Journal of Mathematical Psychology **45**, 99–114 (2001)
doi:10.1006/jmps.1999.1296, available online at http://www.idealibrary.com on IDEAL®

Mathematical Studies of the Information in the Stimulus–Response Matrix

Elad Sagi, Willy Wong, and Kenneth H. Norwich

University of Toronto

This paper considers the information transmitted in absolute judgments as encoded in a stimulus–response matrix (e.g., see Garner and Hake, 1951). When transmitted information is plotted against the number of stimulus categories in the matrix, one obtains a curve that increases monotonically toward a plateau, which is the maximum information transmittable per stimulus for the particular range of stimuli employed. We demonstrate that although the maximum information transmitted is an attribute of the stimulus continuum itself, the shape of the curve is an empirical property of the stimulus–response matrix, which is determined, in part, by maintaining a constant stimulus category width. Therefore, in principle, each curve of information transmitted vs number of stimulus categories can be determined by a single point: the rightmost point on the graph.

INTRODUCTION

The stimulus–response or confusion matrix is a convenient means of summarizing data involving absolute judgments. We can formalize the construction of these matrices in the following way. A set of stimuli, S, consists of discrete and distinct stimulus values, $s_1, s_2, s_3, ..., s_T$, where $s_i < s_{i+1}$. S may be grouped into subsets of adjacent stimuli, $X_1, X_2, X_3, ..., X_j, ..., X_m$, $m \leqslant T$, where, for example, $X_1 = \{s_1, s_2, s_3\}$, $X_2 = \{s_4, s_5, s_6, ..., s_{10}\}$, etc. X_j may be referred to conveniently as the set corresponding to the jth *stimulus category*. In exactly the same way, the set of responses, $R = \{r_1, r_2, ..., r_T \mid r_k < r_{k+1}\}$ may be grouped into subsets or categories $Y_1, Y_2, ..., Y_k, ..., Y_n$, $n \leqslant T$, where Y_k is the set corresponding to the kth *response category*.

This work has been supported by a grant from the Natural Sciences and Engineering Research Council of Canada to K. H. Norwich. Special thanks to Dr. Jack Baird and Dr. Tony Marley for their insightful comments and corrections.

Address correspondence and reprint requests to Dr. Kenneth H. Norwich, Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, Ontario, Canada, M5S 3G9.

0022-2496/01 $35.00

We may note here that the greatest stimulus category cardinality is obtained when $m = T$, that is when $X_1 = \{s_1\}$, $X_2 = \{s_2\}$, ..., $X_m = \{s_T\}$. Similarly, the greatest response category cardinality occurs when $n = T$ and $Y_1 = \{r_1\}$, ..., $Y_n = \{r_T\}$.

It is necessary to distinguish clearly between two types of tasks: the task of identification and the task of categorization. The distinction between the two has been clearly enunciated by Lacouture, Li, and Marley (1998). "In an identification task, each stimulus is assigned a unique 'correct' response." This process of identification finds expression in the set of greatest stimulus and response cardinality, $m = n = T$. In the process of categorization, where m is not necessarily equal to n, we suppose that stimulus s_i lies in stimulus category X_{p_i} and response r_i lies in response category Y_{q_i}. Then the "correct" response to X_{p_i} is Y_{q_i}, as we see illustrated in Fig. 1. As Lacouture *et al.* remark, "identification is a special case of categorization." We shall proceed in exactly this fashion, developing our equations for the general task of categorization, *which includes identification.*

If the stimulus continuum is divided into m discrete categories, and the response continuum into n discrete categories, the resulting $m \times n$ matrix may be called a stimulus–response matrix. The matrix element N_{jk} is equal to the number of times a stimulus category X_j is identified with response category Y_k (Table 1). Garner and Hake (1951) showed how the subject's performance in an absolute identification experiment could be measured by the mutual or transmitted information, I_t, calculated using Shannon's famous equation,

$$I_t = H(Y) - H(Y \mid X), \tag{1}$$

where $H(Y)$ is the response entropy and $H(Y \mid X)$ is the equivocation. Both of these quantities can be evaluated from known values of N_{jk}.

For a given matrix, the value of the transmitted information will depend on the number of trials, N. When N is too small, the value of the calculated information will be biased (Carlton, 1969; Houtsma, 1983). We shall use the term *calculated information* to refer to the value obtained without consideration of the bias. Transmitted information also depends on the stimulus range, $s_T - s_1$ (Braida & Durlach, 1972; Luce, Green, & Weber, 1976). For a given stimulus and response set, S and R, and a constant value of N, we shall be concerned with the changes in the

TABLE 1

A General $m \times n$ Matrix

	1	2	$\cdots$	k	$\cdots$	n
1	N_{11}	N_{22}		N_{1k}		N_{1n}
2	N_{21}	N_{22}		N_{2k}		N_{2n}
$\vdots$						
j	N_{j1}	N_{j2}		N_{jk}		N_{jn}
$\vdots$						
m	N_{m1}	N_{m2}		N_{mk}		N_{mn}

calculated value of I_t as adjacent rows or columns of the matrix are merged. For example, two adjacent columns may be merged if the two response sets $\{r_1, r_2\}$ and $\{r_3, r_4\}$ are joined into one response set $\{r_1, r_2, r_3, r_4\}$. We shall demonstrate below, following Shannon (1948), that whenever a stimulus–response matrix is contracted by the process of merging rows or columns, for example by reducing an $m \times m$ matrix to an $(m-1) \times (m-1)$ matrix, the calculated information can never increase. That is, for any contraction of the matrix by merging, the change in calculated information, ΔI_t, is negative or equal to zero. The converse must also be true; for any expansion of the matrix by subdivision of rows and columns, $\Delta I_t \geqslant 0$.

