 325$.

The Luce Choice-Model Similarity Measure (LCMSM)

The Luce choice-model similarity measure (LCMSM) was computed from the tachistoscopic confusion matrix (TCM) obtained by Townsend (1971a), using six subjects in the noise-free condition. The formula for these computations appeared in Townsend's Appendix:

$$LCMSM_{ij} = \left[\frac{TCM_{ij} \cdot TCM_{ji}}{TCM_{ii} \cdot TCM_{jj}}\right]^{1/2} \quad (i,j = 1, \cdots, 26) \quad (1)$$

where TCM_{ij} is the relative frequency of response j given stimulus i. Note that Equation 1 is intended to remove the effects of response bias, which are represented by another set of parameters in the choice model.

The Subjective Rating Measure (SRMHOL, SRMMIL)

Eight adult subjects provided ratings of each letter pair on an 11-point scale of visual similarity from "not at all similar" (0) to "extremely similar" (10). The subjects were members of a church music group and represented a wide variety of backgrounds. The letter pairs were presented in a 25-line list, 13 pairs to a line, as follows: A___B A___C A___D A___E A___F A___G A___H A___I ... etc. The subject indicated his similarity rating for each pair by placing a number from 0 to 10 on the line connecting its letters. To reduce the effects of order and fatigue, each subject began at a different point in the pair list, worked his way to the end, started back at the beginning, and continued through to the point at which he had begun. The experimenter instructed the subjects to spend 3 sec on each rating and said "next" at 3-sec intervals to provide a guideline as to how fast subjects should proceed. Each letter pair appeared only once in each list in the earlier-letter-first order. This procedure appears to be justified by Kuennapas' (1966) finding of no difference in similarity ratings resulting from the order of presenting letters within each pair.

An informal check on the reliability of this similarity measure for the eight subjects ($k = 1, \cdots, 8$) showed that the sum of these scores

$$SRMHOL_{ij} = \sum_{k=1}^{8} SRMHOL_{ijk} \quad (2)$$

was correlated with the individual $SRMHOL_{ijk}$s at an average of $r = .680$ (the range extending from $r = .512$ to $r = .812$). More importantly, SRMHOL corresponded closely to Miller's (1972, Note 2) subjective rating measure (SRMMIL), which was obtained by summing the 5-point visual similarity ratings of seven subjects. The correlation between SRMHOL and SRMMIL was $r = .702$, suggesting a fair degree of reliability for subjective rating measures, even with relatively few subjects.

Gibson's Matching-Confusion Matrix (MCM)

The matching-confusion measure (MCM) was computed directly from Gibson et al.'s (Note 1) two confusion matrices based on the time-limited (MCMI) and timed (MCMII) matching performance of 4-year-old children. These authors added the cells in the confusion matrices to obtain one overall measure for each letter pair:

$$MCM_{ij} = MCMI_{ij} + MCMII_{ij}. \quad (3)$$

A subsidiary analysis found support for this practice in the fact that MCM gave better predictions of the other letter-similarity measures than either MCMI or MCMII taken separately.

The Distinctive Feature Measure (DFMGIB, DFMGEY)

The first distinctive feature measure (DFMGIB) is drawn from the definition provided by Gibson (1969):

$$DFMGIB_{ij} = \sum_{k=1}^{12} F_{ijk}/TNF_{ij}, \quad (4)$$

where TNF_{ij} is the total number of Gibson's features contained by the letters i and j together and

$$F_{ijk} = \begin{cases} 1 \text{ when letters i and j share Gibson's feature } F_k \\ 0 \text{ otherwise.} \end{cases}$$

A second distinctive feature measure (DFMGEY) used the feature set proposed by Geyer and DeWald (1973). A key difference between Gibson's and Geyer's sets is that the latter permits a

feature (G_k) to be present more than once in the list representing a letter. Accordingly, G_{ijk} could take values greater than one if two letters both scored higher than one on a Geyer feature. Another difference between Gibson's and Geyer's feature lists is that Geyer specified 16 features, Gibson only 12. In computing DFMGEY, however, it was necessary to combine two perfectly correlated features and to eliminate two nondiscriminating features, leaving only 13 Geyer features as a basis for

$$DFMGEY_{ij} = \sum_{k=1}^{13} G_{ijk}/TNG_{ij} \quad (5)$$

where TNG_{ij} is the total number of Geyer features contained by the letters i and j and G_{ijk} is the number of Geyer features of the k-th type shared by these letters.

