Perception & Psychophysics
1984, 35 (1), 85-88

Notes and Comment

An empirical interletter confusion matrix for continuous-line capitals

A. H. C. van der HEIJDEN, M. S. M. MALHAS,
and B. P. van den ROOVAART
University of Leiden, Leiden, The Netherlands

For selecting a subset of letters to be used in a partial-report bar-probe task, we needed a confusion matrix for uppercase English letters. A search of the literature disclosed that only two matrices have been published that approximately met our requirements. Townsend (1971) generated such a matrix, using continuous-line capitals in a conventional tachistoscope and based on 3,900 observations (his condition 1 matrix). Gilmore, Hersh, Caramazza, and Griffin (1979) also published such a matrix, using uppercase letters presented as configurations of dots on the screen of a cathode-ray tube and based on 31,200 observations. In both studies, the letters were presented in a fixed position, that is, on, or slightly above, the fixation point, and the exposure duration was set to limit performance to 50% correct. Unfortunately, however, at present the value of both matrices is questionable.

Mewhort and Dow (1979) studied Gilmore et al.'s results and procedure thoroughly and came up with three rather problematic points. First, they convincingly argue that the exposure times used to limit performance to 50% correct are remarkably long (the range was 10 to 70 msec, and the mean over subjects was 33 msec). Second, they show that, for the main diagonal entries, there is a remarkably high correlation (−.873) between the number of dots in a letter and accuracy of identification. Third, they mention an extremely low correlation (.212) between the main diagonal entries of the matrices reported by Gilmore et al. and Townsend. These three facts are sufficient to give rise to serious doubts about the validity of Gilmore et al.'s matrix. Furthermore, the type font Gilmore et al. used (see their Figure 1) was rather different from the type font we planned to use (see our Figure 1).

In a reply, Gilmore and Hersh (1979) state that, to achieve the 50% accuracy level, they used a low intensity level, which resulted in long exposure durations. They explain the high correlation between identification accuracy and number of dots in a letter by suggesting that letters with few dots are simple figures composed of relatively few features, and that simple letters are relatively unique in their shape. The low correlation between their results and Townsend's can, in their view, be explained either in terms of the differences in type fonts used or in terms of a low validity and reliability of the values reported in one of the two studies. Since Gilmore et al.'s matrix was based on eight times as many observations as the matrix reported by Townsend, there is some reason to question the value of Townsend's data. Furthermore, because of the rather low number of observations, Townsend's matrix may not contain the reliable off-diagonal estimates of probability needed for subsequent calculations.

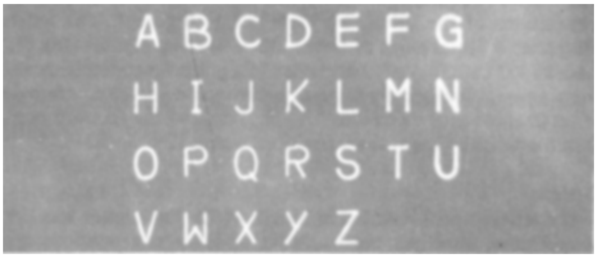

Figure 1. Representations of the letters used.

Because, for our task, a correct choice of letters was essential, we decided to collect sufficient data for a highly reliable confusion matrix that exactly met our requirements. Computer-generated continuous-line uppercase letters were used. The letters were not presented in a fixed position, but rather in one out of five possible positions on the circumference of an imaginary circle at 2.75 deg around the fixation point. In all further respects, we closely followed Gilmore et al.'s procedure. Because the results obtained would be of value to other investigators, we report the matrix here.

METHOD

Subjects

Twenty students from the University of Leiden participated as subjects and were paid Df 6.50/h for their services. All had normal or corrected-to-normal vision.

Apparatus

The letters were presented on a fast display screen (Vector General Graphics Display V11) equipped with P4 phosphor. Stimulus presentation and response registration were controlled by a PDP-11/34 computer.

