Representational Capacity of Face Coding in Monkeys

L. F. Abbott,[1] Edmund T. Rolls,[1] and Martin J. Tovee[2]

[1]Department of Experimental Psychology, Oxford University, Oxford OX1 3UD, United Kingdom, and [2]Department of Psychology, University of Newcastle, Newcastle upon Tyne, NE1 7RU, United Kingdom

We examine the distributed nature of the neural code for faces represented by the firing of visual neurons in the superior temporal sulcus of monkeys. Both information theory and neural decoding techniques are applied to determine how the capacity to represent faces depends on the number of coding neurons. Using a combination of experimental data and Monte Carlo simulations, we show that the information grows linearly and the capacity to encode stimuli grows exponentially with the number of neurons. By decoding firing rates, we determine that the responses of the 14 recorded neurons can distinguish between 20 face stimuli with approximately 80% accuracy. In general, we find that *N* neurons of this type can encode approximately $3(2^{0.4N})$ different faces with 50% discrimination accuracy. These results indicate that the neural code for faces is highly distributed and capable of accurately representing large numbers of stimuli.

The amount of information that can be represented by the firing of a population of neurons depends on the nature of the neural code. In particular, the representational capacity is extremely sensitive to how information is distributed across the population of coding neurons. If each stimulus is represented by the firing of a single neuron or "grandmother cell," the number of stimuli that can be represented is proportional to the number of neurons. If the information about each stimulus is distributed across the full population, the number of stimuli that can be represented grows exponentially with the number of coding neurons. For example if the responses of each neuron can reliably divide the stimuli into two groups of equal size and if N neurons respond independently, the population response should be able to distinguish 2^N different stimuli. Intermediate coding strategies and representational capacities, such as power law dependencies, are also possible.

Demonstrating that individual neurons respond to a wide variety of stimuli or that large numbers of neurons respond to individual stimuli is not sufficient to establish the existence of a truly distributed representation. Distributed coding with its associated exponentially large capacity requires that the differences in the broadly tuned responses of individual neurons are not masked by their trial-to-trial variability. In addition, each neuron must have a distinctive response profile across stimuli so that the population coding is not excessively redundant. In the example of the last paragraph, the capacity will be reduced greatly if the responses of different neurons divide the stimuli into the same two groups. The clearest way to determine how information is distributed in a neural network is to measure how the representational capacity of the network depends on the number of coding neurons.

We use experimental recordings augmented by Monte Carlo simulations to analyze the coding of faces by visual neurons in the temporal cortex of macaque monkeys. Temporal lobe visual areas are at a late stage in the ventral visual pathway (Seltzer and Pandya, 1978; Maunsell and Newsome, 1987; Baizer et al., 1991; Rolls, 1991). In cortical areas of the superior temporal sulcus up to 20% of the neurons with visual responses have selectivity for faces (Desimone and Gross, 1979; Bruce et al., 1981; Perrett et al., 1982; Desimone et al., 1984; Rolls, 1984; Gross et al., 1985; Desimone, 1991). We analyze the capacity of these neurons to represent faces in two ways: by determining how many stimuli can be represented to a given degree of accuracy (Bialek et al., 1991; Salinas and Abbott, 1994) and by computing the amount of information that the responses can convey about the stimuli (Eckhorn and Popel, 1974, 1975; Optican and Richmond, 1987; Richmond and Optican, 1990; Optican et al., 1991; Hertz et al., 1992; Tovee et al., 1993; Kjaer et al., 1994).

Other investigations have found that the information carrying capacity of inferior temporal neurons grows more slowly than a linear function of the number of neurons (Gochin et al., 1994; E. T. Rolls, A. Treves, and M. J. Tovee, unpublished observations). This would suggests that face coding is not fully distributed. The work of Rolls et al. (unpublished observations) is based on the same data examined here. [Other work by Rolls et al. (unpublished observations) examines the information in single neuron responses for this data set and is not directly related to the issue studied here, the dependence of the information on the number of neurons.] However, their analysis differs from the present one in two significant ways. First, the method used to compute the information is completely different in the two articles (see below). Second, the work of Rolls et al. (unpublished observations), like that of Gochin et al. (1994), finds a sublinear growth of the information as a function of cell number. When small numbers of stimuli are involved we obtain a similar result here. However, we find that this slow growth is an artifact of the limited size of the stimulus set. To get around this limit, we introduce a method for simulating additional stimuli on the basis of existing data. When large numbers of stimuli are included by using this method, we find that the information grows linearly with the number of neurons. Similarly, the number of stimuli that can be represented to a given degree of accuracy increases exponentially with the number of coding neurons. Our results reveal a truly distributed code for faces. On the basis of the experimental data, we find that the number of stimuli that can be represented with a 50% discrimination accuracy by N cells is approximately $3(2^{0.4N})$ corresponding to 0.4 bits of information per neuron.

Materials and Methods

Our results are based on recordings of 14 neurons in the superior temporal sulcus of two rhesus macaques, Macaca mulatta. Single neuron responses to 20 images of monkey and human faces were recorded during a visual fixation task. Firing rates were determined by counting spikes over a 500 msec period, starting 100 msec after stimulus presentation. Responses were recorded over an average of 10 trials for each cell and face stimulus. The selected neurons fired in response to faces at more than twice the maximum rate evoked by any of 48 other nonface images (Rolls, 1984). This set of neurons has been described previously (Rolls and Tovee, 1995), and further experimental details and results can be found in Rolls et al. (unpublished observations).

Information represented by neuronal firing has been computed by a number of different techniques (Eckhorn and Popel, 1974, 1975;

Optican and Richmond, 1987; Richmond and Optican, 1990; Optican et al., 1991; Hertz et al., 1992; Tovee et al., 1993; Kjaer et al., 1994). Various methods of correcting for the effects of small sample size have been used ranging from subtracting shuffled data sets (Optican and Richmond, 1987; Richmond and Optican, 1990; Optican et al., 1991; Tovee et al., 1993) to the use of analytic results (Treves and Panzeri, 1995; Rolls et al., 1995a,b), a neural network approach (Hertz et al., 1992), and decoding methods (Kjaer et al., 1994). Here we present and employ a new method based on Monte Carlo integration of extracted probability distributions. The advantages of this approach are that it is simple and direct, that it can be checked by a number of internal consistency tests and that it does not rely on subtraction procedures of unknown validity.