The Optimal Distinctive Feature Measure (OPDFMGIB, OPDFMGEY)

The optimal distinctive feature measures (OPDFMGIB and OPDFMGEY) were formulated as follows:

$$OPDFMGIB_{ij} = \sum_{k=1}^{12} a_{ijk} \cdot F_{ijk}/TNF_{ij} \quad (6)$$

$$OPDFMGEY_{ij} = \sum_{k=1}^{13} b_{ijk} \cdot G_{ijk}/TNG_{ij} \quad (7)$$

where the a_{ij}s and b_{ij}s are the weights for each feature which give an optimal least-squares fit in a regression analysis predicting some other letter-similarity measure. For the special case in which $a_{ij} = b_{ij} = 1$, Equations 6 and 7 reduce to Equations 4 and 5.

RESULTS

Table 1 shows the simple or multiple correlations between the various letter-similarity measures discussed above. Recall that Townsend's mechanical measure (MM) appeared to be a fairly valid approximation to the Luce choice-model similarity measure (LCMSM) as an index of interletter similarity (r = .70). This and all the correlations presented in Table 1 were significant at $p < .00001$ or beyond. Competing measures would have to improve upon the mechanical measure's performance if they were to be accepted as more valid indicants of similarity.

Table 1 shows that a subjective rating measure (SRMHOL) performed about as well as MM in predicting LCMSM: r = .650 (where the difference between this and r = .70 is not significant at p = .10 by Fisher's z test). Miller's subjective rating measure (SRMMIL) did not perform as well as SRMHOL in this respect (r = .539 vs. .650, a difference that is significant at p = .01 by the z test). SRMHOL was therefore adopted as the preferred subjective rating measure in interpreting the remaining results.

Gibson's matching-confusion matrix (MCM) was a weaker indicant of LCMSM than either MM or SRMHOL: r = .485. (The differences are significant at $p < .0001$ and $p < .001$, respectively.) Moreover, all four versions of the feature-analytic measure (DFMGIB, DFMGEY, OPDFMGIB, and OPDFMGEY) performed significantly less well than MM or SRMHOL in predicting LCMSM ($p < .0001$ for all four comparisons with MM; $p < .003$ for all four comparisons with SRMHOL). The comparative validity of SRMHOL was further suggested by the fact that it gave the best prediction of MCM (r = .531) and that, when compared with LCMSM and MCM, it was the most strongly related to DFMGIB (r = .406), DFMGEY (r = .620), OPDFMGIB (R = .520), and OPDFMGEY (R = .665).

In accord with Gibson's speculations, the optimally weighted feature-analytic measure (OPDFMGIB) performed somewhat better ($p < .05$) than the unweighted measure (DFMGIB) as an index of LCMSM (R = .398 vs. r = .256), SRMHOL (R = .520 vs. r = .406), or MCM (R = .456 vs. r = .319). (The z tests for these comparisons required an adjustment for degrees of freedom lost in computing multiple Rs.) Similarly, Geyer's OPDFMGEY improved slightly upon DFMGEY in predicting LCMSM (R = .497 vs. r = .439), SRMHOL (R = .665 vs. r = .620), and MCM (R = .615 vs. r = .437), though only the last of these comparisons reached significance at $p < .10$, thus suggesting that the improvements of OPDFMGEY over DFMGEY were due mostly to the extra degrees of freedom used in computing the individual feature weights.

Table 1
Simple or Multiple Correlations Between Interletter Similarity Measures

	LCMSM	SRMHOL	SRMMIL*	MCM	DFMGIB	DFMGEY
LCMSM	1.000	.650	.539	.485	.256	.439
SRMHOL	.650	1.000	.702	.531	.406	.620
SRMMIL*	.539	.702	1.000	.466	.393	.505
MCM	.485	.531	.466	1.000	.319	.437
DFMGIB	.256	.406	.393	.319	1.000	.489
DFMGEY	.439	.620	.505	.437	.489	1.000
OPDFMGIB	.398	.520	.496	.456	1.000	.582
OPDFMGEY	.497	.665	.525	.615	.645	1.000

Note—The correlations for OPDFMGIB and OPDFMGEY are multiple Rs; all others are simple rs. With N = 325, all correlations are significant beyond p < .00001.