Stimuli

The letters used were "white" Roman capitals on a dark background, generated with the Vector General hardware. With this

The authors thank D. J. K. Mewhort and R. Schreuder for their helpful suggestions. The authors' mailing address is: Department of Psychology, Unit of Experimental Psychology, University of Leiden, Hooigracht 15, 2312 KM Leiden, The Netherlands.

hardware, the letter-refresh time is programmable; no matter what the lengths of the constituting line segments are, a letter is drawn only once during this predefined interval. The refresh time used was 1 msec. The mean letter-writing time equaled 7.5 μsec. Representations of the letters are presented in Figure 1. The letters subtended a visual angle of about .32 × .48 deg at a viewing distance of 125 cm. They were presented in a position randomly chosen from among five equally spaced positions on the circumference of an imaginary circle (radius 2.75 deg) around the fixation point (a dim asterisk). One of these positions was at 12 o'clock.

Design

The letters were presented in blocks of 72. Each letter was presented three times per block. Order of letter presentations and order of letter positions were independently randomized per block by computer. For each subject, responses were collected on 4 days and there were five blocks per day. So for each subject, 60 responses per letter were obtained. Summed over subjects, this amounts to 1200 responses per letter. One day of practice preceded the 4 experimental days. During this practice, the 50% threshold was determined for each subject. During practice and during the experimental days, after each block of trials the exposure duration was increased (or decreased) by 2 msec if performance was lower than 35% (or higher than 65%). A 1-msec correction was applied when performance deviated between 5% and 15% from 50%. On the experimental days, stimulus durations ranged between 3 and 18 msec, with a mean across subjects of 6.42 msec.

Procedure

The subjects were run individually in a dimly illuminated room. The subject was seated in a chair behind a table, and then instructed to look at the display until he or she clearly saw the fixation point. The subject was told to press a button inserted in the table surface, 500 msec after which a letter appeared and the fixation point disappeared. He/she had to respond with a letter name on each trial. The experimenter entered the response into the computer through a terminal. Immediately after the response was recorded, the fixation point reappeared. After each block of trials, there was a rest period of about 2 min, during which the computer calculated the subject's performance and, if necessary, the experimenter adjusted the exposure duration. On each experimental day, the first block of trials was preceded by two blocks of 15 practice trials in order to provide time for adaptation and to reacquaint the subject with the experimental situation. An experimental session lasted about 45 min.

Results

The responses were summed over subjects, days, and blocks into a stimulus × response matrix. Since each letter was presented a total of 1,200 times, the entries in this matrix were divided by this value. The resulting confusion matrix is presented in Table 1. Each entry in this matrix gives the proportion of times that a stimulus letter (row headings) was identified as a response letter (column headings).

Discussion

As an indication that our efforts were really worthwhile, we briefly comment on some similarities and differences between our results and those published by Gilmore et al. and Townsend.

As far as exposure duration is concerned, our results clearly support Mewhort and Dow's argument. With an intermediate intensity level (unfortunately we have no means for exactly specifying the luminance level) and with the letters positioned 2.75 deg from the fixation point, a mean exposure duration of only 6.42 msec (a 3-18-msec range) was sufficient to limit