The process of merging adjacent rows and columns is illustrated in Fig. 1, where a $T \times T$ matrix (fine lines) is reduced to an $m \times n$ matrix (heavy lines).

The importance of this theorem on contractions lies in its implications for sensory channel capacities. In identification tasks, Miller (1956) and many others have shown for a fixed stimulus range that as the number of stimulus categories increases, transmitted information increases monotonically and tends to approach an asymptote not greater than 2.5 bits or 1.75 natural units (Figs. 2 and 3). We shall demonstrate here that the monotonic rise in transmitted information with increasing number of categories is due largely to the mathematical properties of the stimulus–response matrix and is not related strongly to the properties of the sensory system. However, the limiting value toward which the curve rises is not a purely mathematical constant and may take origin from the physical nature of the

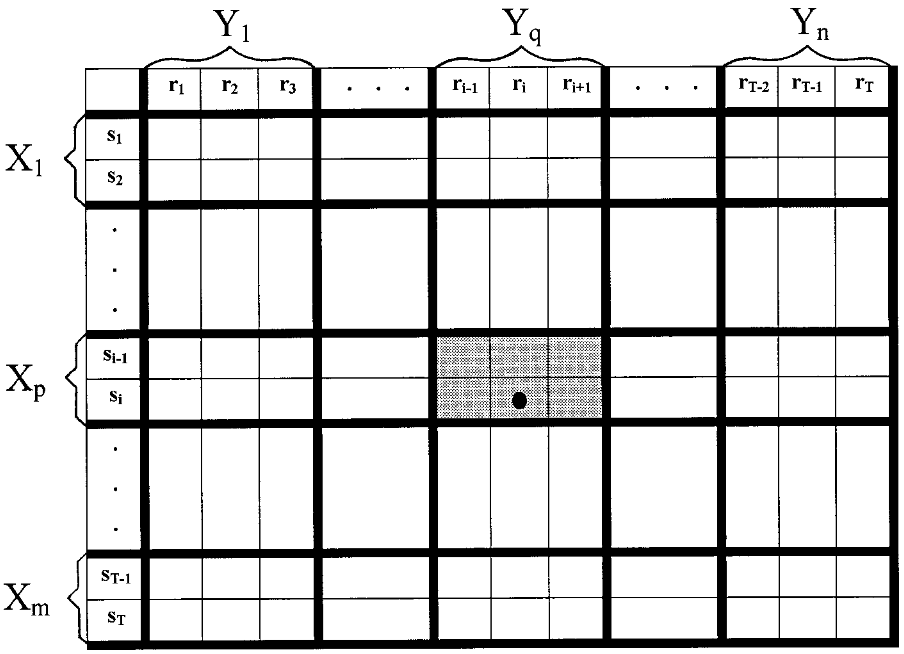

FIG. 1. In a process of categorization, the stimulus set $S = \{s_1, ..., s_T\}$ and response set $R = \{r_1, ..., r_T\}$ are partitioned into stimulus categories $X = \{X_1, ..., X_m\}$ and response categories $Y = \{Y_1, ..., Y_n\}$, respectively. If the correct response to s_i is r_i where $s_i \in X_{p_i}$ and $r_i \in Y_{q_i}$ then the correct response to X_{p_i} is Y_{q_i}.

sensory receptor (Norwich, 1993, pp. 261–264; Baird, 1970a, b). So the psychological issue we are addressing here is the relationship between the shape of the curve in identification tasks obtained by plotting transmitted information against number of categories and the limiting value toward which transmitted information tends.

NUMERICAL DEMONSTRATION

The reader may easily observe the effects of the merging of rows and columns by producing a confusion matrix using any technique whatsoever. The data may be obtained by experiment or by computer simulation; they may be realistic or purely random. The only requirement is that only one data set be used and the respective matrices be formed by merging adjacent rows and columns.

The merging of two adjacent columns into one may be performed quite arbitrarily as shown in Table 2. Table 3 shows the decrease in calculated information when an experimentally obtained 10×10 matrix is reduced to a 9×9 matrix by a merging of the last two rows and columns (Norwich, Wong, and Sagi, 1998). In Table 4, we give another example of a 10×10 matrix. This matrix has been produced by generating uniformly distributed pseudorandom integers representing stimulus–response pairs. That is, the two numbers comprising a stimulus–response pair were generated by the computer as pseudorandom, uniformly distributed integers that lay between 1 and 10. The pair of numbers (p, q) would increment by one the matrix entry N_{pq}. We have then selected two rows arbitrarily and merged them to produce a 9×10 matrix. The transmitted information in the original matrix is very small, since the stimulus and response are uncorrelated. Nonetheless, the merging of two rows decreases the transmitted information very slightly, which is the effect we shall study in the theorem below. It may be seen that associated with each merging, the transmitted information, I_t, decreases or remains constant. This effect holds for all matrices.

The relationship between transmitted information and the number of subdivisions is illustrated in Figs. 2 and 3 using data measured in our laboratory (Norwich *et al.*, 1998). The experimental details will be elaborated in the discussion.