Our analysis consisted of three steps.

(1) We computed a mean firing rate and firing rate variance for each of the 280 possible cell and stimulus combinations by averaging over all trials. From these means and variances we constructed Gaussian distributions for each cell describing the probability that a particular face evokes a given firing rate. We checked the quality of these Gaussian fits using the Kolmogorov-Smirnov test.

(2) We applied two different methods to evaluate the representational capacity of the recorded neurons. First, we used the firing rate probability distributions to compute how much information the responses of different numbers of neurons could carry about the set of face stimuli. In addition, we used decoding methods (described below) to determine how accurately the firing responses could distinguish between different stimuli and how many stimuli could be represented by such responses. The results of these two methods are logarithmically related. If the number of stimuli that can be represented by N neurons is proportional to 2^{aN}, the information content is a bits per neuron.

(3) We modeled how the responses of the recorded neurons varied across stimuli. This allowed us to generate hypothetical responses to additional simulated stimuli not present in the original data set. Through this technique we extrapolated to large numbers of stimuli.

Firing Rate Probability Distributions

Our analysis of the representational capacity of face coding is based on a determination of the probability that a particular stimulus s evokes a set of firing rates $\mathbf{r}$ in the recorded neurons. We denote this probability by $P(\mathbf{r}|s)$. Since neurons in the sample we used were recorded one at a time, trial-to-trial fluctuations are unlikely to be correlated between different neurons. As a result, the firing-rate probability for the population $P(\mathbf{r}|s)$ can be written as the product of probabilities for the individual neurons, which we denote by $P_i(r|s)$. From the measured average firing rates and firing rate variances for each recorded neuron and each stimulus we constructed Gaussian probability distributions $P_i(r|s)$. Since these Gaussian distributions have a finite probability for negative firing rates, we interpreted all negative firing rates as zero.

Kolmogorov-Smirnov Test

The quality of the Gaussian fits to the experimental data was checked using the Kolmogorov-Smirnov test, which is based on a comparison of two cumulative probability curves, $S_1(R)$ and $S_2(R)$. $S_1(R)$ is the integral of the probability distribution $P_i(r|s)$ from zero to R and $S_2(R)$ is a stairstep curve, which is the fraction of trials with rates less than R. The K-S measure for the quality of fit is the maximum distance between these two curves. A separate test was performed for each cell and stimulus combination. Simple corrections were made for the fact that experimental rates were derived from non-negative integer spike counts over a 500 msec interval while the Gaussian probability distribution generates real number rates.

Tables indicating the statistical significance of different K-S distances exist but they do not apply to cases like ours where means and variances are extracted from the data being tested. Instead, we used a Monte Carlo procedure for this purpose (Press et al., 1992). For each cell and stimulus, we used the probability distribution $P_i(r|s)$ to generate simulated trials. We then fit these simulated trials to Gaussian distributions and measured the resulting Kolmogorov-Smirnov distances. We repeated this process 1000 times and recorded how often the Monte Carlo generated K-S measure was greater than the measure obtained from the real data. This percentage gives the probability that fits worse than those for the real data would arise by chance if Gaussian distributions are the correct description. We tested 280 probability distributions this way. If the Gaussian distributions provide an acceptable fit, the number of distributions with K-S distances that are smaller than their Monte Carlo counterparts x% of the time should be about x% of 280 for any value of x.

Information Calculation

The probabilities $P_i(r|s)$ that we extract for each cell and stimulus completely characterize the distributed character of the neural code. The breadth of the probability function for a given neuron and stimulus indicates how "noisy" the encoding is. Comparison of the probability distributions of a given neuron for different stimuli reveals the selectivity of individual neurons, while comparison across neurons for a given stimulus determines the level of redundancy of the code. Of course, these three features are highly interrelated. A useful statistic that accounts for all these effects and features is Shannon's mutual information. The mutual information is a functional of the conditional response probability $P(\mathbf{r}|s)$ given by

$$I_r = \sum_{s,r} P(s)P(\mathbf{r}|s) \log_2\left(\frac{P(\mathbf{r}|s)}{P(\mathbf{r})}\right), \qquad (1)$$

where the sum is over all stimuli and all possible responses. In this equation, $P(s)$ is the probability of a particular stimulus appearing which, in our case, is one over the number of stimuli. $P(\mathbf{r})$ is the probability of the response set $\mathbf{r}$, which is the sum of $P(s)P(\mathbf{r}|s)$ over all stimuli. Note that the information is given by a sum over all responses not just over those in the original data set. Indeed, once the probability distributions $P(\mathbf{r}|s)$ have been extracted, the information does not depend on the responses in the original data set at all. The subscript r in Equation 1 stands for the "raw" information measure to distinguish it from a "cross-validated" form of the information that we will introduce below.

The sum over rates in Equation 1 involves all the possible firing rates for all of the neurons being considered. When we include all 14 recorded cells in our analysis this is a 14-dimensional sum. To handle this high dimensionality we used Monte Carlo methods (Press et al., 1992). In this procedure, we randomly generated firing rates for each cell from the probability distributions $P_i(r|s)$. The logarithm in Equation 1 was then averaged over repeated iterations. For sufficiently large numbers of iterations this procedure is equivalent to doing the sum. Typically, 500–5000 iterations were used, although fairly accurate results could be obtained using a smaller number. In addition to computing the information for all 14 recorded cells, we computed the information with smaller numbers of neurons. To do this we averaged over randomly chosen subsets of the recorded neurons. We also computed the information for single neurons.

Decoding Methods

Decoding is a procedure for determining which stimulus evoked a particular set of firing rates (Bialek et al., 1991; Salinas and Abbott, 1994). It can be used to determine the accuracy with which stimuli are represented by neuronal firing. We define the discrimination accuracy as the percentage of times that a decoding algorithm is able to extract the correct stimulus from a set of rates. To compute the discrimination accuracy, we used the probability distributions $P_i(r|s)$ to generate firing rates corresponding to a given stimulus. We then used a decoding procedure to determine which stimulus was most likely to have produced this particular set of rates. Finally, we compared the decoded stimulus with the actual stimulus used to generate the rates and, repeating the procedure, computed the percentage of correct decodings. The decoding approach used in this analysis was the maximum likelihood method known to be optimal in many cases. The stimulus s that we associated with a given response set $\mathbf{r}$ was the one that maximized the probability $P(\mathbf{r}|s)$. Because the probability distributions we used are Gaussian, the maximum likelihood method is equivalent to a weighted least-squares estimate of the stimulus from the rates. (This estimate is the square of the difference between the actual rate and the mean rate divided by the variance summed over all cells for a given stimulus. The decoded stimulus is the one that minimizes this sum.)

