Biol. Cybern. 78, 377–387 (1998)

Biological Cybernetics

A conceptual approach to the magical number 7

Willy Wong, Shuji Mori

Faculty of Engineering, Toyama Prefectural University, 5180 Kurokawa, Kosugi-machi, Imizu-gun, P-939-03 Toyama, Japan

Received: 12 September 1997 / Accepted in revised form: 13 March 1998

Abstract The traditional paradigm for studying the magical number is questioned and a new approach is sought in order to obtain a better conceptual understanding of this phenomenon. Building on earlier work, a theory is proposed whereby the results of an absolute identification experiment can be characterized by a single parameter to a reasonable approximation. This parameter is the variance in the subject's response to a sensory input. By reducing the magical number to a single parameter, we see that the value of the upper limit in information transmission depends not so much on the absolute magnitude of the response error, but actually on how fast this error grows with range. The theory also predicts a little known characteristic of the magical number. If this prediction can be demonstrated experimentally, we shall need to reinterpret the magical number.

1 Introduction

While a modern measuring device can estimate physical dimensions with astounding accuracy, a human being without additional accessories cannot do nearly as well. It appears that human sensory performance is restricted by certain fundamental limitations. These limits have been the subject of a great deal of interest for sensory scientists, particularly during the 1950s and 1960s, and are still being studied to this day. Beginning with the work of Garner (1953) and Miller (1956), it was demonstrated that humans could, at best, classify single-dimensional stimuli into five or six non-overlapping categories spanning a fixed range. Hence the title of Miller's celebrated paper: "The magical number 7, plus or minus 2: Some limits on the capacity for processing information."

Perhaps the most impressive feature of the magical number is that it is obeyed universally by almost all sensory modalities. The same limit which governs the judgement of line length will, to a good approximation, also limit the judgement of loudness.

To analyse experiments determining the magical number, Garner and Hake (1951) borrowed information theory from communications engineering for application with absolute identification experiments. In such experiments, the subject's task is to correctly identify randomly presented stimuli using a pre-assigned classification scheme. This classification associates a single stimulus to each category: increasing category number corresponds to increasing stimulus magnitude. For example, in the absolute identification of loudness, we might conduct an experiment of four categories where category no.1 corresponds to a stimulus of intensity 10 dB, no.2 to 15 dB, no.3 to 20 dB and no. 4 to 25 dB. The subject, upon hearing a randomly selected tone from any of the four categories, is required to identify the stimulus correctly (1–4).

In the context of Shannon's information theory (Shannon and Weaver 1949), a source communicates information across a noisy channel. The source message is denoted as X and the received message as Y. If transmission were error-free, the source information $H(X)$ would equal the transmitted or received information I_t. However, the presence of noise obfuscates the transmitted message and there is invariably a loss in information. Thus, the transmitted information can be expressed as

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

where $H(X|Y)$ is the loss in information and is calculated from the number of times the input was incorrectly identified. If the source input is random and drawn from a uniform distribution, then $H(X)$ is just $\log m$, the Shannon information for m equally probable events.

Returning to the absolute identification experiments, when the input is correlated to the subject's response, the

Correspondence to: W. Wong, Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge, Madingley Road, Cambridge, CB3 OHE, UK
(e-mail: ww209@phy.cam.ac.uk, Tel.: +44-1223-337460, Fax: +44-1223-337356)

results can be summarized in a confusion matrix where correct responses define the main diagonal. From this matrix, the value of the information can be calculated using (1). This value corresponds to the information required by the subject to classify a single stimulus, and it rarely exceeds 1.6 natural units or 2.3 bits of information corresponding to the $2^{2.3} \simeq 5$ categories. For example, in Fig. 1 the results of Garner (1953) are plotted, showing how information transmitted to the subject varies with the stimulus or input information $H(X)$. We note that when the input information exceeds 2 bits or so, the transmitted information approaches an asymptote and cannot be further increased.

In addition to experimental work, there are several theoretical studies which attempt to explain the magical number phenomenon (Braida and Durlach 1972, 1988; Laming 1984; Luce et al. 1976; Marley and Cook 1984; Treisman 1985). They are all based on a Thurstonian model with signal detection theory (i.e. statistical decision theory) in which sensory effects of the stimulus input are represented by normal distributions on a decision axis. In their models, the limits to absolute identification performance (magical number phenomenon) result from cognitive factors such as imprecise memory of the context by which judgements are made (Braida and Durlach 1972, 1988; Laming 1984; Marley and Cook 1984), fluctuation of attention over the stimulus range (Luce et al. 1976) and trial-by-trial shifting of response criteria (Treisman, 1985). Mainly because they attempt to explain other phenomena in absolute identification as well (such as sequential dependencies and edge effects) their models are complex and require the assessment of several parameters. Recently, Lacouture and Marley (1995) proposed a connectionist neural network model which accounts for the magical number and edge effects in absolute identification.

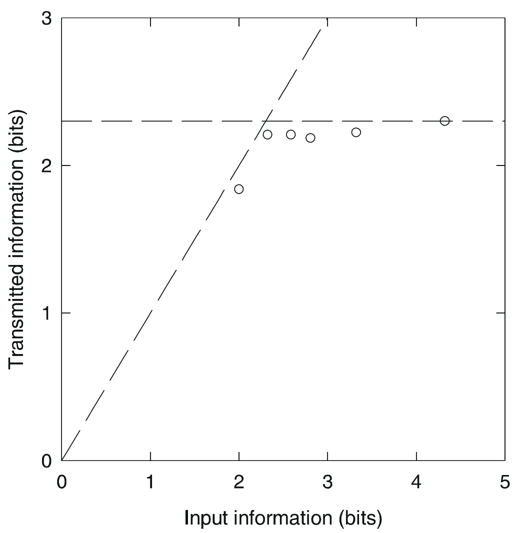

Fig. 1. The loudness data of Garner (1953) showing how transmitted information varies with input information. Input information is defined as the base 2 logarithm of the number of categories. The plateau in information as shown by the *horizontal line* represents an upper limit in transmitted information of approximately 2.3 bits. The *diagonal line* shows transmitted information in the case of error-free transmission. In other words, transmitted information equals input information

In this paper, we propose a new approach to understanding the magical number phenomenon in which its salient features can be explained with only a single parameter. Because this approach is largely conceptual in nature, it is simpler than any of the earlier approaches. Furthermore, the entire theory is based within an information framework, making it easy to compare with existing experimental studies. However, the theory is similar in several respects to many of the earlier studies (particularly signal detection theory), and a paper has been submitted elsewhere dealing with these issues (Wong and Mori, 1998, in press).