TABLE 2
Merging Two Columns

$N_{1,\alpha}$	$N_{1,(\alpha+1)}$		$N_{1,\alpha} + N_{1,(\alpha+1)}$
.	.		.
.	.	$\rightarrow$	.
.	.		.
$N_{m,\alpha}$	$N_{m,(\alpha+1)}$		$N_{m,\alpha} + N_{m,(\alpha+1)}$

TABLE 3

The Effect on Calculated Information of Merging Two Rows of a 10 × 10 Confusion Matrix

Original matrix[a]

20	7	12	9	4	1	0	0	0	0
10	10	15	10	10	1	1	0	0	0
14	11	7	10	6	1	2	1	0	0
1	5	14	16	9	2	4	0	0	0
4	1	4	14	11	8	2	1	0	0
0	0	5	8	14	12	10	6	0	1
0	0	0	2	8	7	8	14	7	1
0	0	0	1	2	7	17	9	9	4
0	0	0	0	3	4	11	18	11	4
0	0	0	1	1	0	2	4	18	13

Reduced matrix[b]

20	7	12	9	4	1	0	0	0
10	10	15	10	10	1	1	0	0
14	11	7	10	6	1	2	1	0
1	5	14	16	9	2	4	0	0
4	1	4	14	11	8	2	1	0
0	0	5	8	14	12	10	6	1
0	0	0	2	8	7	8	14	8
0	0	0	1	2	7	17	9	13
0	0	0	1	4	4	13	22	46

Note. Data are obtained from experiment (Norwich *et al.*, 1998). Stimuli were distributed uniformly in the range of 1–10 dB. The subject (W) was required to identify the stimulus to the nearest decibel.

[a] Calculated information $I_t = 5.956 \times 10^{-1}$ natural units.

[b] Row 9 merged with row 10; column 9 merged with column 10. Calculated information, $I_t = 5.617 \times 10^{-1}$ natural units.

Let us now turn to the primary theorem itself.

Shannon (1948) outlined a proof of the following theorem using *a priori* probabilities for signal transmission and receipt. That is, he assumed that the probabilities of transmission for each signal were known beforehand, as properties, say, of the English alphabet. Here we offer a proof in the language of stimulus–response matrices, where the probabilities must be obtained *a posteriori*. The proof we offer differs from that put forward by Shannon in that it follows from the theory of continuous convex functions.

THEOREM 1. *If an $m \times n$ confusion matrix is reformatted by merging two columns (or rows) into one single column (or row), the calculated information obtained from the resulting $m \times (n-1)$ (or $(m-1) \times n$) matrix will be equal to or less than the calculated information obtained from the original matrix.*

TABLE 4

The Effect on Transmitted Information of Merging Two Rows of a 10×10 Confusion Matrix

Original matrix[a]

95	69	102	107	90	100	103	93	102	83
98	94	107	114	91	104	78	102	118	114
104	91	99	103	101	92	95	91	119	97
113	91	103	98	99	105	99	98	95	97
113	85	87	122	102	104	92	93	102	96
92	83	125	112	91	100	120	101	108	100
100	110	98	97	86	121	99	100	72	105
84	111	99	76	88	107	106	118	82	107
97	102	99	100	83	91	112	106	94	92
89	113	113	112	102	96	128	123	102	98

Reduced matrix[b]

95	69	102	107	90	100	103	93	102	83
98	94	107	114	91	104	78	102	118	114
104	91	99	103	101	92	95	91	119	97
113	91	103	98	99	105	99	98	95	97
113	85	87	122	102	104	92	93	102	96
92	83	125	112	91	100	120	101	108	100
184	221	197	173	174	228	205	218	154	212
97	102	99	100	83	91	112	106	94	92
89	113	113	112	102	96	128	123	102	98

Note. Both stimuli and responses have been generated randomly from the same set of uniformly distributed numbers.

[a] Transmitted information, $I_t = 5.239 \times 10^{-3}$ natural units.

[b] Row 7 merged with row 8. Transmitted information, $I_t = 4.880 \times 10^{-3}$ natural units.

Proof. Recall that a function φ is convex provided that

$$\varphi(q_1 x_1 + q_2 x_2) \leqslant q_1 \varphi(x_1) + q_2 \varphi(x_2),$$

where

$$q_1 + q_2 = 1, \qquad q_1 > 0, \qquad q_2 > 0.$$

It can be shown by induction that

$$\varphi(q_1 x_1 + q_2 x_2 + q_3 x_3) \leqslant q_1 \varphi(x_1) + q_2 \varphi(x_2) + q_3 \varphi(x_3), \tag{2}$$

where

$$\sum_{i=1}^{3} q_i = 1,$$

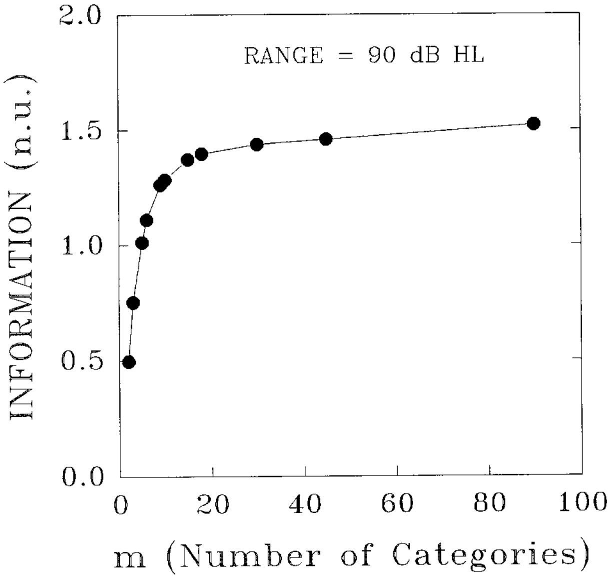

FIG. 2. Transmitted information (natural units) plotted against number of categories. Data were measured in our laboratories (Norwich *et al.*, 1998). An experiment was conducted on absolute identification of the loudness of pure tones at 1000 Hz. The range of stimuli was 1–90 dB. Transmitted information was obtained using a 90×90 matrix with 1-dB category widths and is plotted as the rightmost point on the graph. This matrix was compiled using $N = 500$ trials. The other points on the graph were obtained by merging the rows and columns of the matrix used to plot the rightmost point. The second point from the right was obtained from a 45×45 matrix with 2-dB category widths, the third point from a 30×30 matrix with 3-dB widths, etc. The maximum transmitted information was equal to 1.53 n.u.

etc. The geometric interpretation of this definition is that if every point on a chord lies above the surface which the chord spans, the mathematical function, ϕ, that defines the curve is *convex*.