For comparison purposes, in the discussion section, we also used a linear decoding scheme. In this case we computed how well the actual neural responses aligned with their average values for each stimulus. The set of responses was assemble into a vector, and another vector represented the average firing rates for each stimulus. The alignment was computed from the cosine of the angle between these

two vectors. The decoded stimulus was the one with the best alignment. This is not an optimal coding procedure (Salinas and Abbott, 1994) but, because it is linear, it is one that can easily be implemented by a neural network.

Finite Sample Corrections

Small data sets tend to produce overestimates of the information and discrimination accuracy (Macrae, 1971; Optican et al., 1991; Hertz et al., 1994; Treves and Panzeri, 1995). It is easiest to see why this happens for the case of the discrimination accuracy. In our calculations, the probability distribution $P(\mathbf{r}|s)$ extracted from the data is used both to generate firing rates and to determine the most-likely stimulus to have produced those rates. Inevitably, the Monte Carlo responses fit the distribution $P(\mathbf{r}|s)$ that generated them better than they fit the true probability distribution. This results in an overestimate of the percentage of correct decodings.

The problem of overfitting is far more severe if probability distributions are determined by binning than by fitting to a parameterized curve as is done here. In cases where binning has been used, a subtraction procedure has been applied in an attempt to correct for overfitting (Optican et al., 1991; Treves and Panzeri, 1995). Since we used a two-parameter Gaussian fit of the probability distributions, we do not apply these subtractions. Instead, we bound the actual information and discrimination accuracy by computing upper and lower limits. The upper limit is just the raw calculation discussed above. To generate a lower limit, we used a procedure analogous to the "cross-validation" method of dividing a data set into two parts—one for fitting and one for testing. In this case, rather than dividing our data, we generated a second set by Monte Carlo methods. We did this by using the distribution $P(\mathbf{r}|s)$ to generate simulated data trials that we fit with a second probability distribution $P'(\mathbf{r}|s)$. If the number of simulated trials in this procedure matches the number of trials in the experimental data set, the difference between these distributions is a rough characterization of the difference between $P(\mathbf{r}|s)$ and the true probability distribution for the experimental data.

We generate a lower limit by using the original distribution $P(\mathbf{r}|s)$ to generate Monte Carlo responses while using $P'(\mathbf{r}|s)$ to compute the information or to do the decoding. To compute the cross-validated discrimination accuracy, we generated responses, as before, using the distribution $P(\mathbf{r}|s)$. However, we decoded these responses by determining which stimulus maximized $P'(\mathbf{r}|s)$ for this set of responses rather than $P(\mathbf{r}|s)$. For the case of the information calculation, we computed a "cross-validated" form of the information from the formula

$$I_{cv} = \sum_{s,\mathbf{r}} P(s)P(\mathbf{r}|s) \log_2\left(\frac{P'(\mathbf{r}|s)}{P'(\mathbf{r})}\right). \quad (2)$$

In the Monte Carlo calculation of this information, we used the original distribution $P(\mathbf{r}|s)$ to generate responses for estimating the sum, but average the logarithm involving $P'(\mathbf{r}|s)$ over these responses. Finally, for both the discrimination accuracy and information calculations, a large number of iterations (typically 50) with repeated generation of $P'(\mathbf{r}|s)$ distributions were averaged. The cross-validation procedure produces an underestimate of the true result because the two distributions P and P' tend to differ from each other more than they differ from the true distribution describing the data.

The cross-validated information of Equation 2 does not represent a rigorous lower bound on the information but, in practice and through Monte Carlo simulations (see below), we have found that the actual information is always greater than the information provided by this expression. To verify that the raw and cross-validated results really bound the true answer and to examine the discrepancy between them, we used the probability distribution $P(\mathbf{r}|s)$ to simulate experiments with various numbers of trials. In these simulations, the "true" information could be computed exactly and compared with the results of different computational methods. Simulated data were generated from the probability distribution $P(\mathbf{r}|s)$ and from these "data" new probability distributions were extracted and the information was computed just as it was for the real data. We include the results of both the raw and cross-validation computations in all our figures. The two are easy to distinguish because the raw results always show larger information and higher discrimination accuracy than the cross-validated computations and the two results provide upper and lower bounds on the actual answer.

Generation of Simulated Stimuli

Establishing the exponential growth of the representational capacity that is the hallmark of distributed coding requires knowledge of the responses of cells to a large number of stimuli. In recordings of cortical neurons, it may be virtually impossible to study a large enough stimulus set to achieve this goal. To circumvent this problem, we generated responses to hypothetical stimuli not in the data set. Recall that the responses of the recorded neurons to the face stimuli in the experiment were characterized by Gaussian distributions that depended only on the means and variances of the fire rates across trials. Similarly, the responses to additional simulated stimuli were assumed to have Gaussian statistics. Thus, to produce them we only needed to generate average firing rates and standard deviations corresponding to new hypothetical stimuli. This was done on the basis of the statistical properties of the responses to real stimuli.

Specifically, we generated the average firing rate responses for the simulated stimuli from a Gaussian distribution that satisfied two conditions: (1) each cell had the same average response over all simulated stimuli as it did over all real stimuli, and (2) the 14 by 14 cell-cell correlation matrix summed over real stimuli matched the same matrix summed over simulated stimuli. Average responses for the simulated stimuli were generated by diagonalizing the correlation matrix (using Maple software). In the diagonalized representation, individual components are independent and can be obtained from Gaussian distributions with variances equal to the eigenvalues of the correlation matrix. After these random components were generated, they were transformed back to the original basis yielding Gaussian random variables with the desired correlations. To generate the standard deviations of the responses to simulated stimuli, we incorporated the observation that, in the original data, the standard deviation of the response to a real stimulus grew in proportion to the average firing rate evoked by that stimulus. We extracted the proportionality constant for this relation and also the variance around it from the data for each cell. The firing rate variances for the simulated stimuli were then determined from a Gaussian distribution constructed from these results.