We confine the scope of the paper to a single issue: to understand better the principles behind the magical number phenomenon. Just as we can learn more about geometry by translating the parallel line axiom into other equivalent, but equally valid statements, we hope to learn more about the magical number by simplifying, where possible, the problem at hand, stripping it down to its core and examining it from another perspective. This would involve, among other things, asking questions such as: "In Fig. 1, what aspect of this graph is purely mathematical and what aspect is perceptual in origin?" "How many parameters are required to construct the information curve found in Fig. 1?" Along the way, we will also suggest ways in which future experiments can be conducted more effectively, as well as offer new predictions which can be verified through future experiments.

We make no attempt here to take into account sequential dependencies (tendency of the current response to be correlated to past responses and past stimuli, see Mori 1989) or drift in the subject's response during the course of an experiment (see Treisman 1985). Furthermore, we shall ignore the effects of feedback in absolute identification. These are all factors that can be considered in future studies.

2 Background

We begin with a close examination of how absolute identification experiments are used to study the magical number. We shall illustrate this experiment for the judgement of loudness, keeping in mind that similar results can be obtained with other sensory stimuli and modalities. In this experiment, the range of the stimulus input (i.e. sound intensity) is fixed at R and the continuum is divided into m equal categories in decibel space. For example, a 30–70 dB range can be divided into 5 equally spaced categories by assigning category 1 = 30 dB, 2 = 40 dB, 3 = 50 dB, 4 = 60 dB, 5 = 70 dB, etc. Let us denote the input variable as X . The possible values spanned by X are given by the set $\{x_j\}$, where $j = 1 \ldots m$. They are randomly presented to the subject, who must try to correctly identify the category to which the tone belongs.

The output variable Y spanning the set $\{y_k\}$ can be defined similarly. We shall only consider here the case where there are the same number of output categories as input categories (although they do not, in general, have to be equal; see Eriksen and Hake 1955). Thus, the set $\{y_k\}$ consists of m possible responses ($k = 1 \ldots m$). The input and response pairs obtained from an experiment can be summarized in a confusion matrix with elements n_{jk} being the frequencies of the responses y_k when stimulus x_j is presented. A trial is defined as a single presentation and classification of a tone. Of course, the sum of all the elements in the matrix must equal the total number of trials N carried out in the experiment. Thus,

$$\sum_{j=1}^{m} \sum_{k=1}^{m} n_{jk} = N \tag{2}$$

When a sufficient number of trials have been collected, various probabilities can be estimated from the matrix. Let $p(x_j)$ denote the probability of presenting stimulus x_j and $p(y_k)$ the probability of responding y_k. $p(y_k)$ is evaluated from the expression for total probability

$$p(y_k) = \sum_{j=1}^{m} p(x_j) p(y_k|x_j) \tag{3}$$

where $p(y_k|x_j)$ is the conditional probability of responding y_k when x_j is presented. For later use, we also define $s^2_{\mathrm{eff},j}$ as the sample variance along the jth row of the confusion matrix, namely the variance of the responses y_k given stimulus x_j.

Returning to the equation of transmitted information (1), one can easily demonstrate that

$$\begin{aligned} I_t &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned} \tag{4}$$

where I_t is now calculated from the response information

$$H(Y) = -\sum_k p(y_k) \log\ p(y_k) \tag{5}$$

and equivocation

$$H(Y|X) = -\sum_j p(x_j) \sum_k p(y_k|x_j) \log p(y_k|x_j) \tag{6}$$

We now turn our attention to deriving a simple expression for the transmitted information from (4) in the case where the experiment is carried out over a large number of categories, $m \gg 1$. We present a formal list of assumptions to be used in the derivation. They are introduced in turn followed by certain remarks and comments.

2.1 Assumptions of the theory

(1) *Edge or anchor effects can be neglected.* A typical confusion matrix obtained experimentally is shown in Table 1. The data were collected for the absolute identification of pitch and will be discussed in Sect. 3.1. Edge effects are due to the boundaries of the matrix and can be most prominently observed in the first and last rows. The subject has a greater chance of giving the correct response when the stimulus is located at the edges of the stimulus range.

(2) *$p(y_k|x_j)$ can be approximated by a normal distribution for all ranges R.* With reference to Table 1, we see that the distribution of response along each row can be approximated by a normal distribution except for the very first and last rows. To approximate the discrete response distribution $p(y_k|x_j)$ by a normal distribution, we must first partition the normal distribution into the m categories of response and then integrate the distribution between endpoints calculated from the location of the categories. Mathematically speaking, if y denotes the random variable representing the response, then

$$p(y_k|x_j) = \int_{\alpha_{k-1}}^{\alpha_k} N(y; \mu_j, \sigma_j)\, dy \tag{7}$$

where $N(y; \mu_j, \sigma_j)$ is a normal distribution with mean μ_j and variance σ_j^2. μ_j is the mean response when x_j is presented. When the subject does not drift, the average response equals the correct response, $\mu_j = x_j$. The endpoints of the integration are defined by the midway points between adjacent categories as mapped onto the stimulus range $[0, R]$,

Table 1. A typical confusion matrix summarized from the results of a pitch experiment. The input categories are specified by the variable x_j and the output categories by y_k. Notice that the distribution of x_j is uniform. y_k follows a similar pattern and can be approximated by a uniform distribution. Furthermore, the response along each row is observed to be approximately normally distributed and centred along the main diagonal

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	x_j^{total}
x_1	60	0	0	0	0	0	0	0	0	0	60
x_2	2	53	4	1	0	0	0	0	0	0	60
x_3	0	3	42	11	4	0	0	0	0	0	60
x_4	0	3	7	36	10	4	0	0	0	0	60
x_5	0	0	1	8	33	17	1	0	0	0	60
x_6	0	0	0	0	4	35	18	3	0	0	60
x_7	0	0	0	0	1	4	43	11	1	0	60
x_8	0	0	0	0	0	2	14	38	6	0	60
x_9	0	0	0	0	0	0	1	3	55	1	60
x_{10}	0	0	0	0	0	0	0	0	5	55	60
y_k^{total}	62	59	54	56	52	62	77	55	67	56	

$$\alpha_0 = 0$$
$$\alpha_k = (2k-1)R/2(m-1), \quad k = 1\ldots(m-1) \tag{8}$$
$$\alpha_m = R$$

In general, we note that σ_j may be different for different x_j; however,

(3) *$\sigma_j = \sigma$ for all j.* Both the assumption of the fixed variance and the assumption of the normal distribution of response $N(y;\mu_j,\sigma_j)$ have been used by other investigators in other studies as well (i.e. Durlach and Braida 1969). Since σ_j is the standard deviation of $N(y;\mu_j,\sigma_j)$, we can similarly define $\sigma_{\mathrm{eff},j}$ to be the standard deviation of $p(y_k|x_j)$. Furthermore, we would expect that

(4) *The response variance σ^2_{eff} is the same for all x_j (that is, $\sigma_{\mathrm{eff},j} = \sigma_{\mathrm{eff}}$ for all j).* So $s^2_{\mathrm{eff},j}$, defined earlier as the sample variance along the jth row of the confusion matrix, are estimates of σ^2_{eff}. While the last two assumptions are motivated in part by experimental data (i.e. Table 1 and others), they are expected to hold in general and not just for a particular data set.