Equation (2) can be generalized simply to the form

$$\varphi\left(\sum_{i=1}^{p} q_i x_i\right) \leqslant \sum_{i=1}^{p} q_i \varphi(x_i)$$

or to convex, continuous functions of two variables (Hardy, Littlewood, & Polya, 1964) in the form

$$\Phi\left(\sum_{i=1}^{p} q_i x_i, \sum_{i=1}^{p} q_i y_i\right) \leqslant \sum_{i=1}^{p} q_i \Phi(x_i, y_i). \tag{3}$$

We turn now to the calculation of transmitted information, I_t, from an $m \times n$ confusion matrix with the sum of all entries equal to N. Using the technique of

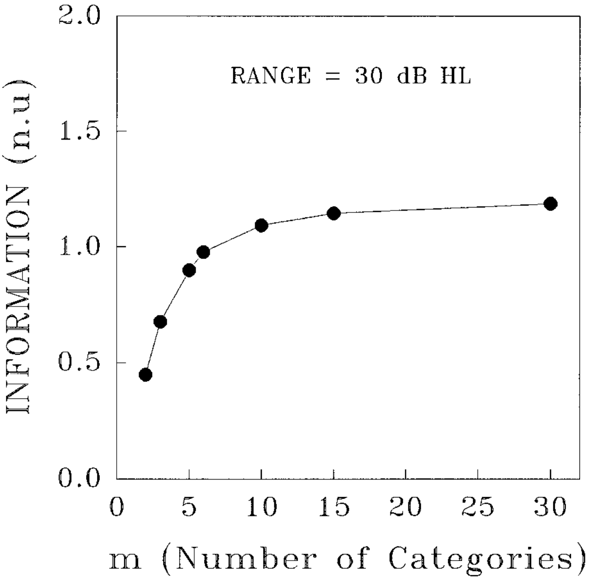

FIG. 3. Transmitted information (natural units) plotted against number of categories. The measurement of the rightmost point has been reported by Norwich *et al.* (1998). The graph has been drawn in exactly the same manner as the graph in Fig. 2. However, the range of stimuli is now 1–30 dB at 1000 Hz ($N = 500$). The rightmost point was calculated from a 30×30 matrix with 1-dB category widths, etc. We notice now that the maximum transmitted information is 1.20 n.u. We see that the shapes of the two curves in Figs. 1 and 2 are very similar, but the asymptotic information is greater with the greater stimulus range.

Garner and Hake (1951), we can express Shannon's Eq. (1) in a simple form consisting of only four terms,

$$N \cdot I_t = N \log N + \sum_{j=1}^{m} \sum_{k=1}^{n} N_{jk} \log N_{jk} - \sum_{j=1}^{m} \left(\sum_{k=1}^{n} N_{jk} \right) \log \left(\sum_{k=1}^{n} N_{jk} \right) - \sum_{k=1}^{n} \left(\sum_{j=1}^{m} N_{jk} \right) \log \left(\sum_{j=1}^{m} N_{jk} \right), \tag{4}$$

where N_{jk} is the value of the element at row j, column k. Expressed simply

$$N \cdot I_t = \text{trial entropy} + \text{element entropy} - \text{row entropy} - \text{column entropy}.$$

We consider now two calculated quantities of information: I_{t_1} is the information calculated from the matrix in which a single column has been formed by the merging of two adjacent columns, α and $\alpha + 1$, of an original matrix (see Table 2). I_{t_2} is the information calculated from the original matrix.

The quantity of relevance to our theorem is $I_{t_2} - I_{t_1}$. Note that combining two adjacent columns does not affect the row entropies (third term on the right-hand side of Eq. (4)), because these depend only on the sum over k of the N_{jk}, which does not change under this operation. Therefore,

$$N \cdot \delta I_t \equiv N \cdot (I_{t_2} - I_{t_1}) = [(\text{element entropy})_2 - (\text{element entropy})_1] - [(\text{column entropy})_2 - (\text{column entropy})_1]. \tag{5}$$

The difference between the element entropies can be written fully as

$$\sum_{j=1}^{m} \left[\sum_{k=1}^{\alpha-1} N_{jk} \log N_{jk} + N_{j\alpha} \log N_{j\alpha} + N_{j(\alpha+1)} \log N_{j(\alpha+1)} + \sum_{k=\alpha+2}^{n} N_{jk} \log N_{jk} \right]$$
$$- \sum_{j=1}^{m} \left[\sum_{k=1}^{\alpha-1} N_{jk} \log N_{jk} + (N_{j\alpha} + N_{j(\alpha+1)}) \log(N_{j\alpha} + N_{j(\alpha+1)}) + \sum_{k=\alpha+2}^{n} N_{jk} \log N_{jk} \right]$$
$$= \sum_{j=1}^{m} [N_{j\alpha} \log N_{j\alpha} + N_{j(\alpha+1)} \log N_{j(\alpha+1)} - (N_{j\alpha} + N_{j(\alpha+1)}) \log(N_{j\alpha} + N_{j(\alpha+1)})].$$