Results

Spike Rate Distributions

Firing rate averages and standard deviations for four of the 14 recorded cells are shown in Figure 1. Six of the recorded cells showed strongly graded responses like those shown in Figure 1*a*. Three cells displayed more weakly graded responses as in Figure 1*b*. Three cells had graded responses that tended to cluster into two or more groups. One such cell is shown in Figure 1*c*. Finally, two of the cells showed "grandmother"-like responses (at least over the limited number of face images used) where one stimulus evoked a considerably different response than all the others in the stimulus set as in Figure 1*d*. For all 14 neurons, the standard deviation across stimuli was only between one and two times the trial-to-trial deviation. These two features have opposite effects on the representational capacity.

Like excessive trial-to-trial variability, correlations between the responses of different neurons to the same stimuli reduce the representational capacity because of the resulting redundancy. To measure the redundancy of the responses we computed the Pearson product-moment correlation matrix of the average responses of the different neurons to the same stimulus, summed over all stimuli. Off diagonal elements, indicating correlations, were as large as 0.9, although most tended to lie between about $+0.5$ and -0.5. Thus, the representation of faces provided by this set of neurons is somewhat redundant. Further analysis presented below determined how much this redundancy reduced the representational capacity for faces and whether it was large enough to preclude the possibility of truly distributed coding with exponential capacity.

As discussed in the Materials and Methods section, the trial-to-trial variability in the neuronal firing responses was fit by Gaussian distributions matching the observed mean firing

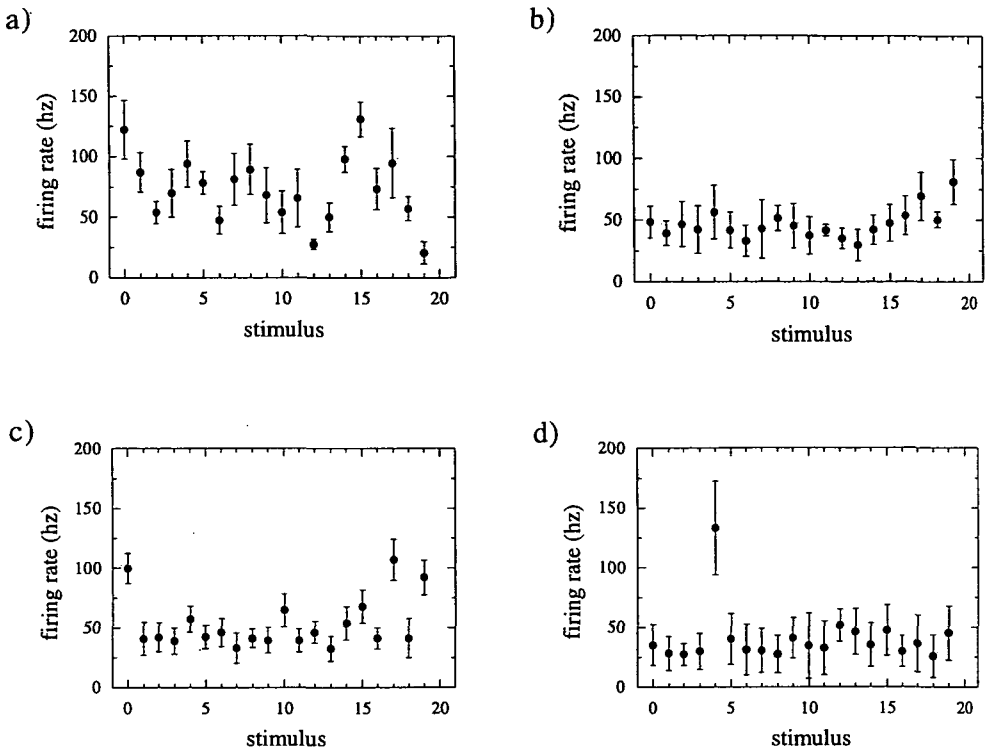

Figure 1. *The average firing rates and* standard deviations for 4 of the 14 recorded cells responding to 20 face stimuli. Stimuli are numbered 0 to 19. Rates are in spikes per second. *a,* One of six neurons with a highly graded response profile. *b,* One of three neurons with a weakly graded response. *c,* One of three neurons that showed graded responses that tended to fall into groups. *d,* One of two "grandmother"-like cells.

rates and variances. The quality of these fits was checked by a Kolmogorov-Smirnov test for all 14 cells and 20 stimuli, resulting in 280 results. Based on the hypothesis that these Gaussian fits are appropriate, we found that, in repeated experiments, the data would produce fits equivalent or worse than the ones we found (as measured by the K-S measure) 45% of the time. This is well within the acceptable range. There was a slight excess of cases with low probabilities. The quality of fit for 33 of the distributions would only have arisen by chance 10% of the time. We would have expected only 28 fits to lie in this range by chance, but again, the excess is acceptable. Only 3 of the 280 Gaussian fits were unacceptably poor by this measure, that is, they had probabilities low enough that we should not have seen them by chance in this number of tests. These occurred for different cells and stimuli. Because all our results are averaged over both cells and stimuli, these three poor fits had minimal impact on the results reported.

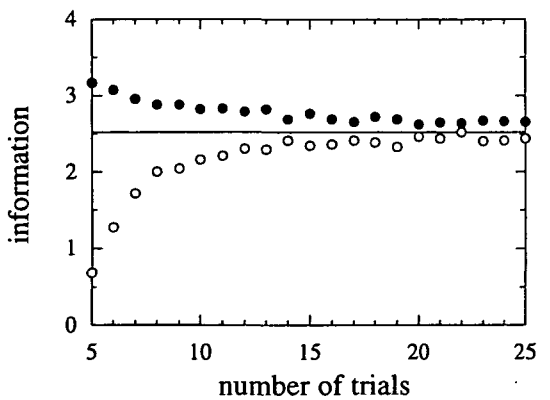

Figure 2. Finite-sample effects on the information calculation. Information was computed from Monte Carlo-generated data to explore the effects of sample size. The results shown are for five cells, but similar results were obtained for other population sizes as well. The *solid circles* show the raw information falling as a function of trial number. The *open circles* indicate the cross-validated information that rises for larger data samples. The *solid line* is the "true" value of the information for this simulated data.