(5) *$p(x_j) = 1/m$, where m is the number of stimulus/response categories.* $p(x_j)$ is a uniform distribution controlled by the experimenter.

2.2 Derivation of an analytical expression for I_t

With these assumptions we can now derive two lemmas.

Lemma 1 *To a good approximation, $p(y_k)$ is given by a uniform distribution*

From the second assumption, if the subject's response does not drift, $p(y_k|x_j)$ can be approximated by a normal distribution with mean response equal to x_j. Recall that a normal distribution is symmetrical about the independent variable (in this case, y_k) and the mean value (in this case,x_j). These two variables can be interchanged without affecting the value of the function. Thus,

$$p(y_k|x_j) = p(x_j|y_k) \tag{9}$$

However, by Bayes theorem we know that

$$p(x_j)p(y_k|x_j) = p(y_k)p(x_j|y_k) \tag{10}$$

and as a consequence of (9) and the last assumption, we have

$$p(y_k) = 1/m \tag{11}$$

The response is uniformly distributed. For example, in Table 1 we observe that the total probability of response is uniform to a good approximation in agreement with (11).

Lemma 2 *$H(Y|X) \simeq -\sum_k p(y_k|\bar{x}) \log\ p(y_k|\bar{x})$ where $\bar{x} = (m+1)/2$ is the middle category of the matrix.*

Returning to (6), we see that the nested equation

$$-\sum_k p(y_k|x_j) \log\ p(y_k|x_j) \tag{12}$$

represents the entropy or information along the jth row of the matrix, and to calculate $H(Y|X)$ we need only to average (12) across all rows. Since $p(y_k|x_j)$ can be approximated by the normal distribution, (12) is, roughly speaking, the entropy of a normal distribution.

When the width of the category $\Delta y = R/m$ is sufficiently small, the sum in (12) can be approximated by an integral. Let us also replace $p(y_k|x_j)$ with the probability density function $p(y|x_j)\Delta y$. Equation (12) would then take the form of

$$-\sum_k p(y_k|x_j) \log\ p(y_k|x_j) \simeq -\int_0^R p(y|x_j) \log p(y|x_j)\, dy - \log(\Delta y) \tag{13}$$

This equation is derived in greater detail in Wong and Norwich (1997), Appendix A. At this point, we refer to the results of Shannon and Weaver 1949 (also Norwich 1993) where it was demonstrated that the entropy of a normal distribution depends only on the variance of the distribution and not on the mean. That is,

$$-\int_{-\infty}^{\infty} p(y|x_j) \log\ p(y|x_j)\, dy = \frac{1}{2}\log(2\pi e \sigma_j^2) \tag{14}$$

where σ_j^2 is the variance of $p(y|x_j)$. We have defined $p(y|x_j) = 0$ outside the range $[0, R]$. Thus, (13) becomes

$$-\sum_k p(y_k|x_j) \log\ p(y_k|x_j) \simeq \log\ (\sqrt{2\pi e}\sigma_j/\Delta y) \tag{15}$$

By an earlier assumption $\sigma_j = \sigma$ for all values of j; equation (15) would imply that the value of (12) is equal for all j and can therefore be calculated using only a single row of the matrix. For simplicity, we shall use the middle row defined by $\bar{x} = (m+1)/2$. Returning to (6) and using $p(x_j) = 1/m$, we have

$$\begin{aligned} H(Y|X) &= -\sum_j p(x_j) \sum_k p(y_k|x_j) \log\ p(y_k|x_j) \\ &= -\frac{1}{m}\sum_j \sum_k p(y_k|x_j) \log\ p(y_k|x_j) \\ &\simeq -\sum_k p(y_k|\bar{x}) \log\ p(y_k|\bar{x}) \end{aligned} \tag{16}$$

That is, the calculation of the equivocation from the confusion matrix, under these approximations, is reduced to the calculation of the entropy of a single, representative row of the matrix (i.e. the middle row).

We conclude this lemma with three additional remarks. The first refers to a limitation in the use of (16). Since $\bar{x} = (m+1)/2$ can take on only integer values, m must be an odd number. Consequently, the use of (16) is restricted to matrices with only an odd number of categories. This is one limitation of the proposed theory. The second remark refers to (13). Since $p(y \mid x_j)$ is a distribution with variance σ^2, a relationship between σ and R has been implicitly assumed in this equation.

While R is a parameter controlled by the experimenter, σ is a parameter characterizing the human subject. We leave the discussion of $\sigma = \sigma(R)$ to Sect. 3.2. Finally, it is important to remember that, in writing (14), we have fixed the base of the logarithm to be $e = 2.718...$ We shall continue to use natural logarithms for the remainder of the paper.

We now proceed to the seminal result of this section: the derivation of an analytical equation governing the transmitted information as defined by (4) in the limit of large m.

Theorem *In the asymptotic limit of large m, the equation for transmitted information takes the form of* $I_t = \log(m/\sigma_{\mathrm{eff}}) - \frac{1}{2}\log(2\pi e)$.

From (4), we evaluate I_t in two parts. First, using $p(y_k) = 1/m$ from Lemma 1 we have

$$H(Y) = -\sum_k p(y_k)\log p(y_k) = \log(m). \tag{17}$$

Next, we substitute (15) into (16) and use $\sigma_j = \sigma$ to obtain

$$H(Y \mid X) = \log\left(\sqrt{2\pi e}\sigma/\Delta y\right) = \log\left(\sqrt{2\pi e}\sigma_{\mathrm{eff}}\right) \tag{18}$$

We have used the result $\sigma_{\mathrm{eff}} = \sigma/\Delta y = \sigma m/R$ to convert between discrete and continuous probabilities. This equation will be derived in the following section (3.1).

Finally, substituting (17) and (18) into (4), we have

$$I_t \simeq \log(m) - \log\left(\sqrt{2\pi e}\sigma_{\mathrm{eff}}\right) = \log(m/\sigma_{\mathrm{eff}}) - \tfrac{1}{2}\log(2\pi e) \tag{19}$$

This equation was first derived in Wong and Norwich (1997). It predicts the unbiased information for large values of m (~ 30) and has been demonstrated to work quite well (see Wong and Norwich 1997; Norwich et al. 1998 (in press)).

Empirically speaking, many investigators have found that I_t typically saturates for $m > 5$ (for example, see Fig. 1). Thus, (19) must also follow this trend and, for large values of m, we would expect σ_{eff} to be a linear function of m. Furthermore, since σ_{eff} is also a function of the full range R, it might appear at first glance that σ_{eff} is a rather complicated function defying simple theoretical description. However, this is not the case. As we shall demonstrate in the following section, σ_{eff}, for the most part, follows a rather simple and elegant equation.