For ease of notation we let $a_j = N_{j\alpha}$ and $b_j = N_{j(\alpha+1)}$. Hence the difference between element entropies can be written succinctly in the form

$$\sum_{j=1}^{m} [a_j \log a_j + b_j \log b_j - (a_j + b_j) \log(a_j + b_j)]. \tag{6}$$

The difference between the column entropies can be written fully as

$$\sum_{k=1}^{\alpha-1} \left[\left(\sum_{j=1}^{m} N_{jk} \right) \log \left(\sum_{j=1}^{m} N_{jk} \right) \right] + \left(\sum_{j=1}^{m} N_{j\alpha} \right) \log \left(\sum_{j=1}^{m} N_{j\alpha} \right)$$
$$+ \left(\sum_{j=1}^{m} N_{j(\alpha+1)} \right) \log \left(\sum_{j=1}^{m} N_{j(\alpha+1)} \right) + \sum_{k=\alpha+2}^{n} \left[\left(\sum_{j=1}^{m} N_{jk} \right) \log \left(\sum_{j=1}^{m} N_{jk} \right) \right]$$
$$- \sum_{k=1}^{\alpha-1} \left[\left(\sum_{j=1}^{m} N_{jk} \right) \log \left(\sum_{j=1}^{m} N_{jk} \right) \right] - \left(\sum_{j=1}^{m} (N_{j\alpha} + N_{j(\alpha+1)}) \right)$$
$$\times \log \left(\sum_{j=1}^{m} (N_{j\alpha} + N_{j(\alpha+1)}) \right) - \sum_{k=\alpha+2}^{n} \left(\sum_{j=1}^{m} N_{jk} \right) \log \left(\sum_{j=1}^{m} N_{jk} \right)$$
$$= \left(\sum_{j=1}^{m} N_{j\alpha} \right) \log \left(\sum_{j=1}^{m} N_{j\alpha} \right) + \left(\sum_{j=1}^{m} N_{j(\alpha+1)} \right) \log \left(\sum_{j=1}^{m} N_{j(\alpha+1)} \right)$$
$$- \left(\sum_{j=1}^{m} (N_{j\alpha} + N_{j(\alpha+1)}) \right) \log \left(\sum_{j=1}^{m} (N_{j\alpha} + N_{j(\alpha+1)}) \right).$$

Making the same substitutions as before, the difference between the column entropies is equal to

$$\left(\sum_{j=1}^{m} a_j\right)\log\left(\sum_{j=1}^{m} a_j\right)+\left(\sum_{j=1}^{m} b_j\right)\log\left(\sum_{j=1}^{m} b_j\right)-\left(\sum_{j=1}^{m} (a_j+b_j)\right)\log\left(\sum_{j=1}^{m} (a_j+b_j)\right). \tag{7}$$

We now consider the function

$$\phi(a, b)=a\log a+b\log b-(a+b)\log(a+b). \tag{8}$$

We see from Eqs. (5)–(8) that

$$N\cdot\delta I_t=N\cdot(I_{t_2}-I_{t_1})=\sum_{j=1}^{m}\phi(a_j, b_j)-\phi\left(\sum_{j=1}^{m} a_j, \sum_{j=1}^{m} b_j\right). \tag{9}$$

By demonstrating that $N\cdot\delta I_t\geqslant 0$, we would prove our proposition. However, we must first prove two lemmas.

LEMMA 1. *If all* $q_i>0$ *are equal* ($q_i=q$), *and* $\sum_i q_i=1$, *then*

$$\phi\left(\sum_i qa_i, \sum_i qb_i\right)=q\phi\left(\sum_i a_i, \sum_i b_i\right).$$

Proof. This equation follows in a straightforward manner from the definition (8).

LEMMA 2. *If* $Q\equiv\phi_{aa}u^2+2\phi_{ab}uw+\phi_{bb}w^2$, *then* $Q\geqslant 0$, *where* $\phi_{ij}=\partial^2\phi/(\partial i\,\partial j)$.

Proof. From the defining function (8),

$$\phi_{aa}=b/[a(a+b)];\qquad \phi_{ab}=-1/(a+b);\qquad \phi_{bb}=a/[b(a+b)].$$

Therefore,

$$Q=\frac{1}{a+b}\left[\sqrt{\frac{b}{a}}\,u-\sqrt{\frac{a}{b}}\,w\right]^2\geqslant 0$$

since a and b are greater than zero by definition, which proves the lemma. Note that $Q=0$ when $u=(a/b)\,w$.

Now Hardy *et al.* (1964) have shown that a necessary and sufficient condition for the convexity of the function ϕ on an open domain is that $Q\geqslant 0$, which we have now proved. Therefore we know from Eq. (3) defining convex functions that

$$\phi\left(\sum_i qa_i, qb_i\right)\leqslant\sum_i q\phi(a_i, b_i). \tag{10}$$

But using Lemma 1 for equal q_i, we can state from Eq. (10)

$$q\phi\left(\sum_i a_i, \sum_i b_i\right) \leqslant q \sum_i \phi(a_i, b_i)$$

so that

$$\sum_{j=1}^{m} \phi(a_j, b_j) - \phi\left(\sum_{j=1}^{m} a_j, \sum_{j=1}^{m} b_j\right) \geqslant 0.$$

We know, however, from Eq. (9) that the latter inequality is the condition for which $\delta I_t \geqslant 0$. Thus the theorem is proved.