Table 1
Confusion Matrix

	Response																									
S	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	.633	.006	.002	.002	.004	.008	.006	.020	.016	.040	.052	.013	.005	.027	.006	.007	.004	.065	.011	.005	.001	.004	.008	.024	.007	.024
B	.018	.393	.007	.053	.013	.002	.150	.014	.001	.010	.007	.002	.002	.010	.037	.012	.064	.092	.087	.002	.009	.002	.007	.000	.002	.004
C	.002	.002	.712	.006	.049	.014	.101	.001	.006	.002	.003	.018	.000	.003	.017	.007	.013	.011	.007	.009	.003	.003	.000	.001	.002	.006
D	.004	.035	.006	.680	.004	.003	.010	.003	.002	.043	.002	.002	.002	.008	.042	.060	.023	.018	.014	.003	.016	.002	.004	.001	.006	.007
E	.004	.042	.022	.006	.350	.183	.034	.021	.038	.008	.020	.029	.002	.016	.007	.055	.008	.067	.023	.032	.007	.004	.002	.003	.010	.007
F	.011	.010	.011	.002	.057	.367	.013	.032	.052	.018	.021	.018	.008	.017	.003	.185	.002	.069	.007	.050	.008	.004	.007	.002	.013	.013
G	.007	.029	.050	.018	.018	.002	.599	.009	.002	.004	.001	.004	.001	.009	.051	.009	.103	.034	.023	.007	.013	.003	.002	.000	.000	.002
H	.015	.018	.002	.010	.007	.011	.008	.414	.035	.032	.015	.013	.101	.108	.007	.012	.006	.067	.004	.023	.018	.006	.052	.002	.013	.003
I	.002	.003	.011	.002	.015	.025	.002	.008	.608	.082	.004	.068	.002	.003	.002	.007	.001	.004	.007	.093	.006	.004	.001	.004	.016	.020
J	.008	.004	.003	.005	.002	.007	.004	.026	.078	.647	.012	.024	.005	.013	.004	.007	.002	.007	.006	.021	.046	.013	.015	.006	.018	.018
K	.002	.002	.009	.002	.008	.023	.002	.010	.014	.015	.385	.011	.011	.014	.000	.015	.001	.047	.005	.023	.006	.022	.022	.173	.114	.063
L	.008	.005	.013	.004	.019	.004	.001	.022	.147	.052	.014	.537	.005	.011	.005	.008	.005	.008	.009	.045	.022	.010	.014	.004	.011	.016
M	.005	.011	.005	.003	.006	.009	.004	.151	.006	.012	.055	.003	.437	.072	.006	.018	.005	.052	.006	.024	.014	.007	.039	.027	.021	.002
N	.019	.012	.001	.006	.001	.002	.004	.022	.005	.008	.108	.013	.027	.494	.001	.004	.003	.108	.013	.006	.005	.020	.026	.060	.028	.003
O	.001	.016	.098	.092	.002	.004	.153	.001	.004	.009	.001	.003	.000	.004	.383	.020	.142	.013	.023	.003	.014	.002	.002	.001	.004	.003
P	.018	.023	.017	.032	.018	.084	.015	.022	.002	.008	.005	.005	.006	.003	.015	.567	.007	.088	.021	.006	.013	.004	.005	.003	.006	.007
Q	.005	.014	.106	.083	.001	.003	.183	.003	.001	.011	.002	.007	.001	.002	.252	.004	.249	.011	.028	.000	.016	.003	.004	.001	.003	.006
R	.022	.022	.015	.019	.015	.036	.046	.040	.000	.006	.013	.002	.005	.023	.023	.131	.042	.444	.044	.011	.008	.002	.007	.007	.013	.003
S	.007	.068	.043	.010	.074	.020	.204	.006	.013	.010	.015	.008	.002	.013	.016	.017	.027	.101	.301	.018	.006	.004	.002	.003	.004	.009
T	.004	.002	.003	.002	.004	.069	.002	.013	.168	.036	.007	.031	.004	.012	.002	.025	.001	.011	.002	.520	.007	.013	.002	.006	.032	.022
U	.004	.003	.012	.013	.002	.002	.037	.011	.004	.028	.005	.021	.012	.025	.021	.007	.014	.005	.002	.006	.628	.049	.082	.002	.006	.000
V	.001	.002	.002	.005	.000	.013	.004	.004	.025	.021	.018	.007	.002	.025	.006	.013	.002	.015	.003	.013	.047	.528	.043	.017	.182	.001
W	.041	.014	.001	.005	.003	.007	.006	.068	.004	.013	.157	.009	.089	.083	.007	.020	.007	.127	.008	.002	.013	.005	.187	.100	.013	.010
X	.002	.001	.000	.001	.002	.005	.002	.002	.009	.016	.086	.007	.005	.023	.001	.003	.000	.022	.004	.013	.002	.037	.011	.503	.211	.034
Y	.002	.001	.000	.002	.000	.023	.000	.005	.020	.023	.012	.004	.003	.016	.001	.022	.001	.007	.002	.031	.004	.027	.007	.059	.657	.074
Z	.009	.001	.002	.002	.007	.006	.002	.003	.011	.028	.015	.013	.001	.004	.002	.003	.001	.007	.010	.014	.002	.006	.003	.027	.058	.760

Note–S = stimulus.

performance to 50% correct. Because Gilmore et al. presented the letters at the point of fixation and nevertheless needed a mean exposure duration of 33 msec (a 10-70-msec range) to limit performance to 50% correct, it follows that they must have employed an extremely low intensity level. Therefore, it is likely that Gilmore et al. measured the visual system's contribution to the process of letter recognition only while it was operating in its scotopic (i.e., rod) range. (Unfortunately, Townsend doesn't mention the exposure times used.)