3 A conceptual approach to the magical number 7

3.1. σ_{eff} as a function of m

At this point, we make one further refinement to the approach we have developed thus far: Corresponding to any stimulus input, we assume that there is an underlying continuous distribution to the subject's response. Furthermore, this response is normally distributed with variance σ^2. The normal distribution $N(y; x_j, \sigma)$ approximating the discrete probability of response $p(y_k \mid x_j)$ in the second assumption would therefore correspond to the newly introduced underlying response distribution.

This distribution is illustrated schematically in Fig. 2. We have defined a new variable a which is one-half the range ($R = 2a$). Since it was established earlier that the entropy of a normal distribution is not dependent on the mean of the distribution [cf. (14)], we have arbitrarily set the mean equal to zero. Thus, the distribution is centred at the origin with variance σ^2. The abscissa is labelled y to represent the continuous response axis. Once again, it is important to bear in mind the difference between σ and σ_{eff}. σ is the standard deviation of a continuous distribution and has the units of the stimulus magnitude (i.e. dB in this case). σ_{eff}, on the other hand, is calculated from the discrete categorical responses from the matrix and is, by definition, unitless.

To explore how σ_{eff} changes with m, we first look at an illustrative example where an experiment is conducted over 5 categories. These 5 categories are placed at equal distances within the range $[-a, a]$. Thus, categories 1 through 5 are located at $y = -a$, $-a/2$, 0, $a/2$ and a (Fig. 3). Next we must establish a set of criteria by which the subject chooses a category for response. Assuming that no systematic response bias is introduced, it would seem plausible to set the criteria at the midway point between adjacent categories (see Treisman 1985). Thus, in the case with 5 categories, the criteria would be placed at $y = -a$, $-3a/4$, $-a/4$, $a/4$, $3a/4$ and a. These are shown by the dotted lines in Fig. 3. To evaluate the probability that the subject would verbally classify the stimulus as category 1, we integrate the normal distribution between $[-a, -3a/4]$. Probability for responding category 2 is defined by the area under $[-3a/4, -a/4]$, and so forth.

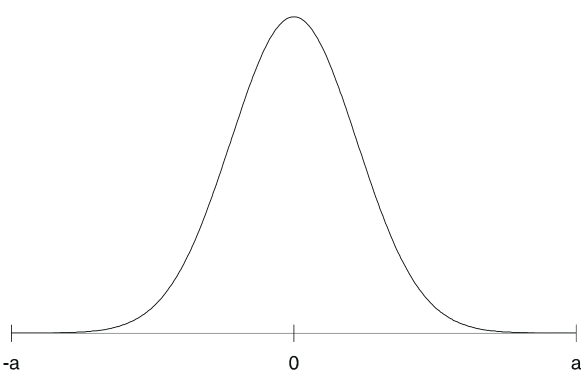

Fig. 2. The underlying continuous response distribution shown to be normally distributed: y is the response variable and has units of the stimulus magnitude (dB in the case of loudness judgements), a is defined as one-half of the total range of the experiment. The response along each row of the matrix is obtained by partitioning the continuous response distribution. See also Fig. 3

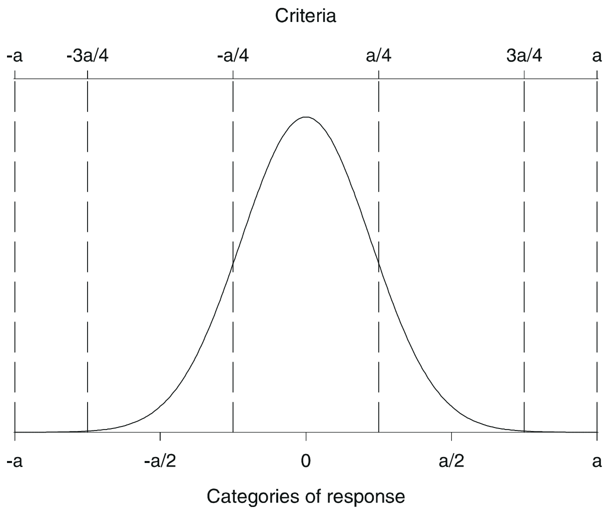

Fig. 3. The same response distribution as in Fig. 2, illustrating the case of 5 categories. We introduce a set of criteria that the subject uses to determine his or her response category. If there is no systematic bias in response, it is reasonable to assume that the criteria are placed at the midway point between adjacent categories. See text for more detail

Generally speaking, we can define a function $g(k)$ $(k = 1 \ldots m)$ which quantifies the probability of response with the kth category (Appendix A). Returning to our previous example with $m = 5$, we see that $g(1)$ equals the area under the normal distribution for the range $[-a, -3a/4]$, $g(2)$ is the area under $[-3a/4, -a/4]$, etc. In fact, $g(k)$ is just $p(y_k\,|\,\bar{x})$, the conditional probability of verbal response. We have identified k with y_k, with both functions centred at $\bar{x} = (m + 1)/2$.

For later use, we rewrite the equivocation, (16), in terms of $g(k)$ to obtain

$$H(Y \mid X) = -\sum_k g(k) \log g(k) \tag{20}$$

and substituting into (4), we find

$$I_t = \log(m) + \sum_k g(k) \log g(k) \tag{21}$$

where $H(Y) = \log(m)$ [cf. (17)]. Next, we calculate σ_{eff} with the function $g(k)$. Recall that k has a mean value equal to $(m + 1)/2$. Using the standard equation for variance, we have

$$\sigma_{\mathrm{eff}}^2 = \sum_{k=1}^{m} \left[k - \frac{1}{2}(m + 1) \right]^2 g(k) \tag{22}$$

This is the form of the expression for σ_{eff}^2 if $g(k) = p(y_k\,|\,\bar{x})$ is assumed to be normal (second assumption). Note that (22) essentially relates σ_{eff} to σ, where σ is embedded within $g(k)$ [see (33) in Appendix A].

Equation (22) is, by itself, a powerful statement. It implies that there is only one "true" value of variance σ^2 associated with the response distribution. For a single subject, σ is the only parameter that requires experimental determination. In principle, all associated values of σ_{eff} for $m = 1, 2, 3 \ldots$ can then be calculated from σ using (22) without conducting further experiments.

Experiments have been conducted by one of us which can be used to demonstrate the compatibility of (22) with experimental data. As described elsewhere (Mori 1991), these experiments were conducted on the absolute identification of pitch. The frequency range spanned 100–8000 Hz and was fixed throughout all experiments. The range was divided in equal logarithmic divisions into 4, 6, 10 and 16 categories with 240, 360, 600 and 960 trials collected for each experiment, respectively. All five subjects participated in the four experiments. σ_{eff} was extracted for each subject at $m = 4$, 6, 10 and 16 (see Appendix B for details). Since the individual data showed quite a bit of scatter, the values were averaged between subjects and are plotted in Fig. 4. Matters of intersubject variability are discussed in Sect. 4.3. The solid-line fit corresponds to the prediction of (22) using a single parameter $\sigma/a = 0.17$. The compatibility of the theory with the experimental data in Fig. 4 confirms our hypothesis that essentially all values of σ_{eff} can be predicted from a single value of σ.