Comparison of Raw and Cross-validated Results

Due to the finite size of our data sample, we cannot compute the information or discrimination accuracy of face coding exactly. However, the raw and cross-validated procedures we use (see Materials and Methods) provide upper and lower limits that bound the exact answer. We verified this by performing Monte Carlo studies of the results of these computations on simulated data where we knew the exact answers. Figure 2 shows a typical case where the raw and cross-validated information calculations are compared with the "true" result for different numbers of simulated data trials. The solid circles in Figure 2 show that for small numbers of trials the raw information is, indeed, an overestimate of the true answer. The results of the cross-validation computation of the information are shown by the lower open circles in Figure 2. Above about 20 trials, the differences between the raw, cross-validated, and "true" information are fairly small, indicating that finite sample effects are under control. Indeed, it appears that the raw information obtained from our Gaussian fits without subtraction or other correction provides a fairly accurate estimate of the true information for 15 or more trials. Below 10 trials the two computations differ significantly from each other and from the "true" result. The raw and cross-validated information provide upper and lower bounds, with the true information always falling between them.

Information and Discrimination Accuracy

The firing response probabilities for each of the recorded neurons characterize their ability to convey information about the stimuli. We first considered their information carrying capacity individually and then in groups of different sizes. In addition, we computed the discrimination accuracy for these groups.

Figure 3*a* shows the information computed from the response probabilities of individual neurons. The neurons have

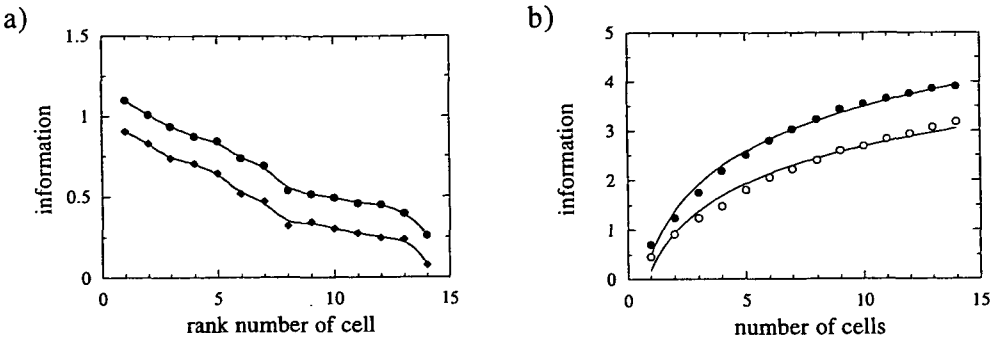

Figure 3. The information and discrimination accuracy for 20 stimuli. In *a* and *b* the *upper curve* is the raw and the *lower curve* the cross-validated information in bits. *a*, The information carrying capacity of single neuron responses for the recorded neurons. The neurons have been ranked in decreasing order of raw information. *b*, The information for random subsets of different numbers of cells. The *curves* are logarithmic fits to the plotted points given by $I_{raw} = 0.47 + 0.9\log_2(N)$ and $I_{cv} = 0.18 + 0.75\log_2(N)$.

been ranked in decreasing order of information so the curves shown are monotonically decreasing. The information from single neurons for the 14 cells recorded using 20 face stimuli range from about one to around 0.1 bits. The cell shown in Figure 1*a* had the highest information measure of any of the recorded cells, which relates well to its highly graded response profile. The six neurons with well-graded responses and the three neurons that showed some grouping in their responses had the nine highest values of information between them. The three neurons with weakly graded responses and the two "grandmother" cells had the lowest information. The cell of Figure 1*c* ranked fifth in terms of information content, the weakly graded neuron of Figure 1*b* ranked eleventh, and the "grandmother" cell shown in Figure 1*d* ranked thirteenth. The information averaged over all single cell results is $I_{raw} = 0.67$ and $I_{cv} = 0.47$.

Figure 3*b* shows the information computed by averaging over different size randomly chosen subpopulations of the 14 recorded cells. Again, both the raw and cross-validated information are shown. The information rises with the number of cells. The curves shown are logarithmic fits. The number of stimuli that can be represented to a given degree of accuracy is proportional to the base-two exponential of the information. The fits in this graph indicate that the number of faces that can be represented by N neurons grows like a power of N and is somewhere between $1.1N^{0.75}$ and $1.4N^{0.9}$. The sublinear fits of the information as a function of the number of cells agree with those found in Goshin et al. (1994) and Rolls et al. (unpublished observations).

Do these results mean that the coding of faces is far from truly distributed coding? There is a complication that prevents us from reaching this conclusion. The information measure is limited by the amount of information contained in the stimulus set, which in our case is $\log_2(20) = 4.32$ bits. The upper points in Figure 3*b* are approaching this limit so that the "roll over" we see in the growth of the information with the number of cells may reflect a saturation effect due to the finite stimulus set rather than the true growth of the representational capacity.

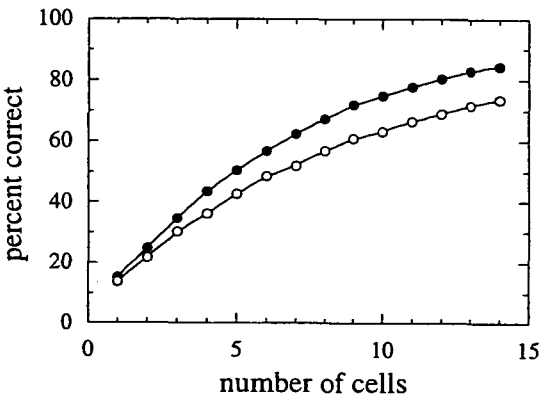

Figure 4. The discrimination accuracy for random subsets of different numbers of cells responding to 20 stimuli. The *upper curve* is the raw result and the *lower curve* the cross-validated result. Discrimination accuracy is defined as the percentage of correctly decoded responses using the optimal maximum likelihood method.

Figure 4 shows the percentage of correct decodings based on the maximum likelihood method when random subsets of different numbers of cells were used. The coding accuracy increases steadily with the number of cells and it ultimately gets fairly close to the bound of 100%. As in the case of the information measure, the rate of increase in the discrimination accuracy may reflect this bound rather than the dependence of the coding on the number of cells.