In evaluating the change in calculated information by merging two columns into one, row entropies cancel (Eq. (5)). Similarly, column entropies cancel in the merging of two rows into one. Hence, merging two rows will produce the same effect on the calculated information as merging two columns. It will also follow from the proof of this theorem that any *increase* in the number of rows or columns of the confusion matrix produced by a process of dividing the set of rows and columns of a previous matrix will increase (or leave unchanged) the calculated information.

DISCUSSION

We have now demonstrated that any reduction of the confusion matrix by a process of merging rows and/or columns will either decrease the calculated information or leave it unchanged. Conversely, subdivision of rows or columns will increase the calculated information or leave it unchanged. We note that this effect is independent of the physical interpretation of the matrix; it is purely a mathematical function of confusion matrices. In this paper, we make no statement about the relative contributions of stimulus or response categories. We state only that a reduction in either will never increase the calculated information. The empirical observation that, in absolute identification (and categorization) experiments, varying the number of response categories has more effect than varying the number of stimuli (in each category) is addressed in Lacouture *et al.* (1998). We must note also that it is *not* true in general that an increase in the number of rows or columns by a process other than subdivision will increase the calculated information.

The effect of the merging of rows and columns is illustrated in Figs. 2 and 3. These figures depict the change in transmitted information obtained from a single experiment on identification of loudness of pure tones at 1000 Hz (Norwich *et al.*, 1998) as categories are merged. In Fig. 2, the range of stimuli is equal to $(s_{90} - s_1) = 90$ dB. The rightmost stimulus set is $\{s_1, s_2, ..., s_{90}\} = \{X_1, X_2, ..., X_{90}\} = \{1\text{ dB}, 2\text{ dB}, ..., 90\text{ dB}\}$. The response set is similarly $\cdots \{1\text{ dB}, 2\text{ dB}, ..., 90\text{ dB}\}$. There are 90 categories ($N = 500$ trials). The stimulus set second from the right is $\{\{s_1, s_2\}, \{s_3, s_4\}, ..., \{s_{89}, s_{90}\}\} = \{X_1, X_2, ..., X_{45}\} = \{\{1\text{ dB}, 2\text{ dB}\}, \{3\text{ dB}, 4\text{ dB}\} \cdots\}, ..., \{89\text{ dB}, 90\text{ dB}\}\}$. The response set is similarly $\cdots \{\{1\text{ dB}, 2\text{ dB}\}, \{3\text{ dB}, 4\text{ dB}\} \cdots\}, ..., \{89\text{ dB}, 90\text{ dB}\}\}$. There are 45 categories. Each reduction gives rise to one confusion matrix, generating one value of transmitted information that is plotted against the number of categories. In Fig. 3, the range of stimuli is equal to $(s_{30} - s_1) = 30$ dB. The graph is plotted analogously to the graph in Fig. 2. In both cases, we see that the transmitted information decreases as we proceed to the left and is maximum for the stimulus set of greatest cardinality. As we shall discuss in more detail below, the graphs can be used to model Garner's (1953) data. We can understand from the above theorem why these graphs assume their characteristic shape. However, we cannot predict, from mathematical considerations alone, the maximum value which the transmitted information will assume for the set of maximum cardinality (the greatest number of categories).

We should also understand that in order to model the type of curve obtained by Garner (1953), or reported by Miller (1956), each condensation of a confusion matrix from the matrix of greatest stimulus category cardinality should attempt to preserve a constant category width; e.g., $\{\{1\text{ dB}, 2\text{ dB}\}, \{3\text{ dB}, 4\text{ dB}\} \cdots\}$, or $\{\{1\text{ dB}, 2\text{ dB}, 3\text{ dB}\}, \{4\text{ dB}, 5\text{ dB}, 6\text{ dB}\} \cdots$. In Garner's experiments, the stimuli, one in each "category," were equally spaced across the stimulus range. Condensing the original matrix into nonuniform categories, in principle, will model experiments whose stimuli were classified into similar nonuniform categories. The particular method of merging determines the rate of decrease of transmitted information. To take an extreme example, condensing an original 90×90 matrix into a 10×10 matrix with nine categories of 1-dB width and one category of 81-dB width will reduce the transmitted information immediately to less than one third of its original value.

It is of some importance to compare matrices produced by the process of merging or fusion, with similar matrices produced experimentally. For example, we can begin with a 90×90 matrix of measured trials (categories 1 dB in width) and merge rows and columns to produce a uniformly spaced 9×9 matrix. We can then generate a second set of measured trials over the same range where the participant produces a 9×9 matrix directly (stimulus categories 10 dB in width, with 10 stimuli in each category). For simplicity, we might call the first 9×9 matrix a merged or secondary matrix and the second 9×9 matrix an unmerged or primary matrix. In the first case, the task is one of identification where the results have been categorized by the experimenter. In the second case, the task is purely one of categorization. Transmitted information can be calculated in each case.

These two values are expected to be equal provided that stimuli are generated using the same statistical distribution in both cases. Suppose we consider the uniform distribution. That is, in the example cited above, a stimulus will be equally likely to assume any integral decibel value between 1 and 90 dB. Thus, when testing on a 90×90 matrix, the correct response to a stimulus of 62 dB would be "62 dB." However, when testing on a primary 9×9 matrix, the correct response to a stimulus of 62 dB would be "category 7," which means "between 61 and 70 dB." Under these conditions, the values of transmitted information calculated from both the primary and the secondary matrices are expected to be equal, because there should be little difference in the behavioral process by which the two matrices are produced. That is, the process by which a subject classifies the intensity of a stimulus in the first instance as belonging to a broad category should be very nearly equivalent to the processes wherein the subject estimates the intensity as precisely as possible, and the estimate is subsequently placed into the broad category within which this estimate fell.