When m is large (corresponding to σ_{eff}/m being small, cf. Fig. 4), there is a very simple and intuitive equation relating σ_{eff} to m and σ and R. As outlined in Appendix A, we can expand σ_{eff} for very large values of m ($m \sim 10$ categories) to obtain

$$\sigma_{\mathrm{eff}} \simeq m\sigma/R. \tag{23}$$

Recall that R/m is the width between each category. Thus, (23) says very simply that σ_{eff} multiplied by the width of each category will give you the true value of the standard deviation as we know from basic

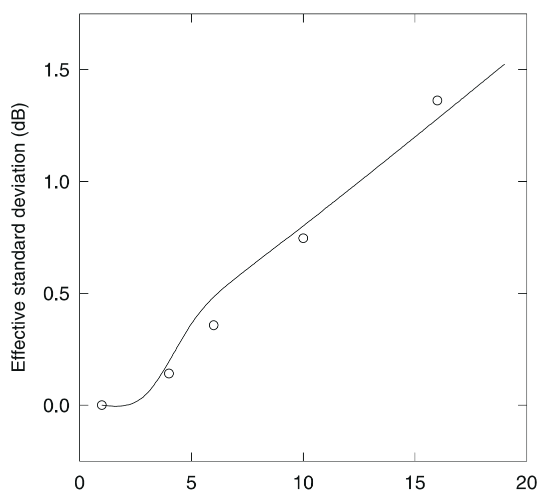

Fig. 4. Graph demonstrating the critical feature of the theory. The effective standard deviation (σ_{eff}) in response was measured at several categories for the absolute identification of pitch and are shown by the points. The theoretical curve was obtained from (22) using only a single parameter (the variance of the continuous response). This figure demonstrates that the standard deviation of response for any number of categories can, in principle, be calculated with only a single parameter σ

statistics. This is the result we used in the proof of Lemma 2. In the case where m is numerically equal to R, R/m is the unit dimension – in our case 1 dB (for loudness) – and σ is numerically equal to σ_{eff}.

Substituting (23) into (19), we find, for large m,

$$I_t = \log(R/\sigma) - \tfrac{1}{2}\log(2\pi e) \tag{24}$$

This is just the information of a continuous system with a uniformly distributed input of range R and normally distributed error of variance σ^2. We could have, essentially, begun the paper with the statement of (24). However, the ideas and tools developed in this section will not be wasted. We shall make full use of them at a later point.

For a fixed range R, I_t defines the unbiased information transmitted from an absolute identification experiment with a large number of categories. However, as it stands, it is not clear how I_t will change as a function of the range. This would depend on how σ changes as a function of R. We shall explore this avenue in the following section.

3.2. *σ as a function of R*

In the previous section we saw how a normal distribution (as illustrated in Fig. 2) can be used to approximate a subject's underlying response distribution. This distribution is characterized by a single parameter σ^2, the variance in response. The mean of the distribution plays no part in the calculations and has been set arbitrarily equal to zero. That is, $N(y; x_j, \sigma) = N(y; \sigma)$, a function with only one free parameter.

We are now interested in determining how the response variance changes with the range of the experiment $R = 2a$. First, we observe that the subject's total response is always normalized regardless of the range of the experiment. That is,

$$\int_{-a}^{a} N(y;\sigma)dy \simeq 1 \tag{25}$$

The equation is written with an approximate equality because the real response distribution is defined over a finite range while the true normal distribution $N(y;\sigma)$ extends over $[-\infty, \infty]$. Next, we recall that from the second assumption the response distribution is always normal-like regardless of the range of the experiment. Since we have assumed that the distribution itself does not vary with range, the standard deviation of response σ must change in order to preserve the relationship in (25) (σ is the only free parameter). From intuition, we would expect the standard deviation to change linearly with the range of the experiment. We now look at a precise mathematical statement of this idea.

Mathematically speaking, we introduce a linear transformation $y = \beta u$ or $u = y/\beta$, where $\beta > 0$ is a scaling constant. Next we choose $\beta = a/\gamma$, where $\gamma > 0$ is another constant and $a = R/2$ is half the range. Thus,

$$\begin{aligned}\text{Var}(y) &= \sigma^2 \\ &= \beta^2 \text{Var}(u) \\ &= \text{Var}(u)a^2/\gamma^2\end{aligned} \tag{26}$$

Since the constants are arbitrary, we can fix the values of both $\text{Var}(u)$ and γ to use to scale σ and a. This would imply that

$$\sigma = AR \tag{27}$$

where $R = 2a$ and $A = \sqrt{\text{Var}(u)}/2\gamma$. The standard deviation is linearly proportional to the range to which the distribution is scaled.

Consequently, if the subject always utilizes the full stimulus range to make his or her response, as a consequence of (25) and the second assumption, the subject's standard deviation must increase linearly with range. Equation (27) has in fact been observed experimentally. However, to examine fully how the standard deviation in response varies with range would require elaboration beyond the scope of this article. As it turns out, (27) is correct only for large values of R. When the range is small, the variance is affected by other factors including edge effects which are not considered in this paper. The resulting response distribution would no longer remain normal-like, violating the assumption of normality. When we include these effects in our calculations, we find that the standard deviation is then given by $\sigma = AR + B$, where B is an intercept. This equation is in good agreement with experimental results. In the case where R is large, σ can be approximated by (27). A publication addressing these issues is currently in preparation.

We can now use (27) to predict the absolute upper bound for the magical number in the limit of large m and large R. Substituting (27) into (24), we find

$$I_t = -\log A - \tfrac{1}{2}\log(2\pi e). \tag{28}$$

This equation predicts the unbiased, asymptotic value of information. Notice that in (28), apart from A, there are no other unevaluated parameters. Since A measures the rate of change of σ with respect to R, we see that the magical number is dependent only on how fast the standard deviation of response grows with stimulus range.

3.3. *Information transmission as a function of the input information*

We are now in a better position to understand what is perhaps one of the most celebrated results of psychophysics: the graph showing the limit of transmitted information with increasing number of stimulus categories. To derive this curve from the theory developed thus far, all we require is to choose a single value for the underlying variance σ^2 . As demonstrated in Sect. 3.1, we can then, in principle, predict the probability of verbal response, $g(k)$, for any number of categories m.

While it is tempting to use (19), this will not in fact work. Recall that (19) was derived in the limit of large m. In the case where $m \lesssim 10$, this equation is a poor approximation. We can, however, use (21) since no assumptions were made in the derivation of this equation regarding the size of m.