**Negative values of SRMMIL are used to keep all correlations positive.*

Geyer and DeWald's (1973) preference for Geyer's own feature list was supported by the significant improvements in performance of DFMGEY and OPDFMGEY over DFMGIB and OPDFMGIB in predicting LCMSM, SRMHOL, and MCM. Even the smallest of these improvements was significant at $p < .07$ by a one-tailed test which adjusted for degrees of freedom lost in computing R.

Table 2 shows the regression coefficients with their t and p values for each feature in the OPDFMGIB and OPDFMGEY predictions of LCMSM. Table 3 shows the features that entered the equation significantly ($p < .10$) in a stepwise regression procedure. Little loss of predictive power occurred when reducing the 12 features used in OPDFMGIB to a set of 3 features (R = .373 vs. .398, n.s.) or when reducing the 13 used in OPDFMGEY to a set of 6 (R = .478 vs. .497, n.s.).

Few generalities emerge from a comparison of the features that appear to be most important in the two models except to note that, in both OPDFMGIB and OPDFMGEY, features representing the presence of curved segments (e.g., B, D, O, P, Q, or R) and straight horizontal lines (e.g., E, F, L, T, or Z) were significant contributors to the prediction of LCMSM. The same relative prominence of these curvature and horizontality features (F5, G4, F12, and G1) appeared when using OPDFMGIB and OPDFMGEY to predict SRMHOL and MCM. Even more dramatically, the contribution of the curvature features (F5 and G4) to the multiple regression prediction of $SRMHOL_k$ ($k = 1, \cdots, 8$) was stronger than that of any other feature for all but one of the eight subjects used to obtain SRMHOL. In these individual analyses, G4 was significant beyond $p < .00001$ for all eight subjects while F5 was significant at that level for all but three subjects (and significant beyond $p < .005$ for those three). It is further encouraging that the curvature feature corresponds to a dimension consistently identified by Gibson et al. (Note 1), Kuennapas (1966, 1967), and Townsend (1971a, b) in multidimensional scaling analysis. It appears, then, that figural curvature and (perhaps) horizontality are two of the more important features involved in perceived interletter similarity.

Table 2
Multiple Regression Results for OPDFMGIB and OPDFMGEY in Predicting LCMSM

		Coefficient	t Value	p Value
OPDFMGIB*				
F1	Straight: horizontal	−63.67	−.69	.491
F2	Straight: vertical	106.62	1.73	.084
F3	Straight: diagonal (/)	95.08	1.02	.307
F4	Straight: diagonal (\)	15.49	.21	.834
F5	Curve: closed	495.42	5.69	.000
F6	Curve: open, vertical	−79.07	−.42	.673
F7	Curve: open, horizontal	91.59	1.17	.242
F8	Curve: intersection	11.65	.17	.867
F9	Redundancy: cyclic change	−23.56	−.18	.856
F10	Redundancy: symmetry	35.76	.87	.388
F11	Discontinuity: vertical	175.61	2.20	.028
F12	Discontinuity: horizontal	425.81	2.70	.007
OPDFMGEY**				
G1	External: horizontal	309.03	3.26	.001
G2	External: vertical	219.26	5.54	.000
G3	External: slant (/) (\)	89.00	1.00	.316
G4	External: convex segment	293.56	7.39	.000
G5	Open: horizontal	−173.10	−.69	.492
G6	Open: vertical	−582.93	−1.64	.103
G7	Open: wedged, horizontal	185.34	1.66	.099
G8	Open: wedged, vertical	203.30	2.55	.011
G9	Open: internal protrusion	527.83	1.17	.245
G10	Open: intersection, internal	238.65	2.05	.042
G11	Open: bar horizontal	−42.37	−.25	.802
G12	Open: symmetry, vertical	131.22	1.76	.079
G13	Open: symmetry, horizontal	13.33	.16	.870