Because in Townsend's, Gilmore et al.'s, and the present study rather similar procedures were used—the essential factor being that in all three studies performance was limited to 50% correct—it is possible to explicitly compare subsets of the data. Of special importance here are the proportions correct in the main diagonal entries of the matrices, because these values measure the difficulty of the letters relative to each other. We restrict our further discussion to these proportions correct.

In order to establish the correspondence between óur matrix and the other two, we treated each letter as the unit of analysis and calculated the correlation between the proportions correct. The correlation between Townsend's data and the data here presented was .477 ($p < .05$) and the correlation between Gilmore et al.'s proportions correct and the data in Table 1 was .503 ($p < .01$). While these correlations are appreciably higher than the correlation of the proportions in Gilmore et al.'s and Townsend's main diagonals (.212; n.s.), the values are not too impressive. The variation in the data here reported explains only about 25% of the variation in the other two matrices.

Further inspection of the differences between the main diagonal proportions in Townsend's matrix and the matrix reported here revealed no obvious trends. There were, however, clear and systematic differences between Gilmore et al.'s data and the present data. Simple letters (e.g., I, L, and J) in Gilmore et al.'s matrix had appreciably higher proportions correct and complex letters (e.g., A, B, and M) appreciably lower proportions correct than does our matrix.

In order to bring out this feature of the data more clearly, we resorted to Mewhort and Dow's technique. We counted the number of dots for each letter in Gilmore et al.'s (1979) representation of the characters (their Figure 1), and determined the linear function relating the *differences* between the main diagonal proportions in Gilmore et al.'s matrix and ours (ΔP) with the number of dots in each of Gilmore et al.'s characters (n). The obtained function was

$$\Delta p = .836 - .056n \ (r = -.709, p < .01).$$

Because the mean number of dots ($\bar{n}$) used by Gilmore et al. equals 14.77, this function can also be written as

$$\Delta p = .014 - .056(n - \bar{n}). \qquad (1)$$

We determined the corresponding function for Gilmore et al.'s and Townsend's matrices and obtained

$$\Delta p = 1.029 - .068n \ (r = -.758, p < .01),$$

which can be written as

$$\Delta p = .026 - .068 \ (n - \bar{n}). \qquad (2)$$

Functions 1 and 2 clearly show how Gilmore et al.'s main diagonal proportions deviate from those reported by Townsend and those reported in this paper. Characters composed of a below-average number of dots are reported much more accurately and characters composed of an above-average number of dots much less accurately than the corresponding continuous-line letters in the other two studies. Each additional dot seems to impair performance in Gilmore et al.'s study by the substantial proportion of about .06.

As stated in the introduction, Mewhort and Dow have already pointed to the remarkably high correlation in Gilmore et al.'s matrix (−.873) between the number of dots in a letter and its identification accuracy, as given by the *absolute* proportions correct in the main diagonal. They suggest that there was possibly something wrong with the dot-refreshment procedure used. If dots are refreshed continuously, a letter formed from a small number of dots would be brighter than one formed with a large number of dots. In their reply, however, Gilmore and Hersh responded that there was nothing wrong with the dot-brightening algorithm and suggested that the reason for the high negative correlation between number of dots and identification accuracy was that letters with few dots are simple figures composed of relatively few features and that the simpler letters are relatively unique in their shape. In other words, in their view, number of dots is a measure for letter complexity and it is this complexity that is responsible for the correlation, not the number of dots as such.

At first sight, Functions 1 and 2 seem to support Mewhort and Dow's position and not Gilmore and Hersh's interpretation. A rather high negative correlation is also found between numbers of dots and differences in proportions between Gilmore et al.'s matrix and the other two matrices, that is, after subtracting the effects due to letter complexity, as estimated in Townsend's (1971) and in the present study. Because the subtraction procedure ought to remove a letter-complexity effect—simple letters remain simple and unique letters remain unique, whether presented as dots or as continuous lines—Gilmore and Hersh's interpretation is at least not sufficient.