By way of direct experimental verification, we merge a primary 90×90 matrix to produce a secondary 9×9 matrix and a secondary 2×2 matrix and then compute the transmitted information "secondary $I_t(9, 90)$" and "secondary $I_t(2, 90)$." We can then compare these computed values with the transmitted information obtained from a primary 9×9 matrix and a primary 2×2 matrix; i.e. "primary $I_t(9, 90)$" and "primary $I_t(2, 90)$," respectively. In performing these comparisons, we must be careful to compare matrices compiled from data gathered over the same range of stimuli and using the same number of trials. The experiments were performed using tones with intensities uniformly distributed over 1 to 90 dB, in the manner described by Norwich *et al.* (1998) for one subject. Five hundred trials were used for every 9×9 matrix and 100 trials for every 2×2 matrix. In the case of secondary matrices, our subject was required to identify 90 stimulus tones using 90 response categories over 1–90 dB. Subsequently, stimulus–response pairs were categorized, by us, into the appropriate 9×9 or 2×2 matrix while maintaining a constant category width, respectively. In the case of primary matrices, subjects were required to categorize 90 possible stimulus tones into nine or two response categories and the transmitted information obtained directly from the resulting 9×9 or 2×2 matrices, respectively.

The results are tabulated in Table 5. Notice that primary and secondary transmitted information values are nearly equivalent, which demonstrates that the decrease in transmitted information resulting from the use of a smaller number of categories over a fixed range is in agreement with our view of the underlying processes.

That is, we have adopted the following theoretical viewpoint regarding the equivalence of primary and secondary $m \times m$ matrices measured over the same stimulus

TABLE 5

Comparison of Transmitted Information from Primary and Secondary Matrices

$I_t(m, R_0)$ [n.u.]	$I_t(9, 90)$	$I_t(2, 90)$
Primary	1.205 ± 0.04	0.414 ± 0.05
Secondary	1.153 ± 0.04	0.391 ± 0.04

Note. Five hundred trials were used to compile each 9×9 matrix and 100 trials for each 2×2 matrix. Transmitted information values computed for primary and secondary matrices match closely, demonstrating that for uniformly distributed stimuli, a matrix obtained by merging rows and columns (secondary) will yield the same value of transmitted information as matrices obtained from experiments performed with a reduced number of rows and columns (primary). The error bars represent an estimate based on the chi-square distribution of the 90% confidence interval about the reported value.

range. Consider an absolute identification experiment with, say, 90 correct stimulus–response pairs, and (for simplicity) with the stimuli equally spaced. Also, consider an absolute identification experiment with 30 correct stimulus–response pairs where these 30 stimuli are also equally spaced and cover the same range as the earlier 90 stimuli. Then the information transmitted in the absolute identification experiment with the 30 correct stimulus–response pairs (primary 30×30 matrix) can be obtained from the information transmitted in the absolute identification experiment with the 90 correct stimulus–response pairs by collapsing the 90×90 matrix associated with the latter experiment to a 30×30 (secondary) matrix in the manner described in this paper. This is not a purely mathematical result and would benefit from further experimental studies.

As a further demonstration of our theoretical viewpoint regarding the equivalence of primary and secondary $m \times m$ matrices measured over the same stimulus range, and using the same number of trials, we analyze the date of Garner (1953). Absolute identification experiments were conducted over a fixed range of 15 to 110 dB, i.e., $R_0 = 95$ dB, using $m = 4$, 5, 6, 7, 10, 20 stimulus categories with one stimulus in each category. Casting Garner's data into our terminology, the transmitted information, $I_t(m, R_0)$, was measured from the resulting primary confusion matrices. To overcome a small sample bias, data were pooled over several subjects in such a way that the number of trials used corresponded to $N = 600m$. Garner's paradigm was recreated through Monte Carlo simulation (Wong and Norwich, 1997). Consider a constant parameter $\sigma(R_0)$ that corresponds to subject error over the fixed stimulus range, R_0. Using $\sigma(R_0)$, stimulus–response pairs can be generated to any desired number of trials, N, and categorized into any desired secondary $m \times m$ matrix. A best value was found for $\sigma(95)$ such that after $N = 12,000$ simulated trials, the resulting transmitted information corresponded to Garner's primary $I_t(20, 95)$. Using $\sigma(95) \simeq 4.8$ dB and the same seed for pseudorandom number generation, secondary $I_t(m, 95)$ were calculated with $N = 600m$ and $m = 4$, 5, 6, 7, 10. The comparison of Garner's primary $I_t(m, 95)$ values with our simulated secondary $I_t(m, 95)$ values is shown in Fig. 4. Again, primary and secondary transmitted information values are nearly equivalent.

We should note that for $I_t(20, 95)$, our simulation is a model of Garner's identification experiment using 20 stimuli, equally spaced over R_0. In particular, each stimulus category, $X_j = \{s_j\}$, and response category, $Y_k = \{r_k\}$, is represented by single numbers, s_j, $r_k \in \{15, 20, 25, ..., 110\}$. In a simulated trial, s_j was paired with a response α normally distributed about s_j such that $\alpha \in [15, 110]$. Subsequently, s_j was paired with the r_k closest to α and hence the stimulus–response pair (X_j, Y_k) was generated.