Extrapolation to Large Numbers of Stimuli

As we saw from Figures 3*b* and 4, the limited number of stimuli in our data set prevents us from determining directly how the representational capacity depends on the number of coding neurons. To overcome this problem we used the statistical properties of the recorded responses for 20 stimuli to generate simulated responses to additional hypothetical stimuli. As discussed in the introductory paragraphs, correlations between the responses of different neurons to the same stimuli play a critical role in determining if the redundancy in the neural representation is too high to produce an exponentially growing representational capacity. Because of this, we have made sure that the cell-to-cell correlations of the responses to simulated stimuli match those of the real stimuli exactly (see Materials and Methods). Including stimulus-stimulus correlations between neurons in the simulated responses allows us to test directly whether the resulting redundancy is large enough to destroy exponential growth of the coding capacity.

Figure 5 shows a comparison of the measured response statistics for the 20 original stimuli and simulated response statistics for 20 more hypothetical stimuli generated as outlined above. For the nine cells that showed either weakly or strongly graded responses the simulated response statistics were indistinguishable from those generated by the real stimuli as is seen in Figure 5, *a* and *b*, for the cells shown in Figure 1, *a* and *b*. We did not attempt to include any clumping of the responses or any "grandmother" responses in our simulated results. As a result, the simulated responses for the three neurons that showed some grouping tendency and the two "grandmother"-like cells did not match quite as well. This is seen in Figure 5, *c* and *d*, which correspond to the neurons shown in Figure 1, *c* and *d*. However, even in these cases the simulated responses are similar to the real responses. We found that removing these worse fits did not significantly affect our results.

As a final check of our procedure for generating hypothetical stimuli we compared the information computed from the 20 real stimuli with that computed using 20 simulated

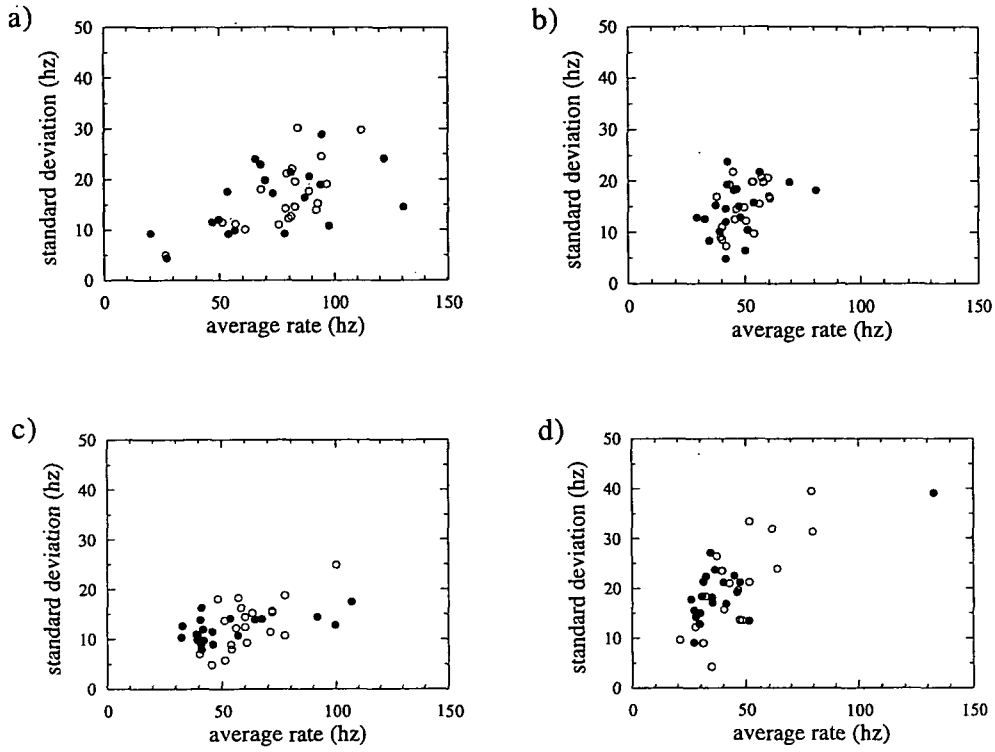

Figure 5. Scatter plots of average firing rates and standard deviations for the 20 original stimuli and for 20 simulated stimuli. *Solid circles* are the actual stimuli in the data set and *open circles* are the simulated points. *a,* The same neuron as in Figure 1*a*. *b,* The same neuron as in Figure 1*b*. *c,* The same neuron as in Figure 1*c*. *d,* The same neuron as in Figure 1*d*.

stimuli. As Figure 6 indicates, the information contained in the generated responses to the simulated stimuli is in good agreement with the information computed from the responses to the real stimuli.

Representational Capacity

With arbitrary numbers of simulated stimuli at our disposal, we repeated the information and discrimination accuracy calculations for different numbers of neurons. The results for the information are given in Figure 7. Figure 7*a* shows the raw information results and 7*b* the cross-validated results. From these figures, it is clear that the finite size of the stimulus set was the factor limiting the rise of the information in Figure 3*b* as the number of coding neurons increased. For large numbers of stimuli, the curves in Figure 7 approach a straight line,

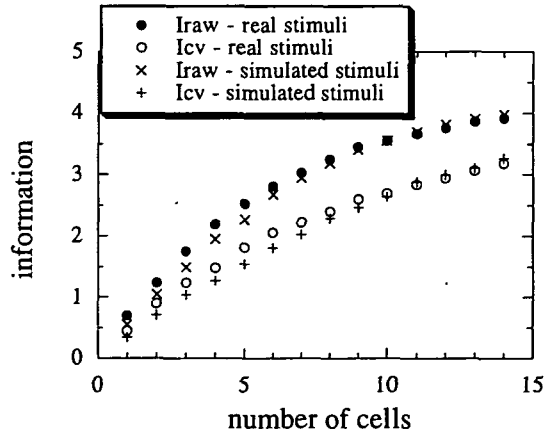

Figure 6. Comparison of the information for the 20 real stimuli and for 20 simulated stimuli. The *solid circles* show the raw information for the real stimuli and the *x*s are the analogous results for the simulated stimuli. The *open circles* show the cross-validated information for the real stimuli and the +s are the corresponding points for the simulated stimuli.

which for the raw information corresponds to 0.5 bits per cell and for the cross-validated information is 0.35 bits per cell.