One can, of course, fit this equation to the data to obtain a "best-fit" value of σ. However, it would be much more interesting to go about it by another approach. Let us return to Fig. 1 and the data of Garner (1953). Input information on the abscissa is defined as $\log_2(m)$. If we assume that the transmitted information has reached the asymptotic value by $m = 20$ (the very last point), we can use (24) to solve for the value of σ with the range used by Garner ($R = 95$ dB). Solving backwards, we find $\sigma = 4.66$ dB. Substituting this value into $g(k)$ (see Appendix A) and then calculating I_t for different values of m using (21), we obtain the result shown in Fig. 5. Once again we emphasize that the theoretical curve was obtained without curve-fit, using only a single parameter calculated from the asymptote. Surprisingly, the theory even predicts a nadir and apex in the information curve in agreement with Garner's data.

4 Discussion

4.1. Is information transmission the primary variable?

Having involved ourselves with so many different equations and calculations, it is easy to lose sight of the original purpose of the paper. We sought primarily to ask the question: What is the nature behind the magical number phenomenon? Is it the changing value of the variance in the verbal response corresponding to experiments conducted at $m = 1, 2, 3, \ldots$ (i.e. σ_{eff} as a function of m)? Does the magical number phenomenon arise simply because there is error in the subject response (i.e. $p(y_k|x_j)$ quantifies the non-zero probability of incorrect responses)? Should information transmission I_t be the primary variable of investigation?

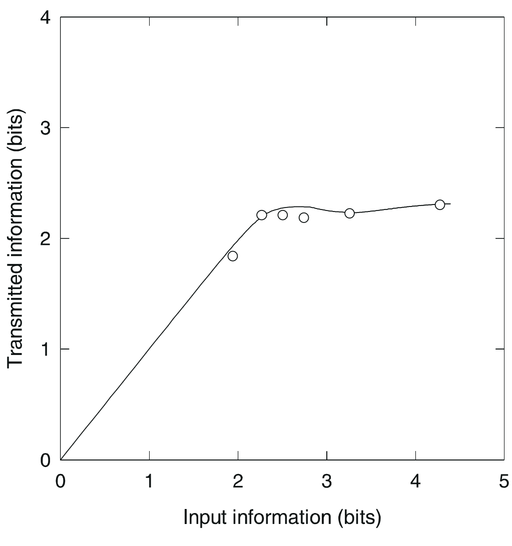

Fig. 5. The data of Garner (1953) along with the prediction by the theory. The theoretical curve was obtained without curve-fit. Assuming that the data have reached the asymptotic value by the very last point, we used (24) to solve for the value of σ. This value was then used to calculate the theoretical curve from (21). Note that the entire information curve was calculated from a single parameter alone. The theory predicts the existence of critical points in the information curve in agreement with Garner's data

The results of this study indicate that the answer to all of these questions is "No".

As we have seen from Fig. 5, I_t is actually a function only of a single parameter σ. From σ, all values of σ_{eff} can in principle be calculated. Moreover, it is not the absolute magnitude of σ which determines the magical number but actually how fast it grows with range [cf. (24)]. Thus, the single most important variable determining the magical number is the rate of change of the error. Quite simply, all we really need to obtain is

$$\frac{d\sigma}{dR} = A \tag{29}$$

which follows from the linear relationship between σ and R [cf. (27)]. It is then a simple matter to calculate the upper bound for I_t using (28). This would imply that σ is the most fundamental variable for study and not information transmission as previously assumed.

One can also derive this conclusion from quite a different perspective. Throughout the entire paper we have been dealing with unbiased estimates of information transmission. To obtain these values experimentally one would require many thousands of trials – the reason being that, for small sample sizes (e.g. $\lesssim 250$ points per row), the calculated information is biased and will overestimate the "true" upper limit of I_t, defeating the purpose of the experiment. While several attempts have been made to address the statistical bias (Miller and Madow 1954; Rogers and Green 1955; Carlton 1969), a viable method for obtaining unbiased estimates of I_t is still lacking to this day.

More recently, with the widespread availability of computers, various investigators have proposed that such "missing" data can in fact be simulated by computer (i.e. Houtsma 1983; Mori 1991; Wong and Norwich 1997). Wong and Norwich proposed a normal distribution for the conditional probability $p(y_k|x_j)$, demonstrating that real data can be simulated quite accurately using only a single value of variance σ^2_{eff}. All that is required experimentally are a sufficient number of trials (MacRae 1970) or data points to estimate σ_{eff} for a given number of categories, and then the remainder of the "missing" data can be simulated using σ_{eff}. This approach underwent further refinements in which a chi-squared test was introduced to analyse the variability in the estimate of σ_{eff} from a limited number of trials (Norwich, 1997 Pers. Comm.). It was demonstrated that the variability in σ_{eff} is far smaller than the bias in the information for an equivalent number of trials.

We have continued this approach in this paper, albeit from a theoretical rather than a computational point of view. Since it was demonstrated that σ_{eff} for any number of categories can, in principle, be derived from a single value of σ, the determination of the magical number ideally involves only the determination of σ with a relatively few number of trials (say only 500 for σ, compared with potentially 10^5 or more to determine an entire matrix).

4.2. Violation of the magical number?

Contrary to the common belief that transmitted information should increase monotonically with increasing number of categories, we saw in Fig. 5 that the theory actually predicts that there is a point which transmits a higher quantity of information than the asymptote. To demonstrate theoretically that such a critical point can exist under nominal parameter values turned out to be very difficult. Furthermore, we cannot, at present, confirm whether these critical points will remain upon further refinement of the theory to include edge-effects.

However, there does appear to be some experimental evidence which confirms our prediction. We examine several results quoted by Miller in his 1956 paper: the experiment of Beebe-Center et al. (1955), and of Pollack (1952). As with Fig. 5, we fit transmitted information as a function of m with a single parameter σ/a using (21). Both the data and the theoretical predictions are shown in Fig. 6a ($\sigma/a = 0.14$) and b ($\sigma/a = 0.094$). Except for the last point in Fig. 6a, the theory seems to do exceedingly well despite its simplicity.

To our knowledge, many if not all of the earlier studies on the magical number phenomenon predict that transmitted information should rise monotonically with increasing input information. Therefore, it would be of some interest to compare the various studies to understand what is the cause for this difference. However, it has been noted that several of the older experiments have not yet been corrected for statistical bias due to an insufficient number of trials (MacRae 1970). This is an important point to keep in mind when our theory is compared to experimental data.