*$df = 312$ **$df = 311$

Table 3
Stepwise Regression Results for OPDFMGIB and OPDFMGEY in Predicting LCMSM

		Coefficient	t Value	p Value
OPDFMGIB*				
F5	Curve: closed	501.13	5.89	.00000
F11	Discontinuity: vertical	242.01	3.72	.00023
F12	Discontinuity: horizontal	371.12	2.85	.00463
OPDFMGEY**				
G1	External: horizontal	304.94	3.22	.00142
G2	External: vertical	206.67	5.30	.00000
G4	External: convex segment	276.66	7.37	.00000
G8	Open: wedged, vertical	255.94	3.40	.00077
G10	Open: intersection, internal	244.12	2.32	.02081
G12	Open: symmetry, vertical	140.78	1.93	.05439

*$df = 321$ **$df = 318$

DISCUSSION

The purpose of this paper has not been to develop new measures of interletter similarity, but rather to compare those advocated by other researchers. The adoption of the Luce choice-model measure (LCMSM) as the most valid index of interletter similarity suggested that the performance of a mechanical measure (MM) was as good as that of a subjective rating measure (SRMHOL) and that both measures performed significantly better than Gibson's matching-confusion measure (MCM) or any version of the feature analysis (DFMGIB, DFMGEY, OPDFMGIB, or OPDFMGEY). The relative failure of Gibson's matching-confusion measure might be explained by the fact that it was based on the behavior of nursery-school children (whose perceptual performance is undoubtedly very different from that of adults) and that, unlike the Luce choice-model measure, it made no correction for the confounding of letter similarity with the toddlers' response biases.

Both these factors seem more important than Gibson et al.'s use of a simplified type style since the features used to distinguish type faces are presumably not criterial for distinguishing letters.

It is more difficult to account for the failure of the feature-analytic models except to note (a) that their performance may be improved by optimally weighting the features and by substituting Geyer's feature list for Gibson's, and (b) that their performance is not significantly damaged by omitting all but three Gibson features or six Geyer features from the optimally weighted sets. These points suggest the desirability of developing a (theoretically justified) set of the few "most important" features defined more appropriately than in Gibson's original intuitive specification. Until such a refined set becomes available, it cannot reasonably be argued that feature analysis has improved upon the simpler template-matching and subjective rating measures of interletter similarity.

REFERENCE NOTES

1. Gibson, E. J., Osser, H., Schiff, W., & Smith, J. An analysis of critical features of letters, tested by a confusion matrix. In: A basic research program on reading. Cooperative Research Project No. 639, U.S. Office of Education, 1963.

2. Miller, L. K. Visual and auditory similarity-ratings for capital letters. Unpublished manuscript, University of Illinois, no date.

REFERENCES

Dunn-Rankin, P., Leton, D. A., & Shelton, V. F. Congruency factors related to visual confusion of English letters. *Perceptual and Motor Skills*, 1968, **26**, 659-666.

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: Meredith, 1969.

Holbrook, M. B. Note on validity of a mechanical measure of interletter similarity. *Perceptual and Motor Skills*, 1973, **36**, 298.

Holbrook, M. B. A study of communication in advertising. Unpublished doctoral dissertation, Columbia University, 1975.

Kuennapas, T. Visual perception of capital letters. *Scandinavian Journal of Psychology*, 1966, **7**, 189-196.

Kuennapas, T. Visual memory of capital letters: Multidimensional ratio scaling and multidimensional similarity. *Perceptual and Motor Skills*, 1967, **25**, 345-350.

Kuennapas, T., & Janson, A.-J. Multidimensional similarity of letters. *Perceptual and Motor Skills*, 1969, **28**, 3-12.

Miller, L. K. Letter recognition: Effects of interitem similarity and report requirements. *Perception & Psychophysics*, 1972, **11**, 252-256.

Townsend, J. T. Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, 1971, **9**, 40-50. (a)

Townsend, J. T. Alphabetic confusion: A test of models for individuals. *Perception & Psychophysics*, 1971, **9**, 449-454. (b)

(Received for publication November 22, 1974;
revision received February 7, 1975.)