CONCLUSIONS

For identification tasks, Miller (1956) drew attention to the characteristic relationship between transmitted information and the number of stimulus categories, each containing only one stimulus, for which this information was calculated. Transmitted information rose rapidly, as the number of stimulus categories increased from two to three or four and then increased less rapidly as the

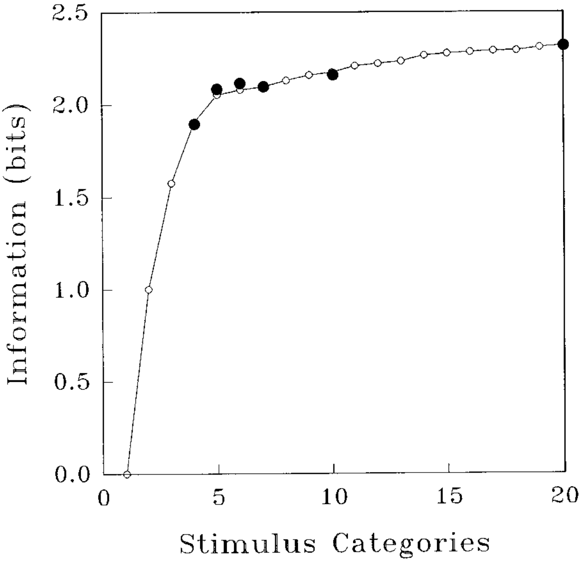

FIG. 4. Transmitted information (bits) plotted against number of categories. Filled circles: Data from Garner (1953). Open circles: Garner's data recreated through computer simulation. A constant value of 4.8 dB was used corresponding to subject error for the fixed range of 95 dB. Stimulus–response pairs were generated over the 95-dB range and progressively reorganized into decreasing numbers of categories of uniform width. Refer to main text for a more complete description.

number of categories continued to increase. The transmitted information soon approached its asymptotic value of about 1.7 natural units, or 2.5 bits. Results such as these were obtained for many one-dimensional sensory modalities.

We have demonstrated here that the characteristic shape of the transmitted information curves can be understood on a purely mathematical basis. If one begins with a stimulus set of maximum cardinality, corresponding physically to a set of measurements performed with a high degree of accuracy, one can then construct the general shape of the curve that would be obtained as the number of rows and columns is reduced progressively by a process of merging, while maintaining equal widths of all categories. That is, the entire shape of the curve is determined in principle by the matrix of maximum cardinality, or the rightmost point. The larger mystery underlying these graphs of transmitted information against category number is the value of the asymptotic information, which is not addressed here.

APPENDIX: NOMENCLATURE

S Stimulus set
s_i ith stimulus (dB HL)
X_j jth stimulus category
R Response set
R_o Range of stimulus values (dB)
r_k kth response (dB HL)
Y_k kth response category
m Number of rows

n	Number of columns
T	Maximum number of rows and/or columns
N	Total number of trials
N_{jk}	Matrix element representing number of times X_j was represented as Y_k.
I_t	Transmitted information
$H(Y)$	Response entropy
$H(Y\mid X)$	Response equivocation
$I_t(m, R_o)$	Transmitted information from a matrix with m rows, with stimuli selected over a range of R_0 dB.
δI_t or ΔI_t	Information difference between two stimulus-response matrices
φ	Convex function of one variable
ϕ, Φ	Convex function of two variables

REFERENCES

Baird, J. C. (1970a). A cognitive theory of psychophysics: I. Information transmission, partitioning, and Weber's law. *Scandinavian Journal of Psychology*, **11**, 35–46.

Baird, J. C. (1970b). A cognitive theory of psychophysics: II. Fechner's law and Stevens' law. *Scandinavian Journal of Psychology*, **11**, 89–102.

Carlton, A. G. (1969). On the bias of information estimates. *Psychological Bulletin*, **71**, 108–109.

Braida, L. D., & Durlach, N. I. (1972). Intensity Perception. II. Resolution in one-interval paradigms. *The Journal of the Acoustical Society of America*, **51**, 483–502.

Garner, W. R., & Hake, H. W. (1951). The amount of information in absolute judgments. *Psychological Review*, **58**, 446–459.

Garner, W. R. (1953). An information analysis of absolute judgments of loudness. *Journal of Experimental Psychology*, **46**, 373–380.

Hardy, G. H., Littlewood, J. E., & Polya, G. (1964). *Inequalities*. Cambridge, UK: Cambridge Univ. Press.

Houtsma, A. J. M. (1983). Estimation of mutual information from limited experimental data. *Journal of the Acoustical Society of America*, **74**, 1626–1629.

Lacouture, Y., Li, S.-C., & Marley, A. A. J. (1998). The roles of stimulus and response set size in the identification and categorisation of unidimensional stimuli. *Australian Journal of Psychology*, **50**, 1–10.

Luce, R. D., Green, D. M., & Weber, D. L. (1976). Attention bands in absolute identification. *Perception & Psychophysics*, **20**, 49–54.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, **63**, 81–97.

Norwich, K. H. (1993). *Information, sensation and perception*. San Diego: Academic Press.

Norwich, K. H., Wong, W., & Sagi, E. (1998). Range as a factor in determining the information of loudness judgments: Overcoming small sample bias. *Canadian Journal of Experimental Psychology*, **52**, 63–70.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, **27**, 623–656 (Appendix 7).

Wong, W., & Norwich, K. H. (1997). Simulation of human sensory performance. *BioSystems*, **43**, 189–197.

Received: June 23, 1998; revised: April 28, 1999