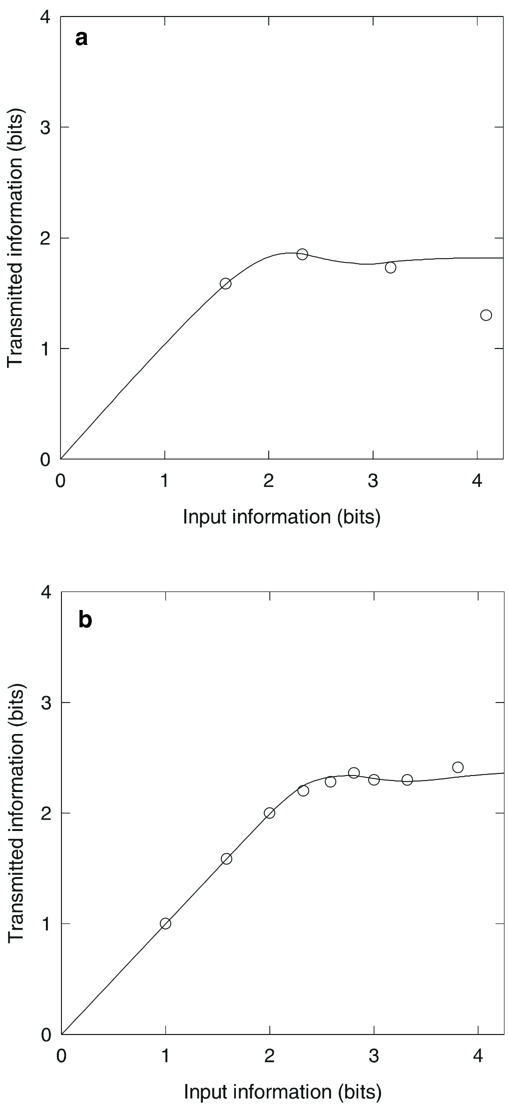

Fig. 6 **a** The data of Beebe-Center, Rogers and O'Connell (1955) for taste. The theoretical curve was obtained by curve-fitting the data with (21) for a single parameter σ. **b** The data of Pollack (1952) for pitch. Again, the theoretical curve was obtained by curve-fitting the data with (21) for σ. Apart from the very last point in Fig. 6a, we see that the theory does quite well

4.3. Sources of error

By and large, we believe that many of the errors that limit the comparison between theory and experiment enter at the level of the single subject. While the estimation of information or variance is constrained by certain basic principles of statistics (like chi-squared statistics), there is a substantial variability not accounted for by statistics alone, including drift in the subject's response (mean response changing with time), shifting of the response criteria and differences in performance level – say, σ improving with experience (see, for example, Weber et al. 1977) or worsening with fatigue as the experiment proceeds. Thus, the consistency of response (between trials or between different sessions) is difficult to quantify and would not only affect the results of the measurements, but also lead to questioning the validity of the assumptions of the theory (fixed criteria, fixed mean response, etc.). Furthermore, the effects of sequential dependencies to past inputs and responses remain difficult to quantify or to take into consideration, but may very well affect the estimation of transmitted information.

With respect to the theory, we have ignored all effects of the boundary on the subject response. This is certainly not a realistic assumption. However, the results tend to indicate that this is in fact not a bad approximation. Future studies can perhaps address this issue and offer a more realistic approach to edge-effects although perhaps at the expense of even greater mathematical complexity.

5. Conclusions

We have attempted to simplify the paradigm by which the upper limit in information transmission has been explored in order to better understand the perceptual characteristics determining the magical number. By proposing a theory which assumes that the subject's response to a sensory input is both continuous and normally distributed, we have demonstrated that an entire categorical matrix can, under approximate conditions, be characterized by a single value of variance associated with the response distribution. Thus, the results of almost any experiment spanning any number of categories can be predicted with this single value of variance provided that the range of the experiment remains unchanged. As demonstrated theoretically, the upper limit in transmitted information defining the magical number for both large range and large number of categories is a function only of how quickly the variance of response grows with range.

We have also made several proposals on how experiments on absolute identification (particularly for the determination of the magical number) can be more effectively conducted in future investigations. Furthermore, the theory predicts a little known characteristic of the magical number in which the transmitted information does not follow a monotonic rise with increasing input information.

Acknowledgments. We would like to thank Prof. A.E. Mortensen for a very careful and methodical review of our manuscript. Furthermore, we would like to thank Prof. Kenneth Norwich and Mr. Elad Sagi for their many helpful comments on an earlier draft of the paper. W.W. would like to acknowledge the support of the Japan Society for the Promotion of Science for his visiting fellowship to Japan. Both authors acknowledge the support of operating grants from the Japanese Ministry of Education, Science and Culture (nos. 09710054 and 97183).

Appendix A

The integrated response, $g(k)$

In Fig. 2, recall that the response distribution was assumed to be normally distributed with zero mean and variance σ^2. Thus, we define the response distribution $N(y;\sigma)$ to be

$$N(y;\sigma) = \exp\left(-y^2/2\sigma^2\right)/\sqrt{2\pi\sigma^2} \tag{30}$$

We now set up the criteria (cf. Fig. 3) for the case of m categories. Let us denote categories $1, 2 \ldots m$ by the index k. The criteria are placed at the midway point between adjacent stimulus categories,

$$\begin{array}{cc} \text{Criteria} & \text{Category} \\ \left[-a, -a + \frac{a}{m-1}\right] & k = 1 \\ \left[-a + \frac{a}{m-1} + \frac{2(k-2)a}{m-1}, -a + \frac{a}{m-1} + \frac{2(k-1)a}{m-1}\right] & k = 2, 3, \ldots m-1 \\ \left[a - \frac{a}{m-1}, a\right] & k = m \end{array} \tag{31}$$

Next, we define the function $g(k)$, which is the area underneath the normal distribution $N(y;\sigma)$ between the points α_{k-1} and α_k,

$$g(k) = \int_{\alpha_{k-1}}^{\alpha_k} N(y;\sigma)dy = \frac{1}{2}\text{erf}\left(\alpha_k/\sqrt{2\sigma^2}\right) - \frac{1}{2}\text{erf}\left(\alpha_{k-1}/\sqrt{2\sigma^2}\right) \tag{32}$$

where α_{k-1} and α_k are the endpoints defined by the criteria. Thus,

$$g(k) = \begin{cases} \frac{1}{2}\text{erf}\left(\frac{a}{\sqrt{2\sigma^2}}\right) - \frac{1}{2}\text{erf}\left(\frac{a(m-2)}{\sqrt{2\sigma^2}(m-1)}\right), & k = 1 \text{ and } k = m \\ \frac{1}{2}\text{erf}\left(\frac{a(2k-m)}{\sqrt{2\sigma^2}(m-1)}\right) - \frac{1}{2}\text{erf}\left(\frac{a(2k-m-2)}{\sqrt{2\sigma^2}(m-1)}\right), & k = 2, 3 \ldots m-1 \end{cases} \tag{33}$$

Recall that $g(k)$ is a probability function centered about $k = (m+1)/2$. Since k is an integer, we see that $g(k)$ is defined only for odd values of m. Thus, at best, we can calculate $g(k)$ at odd values and then introduce a spline to obtain the even values. This is, unfortunately, one limitation of the theory. The same problem is encountered in the calculation of the variance σ^2_{eff} from $g(k)$.