Figure 8 shows a confirmation of these results using decoding rather than information theory. We determined the number of stimuli that could be decoded to a given accuracy as a function of the size of the neural population. Figure 8 shows the result for 50% discrimination accuracy (half the decoded stimuli matched the true stimuli) with and without cross-validation. A 50% discrimination accuracy is well above chance levels, which would only be one over the number of stimuli. The number of stimuli that can be represented to a given level of accuracy increases exponentially with the number of coding neurons. This was found for other decoding accuracies as well as for the 50% case shown in Figure 8. Furthermore, the exponential fits in Figure 8 agree with the results from the information theoretic analysis. The solid curve fitting the data points for a raw correct percentage of 50% is $2.9(2^{0.47N})$, while the curve for the cross-validated case is $2.9(2^{0.36N})$. These fits indicate that 0.47 and 0.36 bits of information are carried per cell in these two cases, numbers that are in excellent agreement with the results from Figure 7, indicating 0.5 and 0.35 bits per cell. This agreement provides an excellent consistency check on our results since the two methods are quite distinct.

Discussion

Taken together, Figures 7 and 8 provide strong evidence that the coding of face cells by temporal visual neurons in the macaque monkey is truly distributed. This results in an exponential dependence of the representational capacity on the number of cells. Choosing a number midway between the raw and cross-validated results, we estimate that N neurons can represent about $3(2^{0.4N})$ faces with 50% discrimination accuracy. This means that the 14 neurons recorded responding to 20-face stimuli could potentially represent up to 145 faces with this accuracy. The population as a whole carries about 0.4 bits of information about the stimuli per cell. The

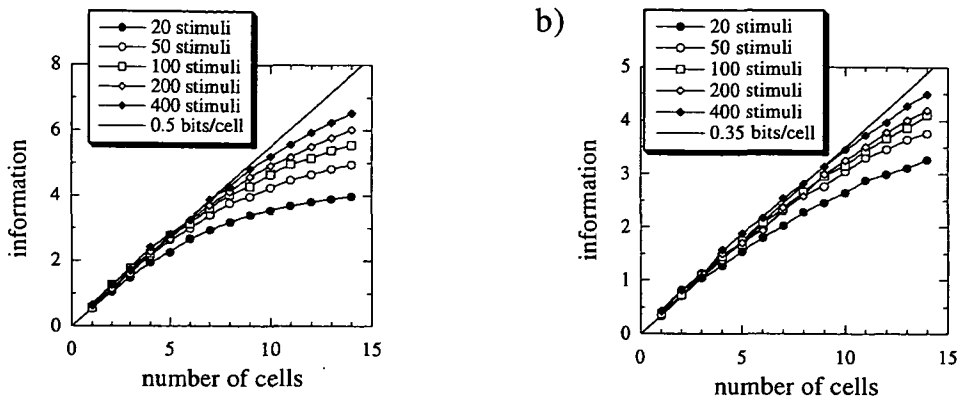

Figure 7. The information as a function of the number of cells computed for different numbers of stimuli. *a*, The raw information is plotted for 20, 50, 100, 200, and 400 simulated stimuli. The *straight line* corresponds to 0.5 bits per neuron. *b*, The cross-validated information is plotted for the 20, 50, 100, 200, and 400 simulated stimuli. The *straight line* corresponds to 0.35 bits per neuron.

average information carried by each neuron considered individually, as in Figure 3*a*, is about 0.6 bits per cell. Redundancy has therefore reduced the information per cell from 0.6 bits to the 0.4 bits seen for the full population. Although this may seem a small change, it has a large impact on the representational capacity because of the exponential dependence. Of course, some redundancy can be useful for combating the effects of noise.

A simple model provides a rough interpretation of these results. Suppose that a neuron responds in only two distinct ways so that a fraction f of the stimuli elicit one response while the remaining fraction $(1 - f)$ evoke the other response. The information for this neuron is $-f \log_2(f) - (1 - f)\log_2(1 - f)$. The value of 0.6 for the average information in single neuron responses corresponds to $f = 0.15$ and the 0.4 bits per neurons across the population gives $f = 0.08$. These values suggest that the responses of individual neurons distinguish about 15% of the stimuli from the others, but in the full population about half of the 15% distinguished by one neuron are distinguished by other neurons as well.

An exponentially growing representation gives remarkably large capacities for even modest numbers of neurons. Indeed, the capacity for faces may seem more that would be required. Even 25 neurons could code around 3000 faces according to our formula. It should be stressed that the representational capacity we are discussing does not imply recognition or identification of an image. The discrimination accuracy only implies that stimuli are represented in such a way that they can be perceived as different from each other. This does not imply that any significance or meaning has been attached to the stimuli. Furthermore, we should probably think of faces as representing a continuum of possible images with graded differences between them rather than a discrete set of images. In this case, the extremely large capacities we have found indicates that the neural coding is capable of representing very subtle differences in this continuum.

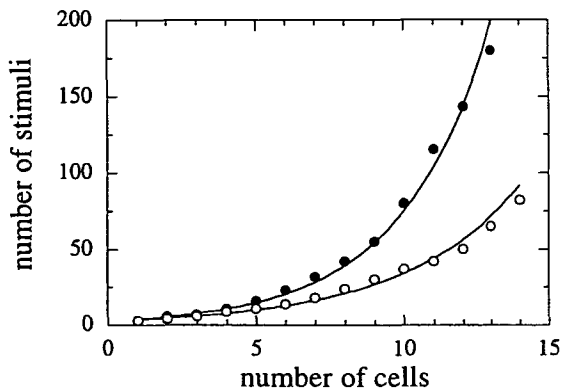

Figure 8. The number of stimuli that can be decoded at 50% accuracy as a function of the number of coding neurons. The *solid circles* are the result of the raw calculation, and the *open circles* correspond to the cross-validated case. The *curves* are exponential fits to the data points.

If our results are extrapolated to larger numbers of neurons than the 14 cells studied, as in the last paragraph, we must, of course, assume that these 14 are representative of the full population. The cells we used had properties that were in complete accord with those of face cells studied previously (Desimone and Gross, 1979; Bruce et al., 1981; Perrett et al., 1982; Desimone et al., 1984; Rolls, 1984; Gross et al., 1985; Desimone, 1991) and they were recorded from different locations within the superior temporal sulcus and in more than one animal. Furthermore, our results are not sensitive to which particular set of cells or stimuli were analyzed when we examined random subsets of the 14 cells or 20 face stimuli used in the experiment.