But it appears that Gilmore and Hersh's interpretation is partly valid. Gilmore et al. (1979, p. 425) reported that, under the viewing conditions they employed, the letters composed of dots appeared like green continuous-line figures against a dark background. So we do not have to interpret n as number of dots, but can also interpret this parameter as a rough measure of letter complexity, that is, as the number of dots that would be needed to write each continuous-line letter in dots. If we interpret n in this way, it makes sense to relate not only Gilmore et al.'s results (P_G) to n, but also Townsend's (1971) results (P_T) and those reported here (P_p). Using the same format as for Equations 1 and 2, the linear relations are, respectively:

$$P_G = .514 - .079(n - \bar{n}) \quad (r = -.873,\ p < .01) \quad (3)$$

$$P_P = .499 - .023(n - \bar{n}) \quad (r = -.424,\ p < .05) \quad (4)$$

$$P_T = .488 - .011(n - \bar{n}) \quad (r = -.296,\ \text{n.s.}). \quad (5)$$

Equations 3, 4, and 5 suggest similar relations between identification accuracy and n in the three studies, although different in strength. Because there was nothing wrong with our letter-refreshment algorithm and because Townsend used conventional tachistoscopic exposures, this similarity in relations suggests that there is no reason to assume that there was something wrong with Gilmore et al.'s dot-refreshment algorithm. Rather, it appears that n really measures an aspect of letter difficulty, as suggested by Gilmore and Hersh (1979). It furthermore appears that this aspect has a strong effect in Gilmore et al.'s study, an intermediate effect in the present study, and a rather small effect in Townsend's.

Another related way of showing that n really measures an aspect of letter-complexity is by means of partial correlations between proportions correct in the three matrices, that is, correlations with the effect of n eliminated. If n is a relevant variable, eliminating n results in partial correlations that are lower than the original correlations. If n is an irrelevant variable, introducing random variation only, as suggested by Mewhort and Dow, then eliminating the effect of n results in a higher partial correlation. The partial correlation between Gilmore et al.'s data and the data reported here equals .300 (the correlation was .503) and between Gilmore et al.'s data and Townsend's, −.102 (.212). The partial correlation between the data reported here and Townsend's equals .406 (.477). So it again appears that letter complexity, as measured by n, is indeed a relevant variable, responsible for a part or the whole of the original correlations.

The problem that remains is how to explain the differences in letter-complexity effects in the three studies. (It is these differences that are largely responsible for the rather low original correlation between the three studies.) The beginning of an explanation can be found if we look at the differences between Townsend's study and the present one. Of importance is the fact that Townsend presented the letters 10 min above the fixation point, whereas in our study the letters were positioned on the circumference of a circle, 2.75 deg from the fixation point. It is reasonable to assume that the decrease in visual acuity and resolution power and the increase in severity of lateral masking from fovea to periphery more strongly impairs recognition performance with complex letters than with simple letters. It is even possible that, with the luminance levels used in the two studies, differences in relative contributions of the cone system and the rod system at 10 min and at 2.75 deg can account for the difference in results.

In line with the above argument, we have to suppose that Gilmore et al. presented their letters still further in the periphery. Actually, however, they presented the letters at the point of fixation. We therefore guess that they found the far periphery near the center of the eye, that is, we guess that they measured exclusively the contribution of the rod system to the process of letter recognition.

The above analysis underlines Mewhort and Dow's conclusion that one has to be cautious in interpreting and using Gilmore et al.'s data. The matrix presented in this paper and, for some purposes (see our introduction), the matrix reported by Townsend, can be used by investigators as valid and reliable instruments.

REFERENCES

Gilmore, G. C., & Hersh, H. Multidimensional letter similarity: A reply to Mewhort and Dow. *Perception & Psychophysics*, 1979, **26**, 501-502.

Gilmore, G. C., Hersh, H., Caramazza, A., & Griffin, J. Multidimensional letter similarity derived from recognition errors. *Perception & Psychophysics*, 1979, **25**, 425-431.

Mewhort, D. J. K., & Dow, M. L. Multidimensional letter similarity: A confound with brightness? *Perception & Psychophysics*, 1979, **26**, 325-326.

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

(Manuscript received October 3, 1983;
accepted for publication October 5, 1983.)