Demonstrating equation (23)

The variance σ^2_{eff} as calculated from $g(k)$ takes on the form

$$\sigma^2_{\text{eff}} = \sum_k \left[k - \frac{1}{2}(m+1)\right]^2 g(k) \tag{34}$$

Using the Euler-MacLaurin summation formula, we have, approximately,

$$\sum_{k=1}^{m} \left[k - \frac{1}{2}(m+1)\right]^2 g(k) \simeq \int_1^m \left[k - \frac{1}{2}(m+1)\right]^2 g(k)\,dk \tag{35}$$

Next, we take an asymptotic expansion of this equation for large m. While this calculation is simple conceptually, it is exceedingly tedious to compute and can be best carried out with a symbolic manipulator like Maple or Mathematica. An expansion to highest order in m would yield

$$\sigma^2_{\text{eff}} \simeq \left[\frac{\sigma^2\text{erf}\left(\frac{a}{\sqrt{2\sigma^2}}\right)}{4a^2} - \frac{\sqrt{2\sigma^2}}{4\sqrt{\pi}a\exp(a^2/2\sigma^2)}\right] m^2 \tag{36}$$

where we have ignored all terms of order m or lower. We then observe that since almost all of the response distribution lies between the boundaries of the matrix, we have $\sqrt{2}\sigma/a \ll 1$ (standard deviation is much smaller than the range). Thus, $\text{erf}(a/\sqrt{2\sigma^2}) \simeq 1$ and $\exp(-a^2\ /2\sigma^2) \simeq 0$, and (36) becomes

$$\sigma^2_{\text{eff}} \simeq m^2\sigma^2/4a^2 \tag{37}$$

Using $R = 2a$, we can write

$$\sigma_{\text{eff}} = m\sigma/R \tag{38}$$

which is (23).

We can now check the validity of our assumption $\sqrt{2}\sigma/a \ll 1$. For example, in Fig. 5, the value of $\sigma = 4.66$ dB was used to generate the theoretical curve. Since the experiment was conducted over a $R = 95$ dB range, we have $\sqrt{2}\sigma/a \simeq 0.14$ with $\text{erf}(a/\sqrt{2\sigma^2}) \simeq \text{erf}(7) \simeq 1$ and $\exp\left(-a^2/2\sigma^2\right) \simeq \exp(-52) \simeq 0$.

Appendix B

To extract the average row variance σ^2_{eff}, we used the following method. Since it is assumed that the variance along each row is constant, we ignored the edge-effects and pooled the data from all rows to obtain a more robust estimation of σ^2_{eff}. In doing so, the values of y_k had to be adjusted for the different mean x_j along each row. Thus, we took the response y_k and subtracted away the value of the mean response x_j. The data from all rows were then pooled to obtain a single distribution $p_{k,\text{pooled}}$ with zero mean. k spans the range $-(m-1)\ldots(m-1)$ to account for all the data that have been shifted. The variance was then calculated by the usual formula,

$$\sigma^2_{\text{eff}} = \sum_{k=-(m-1)}^{m-1} k^2 p_{k,\text{pooled}} \tag{39}$$

7 References

Beebe-Center JG, Rogers MS, O'Connell DN (1955) Transmission of information about sucrose and saline solutions through the sense of taste. J Psychol 39:157–160

Braida LD, Durlach NI (1972) Intensity perception. II. Resolution in one-interval paradigms. J Acoust Soc Am 51:483–502

Braida LD, Durlach NI (1988) Peripheral and central factors in intensity perception. In: Edelman GM, Gall WE, Cowan WM (eds) Auditory function. Wiley, New York, pp 559–583

Carlton AG (1969) On the bias of information estimates. Psychol Bull 71:108–109

Durlach NI, Braida LD (1969) Intensity perception: I. Preliminary theory of intensity resolution. J Acoust Soc Am 46:372–383

Eriksen CW, Hake HW (1955) Absolute judgments as a function of stimulus range and number of stimulus and response categories. J Exp Psychol 49:323–332

Garner WR (1953) An information analysis of absolute judgments of loudness. J Exp Psychol 46:373–380

Garner WR, Hake HW (1951) The amount of information in absolute judgments. Psychol Rev 58:446–459

Houtsma AJ (1983) Estimation of mutual information from limited experimental data. J Acoust Soc Am 74:1626–1629

Lacouture Y, Marley AAJ (1995) A mapping model of bow effects in absolute identification. J Math Psychol 39:383–395

Laming D (1984) The relativity of "absolute" judgments. Br J Math Stat Psychol 37:152–183

Luce RD, Green DM, Weber DL (1976) Attention bands in absolute identification. Percept Psychophys 20:49–54

MacRae AW (1970) Channel capacity in absolute identification tasks: an artifact of information bias? Psychol Bull 73:112–121

Marley AAJ, Cook VT (1984) A fixed rehearsal capacity interpretation of limits on absolute identification performance. Br J Math Stat Psychol 37:136–151

Miller GA (1956) The magical number seven plus or minus two: some limits on our capacity for processing information. Psychol Rev 65:81–97

Miller GA, Madow WG (1954) On the maximum likelihood estimate of the Shannon-Wiener measure of information. (AFCRC-TR-54-75) Operational applications laboratory, Air Force Cambridge Research Center

Mori S (1989) A limited-capacity response process in absolute identification. Percept Psychophys 46:167–173

Mori S (1991) Dynamic information model of identification performance. PhD thesis, University of British Columbia, Vancouver

Mori S, Ward LM (1995) Pure feedback effects in absolute identification. Percept Psychophys 57:1065–1079

Norwich KH (1993) Information, sensation and perception. Academic Press, San Diego

Norwich KH, Wong W, Sagi E (1997) Range as a factor determining the information of loudness judgments: Overcoming small sample bies. Can J Exp Psych (in press)

Pollack I (1952) The information of elementary auditory displays. J Acoust Soc Am 24:745–749

Rogers MS, Green BF (1955) The moments of sample information when the alternatives are equally likely. In: Quastler H (ed) Information theory in psychology: problems and methods. Free Press, Glencoe, pp 101–108

Shannon CE, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Urbana

Treisman M (1985) The magical number seven and some other features of category scaling: properties of a model for absolute judgments. J Math Psychol 29:175–230

Weber DL, Green DM, Luce RD (1977) Effects of practice and distribution of auditory signals on absolute identification. Percept Psychophys 22:223–231

Wong W, Mori S (1998) The magical "wave" seven, plus or minus two? J Acoust Soc Am (in press)

Wong W, Norwich KH (1997) Simulation of human sensory performance. Biosystems 43:189–197