It may be important that significant amounts of information can be extracted from a small subset of the neurons encoding a particular signal. It has been suggested that correlations in the noise between different neurons severely limits the size of the neuronal pools that can be effectively decoded as a group (Gawne and Richmond, 1993; Shadlen and Newsome, 1994; Zohary et al., 1994). Furthermore, neurons in downstream networks may synapse with only a small fraction of the coding neurons. These downstream neurons must therefore react to the responses of a limited number of coding neurons. In this case, the number N in our formula for the representational capacity corresponds to the number of coding neurons being readout by neurons in the downstream network not to the full size of the coding population. It is difficult to assess the impact of correlated noise (Gawne and Richmond, 1993; Shadlen and Newsome, 1994; Zohary et al., 1994) on decoding accuracy. If correlations in the noise represent variations in the overall excitability of the coding neurons, this can easily be corrected in the decoding scheme to produce even more accurate results than those we have reported. However, if the decoding scheme cannot account for the effects of correlations, the accuracy will be reduced.

Finally, the discrimination accuracy we have computed refers to an optimal decoding scheme. Downstream neurons may not be as efficient at interpreting the output of the coding neurons as our mathematical procedure (although see Salinas and Abbott, 1995). To examine this, we measured the discrimination accuracy when we required that the decoding scheme be based on a linear function of the firing rates (see Materials and Methods). This is what could be achieved, for example, by a neural network summing products of firing rates times synaptic weights. For linear decoding, we found that about $2.2(2^{0.35N})$ stimuli could be discriminated by N neu-

rons with 50% accuracy. While smaller than the result for optimal coding, this still represents a large number of stimuli for even modest numbers of neurons. Thus, downstream neurons can obtain large amounts of information about the represented stimuli even if they decode the responses of the coding neurons fairly inefficiently using a limited number of synaptic inputs.

Notes

We thank A. Treves and R. Baddeley for helpful comments and D. Foster for assistance with data preparation. This research was supported by the National Science Foundation Grant DMS9208206, the McDonnell-Pew Centre for Cognitive Neuroscience at Oxford (L.A.), the Oxford MRC Interdisciplinary Research Centre in Brain and Behavior, the Human Frontier Science Program, and Medical Research Council Grant PG851379 (E.R.).

Address correspondence to L. F. Abbott, Center for Complex Systems, Brandeis University, Waltham, MA 02254.

References

Baizer JS, Ungerleider LG, Desimone R (1991) Organization of visual inputs to the inferior temporal and posterior parietal cortex in macaques. J Neurosci 11:168–190.

Bialek W, Rieke F, de Ruyter van Steveninck RR, Warland, D (1991) Reading a neural code. Science 252:1854–1857.

Bruce C, Desimone R, Gross CG (1981) Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J Neurophysiol 46:369–384.

Desimone R (1991) Face-selective cells in the temporal cortex of monkeys. J Cognit Neurosci 3:1–8.

Desimone R, Gross CG (1979) Visual areas in the temporal lobe of the macaque. Brain Res 178:363–380.

Desimone R, Albright TD, Gross CG, Bruce C (1984) Stimulus-selective properties of inferior temporal neurons in the macaque. J Neurosci 4:2051–2062.

Eckhorn R, Popel B (1974) Rigorous and extended application of information theory to the afferent visual system of the cat. I. Basic concepts. Kybernetik 16:191–200.

Eckhorn R, Popel B (1975) Rigorous and extended application of information theory to the afferent visual system of the cat. II. Experimental results. Biol Cybern 17:7–17.

Gawne TJ, Richmond BJ (1993) How independent are the messages carried by adjacent inferior temporal cortical neurons? J Neurosci 13:2758–2771.

Gochin PM, Colombo M, Dorfman GA, Gerstein GL, Gross CG (1994) Neural ensemble coding in inferior temporal cortex. J Neurophysiol 71:2325–2337.

Gross CG, Desimone R, Albright TD, Schwartz EL (1985) Inferior temporal cortex and pattern recognition. Exp Brain Res Suppl 11: 179–201.

Hertz JA, Kjaer TW, Eskandar EN, Richmond BJ (1992) Measuring natural neural processing with artificial neural networks. Int J Neural Syst 3:91–103.

Kjaer TW, Hertz JA, Richmond BJ (1994) Decoding cortical neuronal spike signals: network models, information estimation and spatial tuning. J Comput Neurosci 1:109–139.

Macrae AW (1971) On calculating unbiased information measures. Psychol Bull 75:270–277.

Maunsell JHR, Newsome WT (1987) Visual processing in monkey extrastriate cortex. Annu Rev Neurosci 10:363–401.

Optican L, Richmond BJ (1987) Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. Information theoretic analysis. J Neurophysiol 57:132–146.

Optican LM, Gawne TJ, Richmond BJ, Joseph PJ (1991) Unbiased measures of transmitted information and channel capacity from multivariate neuronal data. Biol Cybern 65:305–310.

Perrett DI, Rolls ET, Caan W (1982) Visual neurons responsive to faces in the monkey temporal cortex. Exp Brain Res 47:329–342.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes. Cambridge: Cambridge UP.

Richmond BJ, Optican LM (1987) Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex I. Information transmission. J Neurophysiol 57:163–178.

Rolls ET (1984) Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. Hum Neurobiol 3:209–222.

Rolls ET (1991) Neural organisation of higher visual functions. Curr Opin Neurobiol 1:274–278.

Rolls ET, Tovee MJ (1995) The sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. J Neurophysiol 73:713–726.

Salinas E, Abbott LF (1994) Vector reconstruction from firing rates. J Comput Neurosci 1:89–107.

Salinas E, Abbott LF (1995) Transfer of coded information from sensory to motor networks. J Neurosci 15:6461–6474.

Seltzer B, Pandya DN (1978) Afferent cortical connections and architectonics of the superior temporal sulcus and surrounding cortex in the rhesus monkey. Brain Res 149:1–24.

Shadlen MN, Newsome WT (1994) Noise, neural codes and cortical organization. Curr Opin Neurobiol 4:569–579.

Tovee MJ, Rolls ET, Treves A, Bellis RP (1993) Information encoding and the responses of single neurons in the primate temporal visual cortex. J Neurophysiol 70:640–654.

Treves A, Panzeri S (1995) The upward bias in measures of information derived from limited data samples. Neural Comput 7:399–407.

Zohary E, Shadlen MN, Newsome WT (1994) Correlated neuronal discharge rate and its implications for psychophysical performance. Nature 370:140–143.