Neural Selectivity in Anterior Inferotemporal Cortex for Morphed Photographic Images During Behavioral Classification or Fixation

Yan Liu and Bharathi Jagadeesh
Department of Physiology and Biophysics, University of Washington, Seattle, Washington

Submitted 14 December 2007; accepted in final form 29 January 2008

Liu Y, Jagadeesh B. Neural selectivity in anterior inferotemporal cortex for morphed photographic images during behavioral classification or fixation. J Neurophysiol 100: 966–982, 2008. First published January 30, 2008; doi:10.1152/jn.01354.2007. Anterior inferotemporal cortex (aIT) contributes to the ability to discriminate and classify complex images. To determine whether and what proportion of single neurons in aIT cortex can yield enough information to classify complex images, we recorded from aIT neurons during the presentation of morphed photographic images in sessions in which monkeys classified images in a two alternative forced-choice—delayed-match-to-sample (2AFC-DMS) task or in sessions in which they performed a fixation task. The sample stimuli were chosen from a sequence in which one image was gradually morphed into another in a pair, while the original pair of images served as choices. Responses of many individual neurons in aIT cortex during the behavioral classification of the images, decoded using an ideal observer analysis, were sufficiently selective to account for the observed behavioral classification of the images. The responses of a separate population of neurons in aIT cortex recorded in subsequent sessions while the monkeys viewed the same images, were less selective than neural responses measured during sessions in which the 2AFC-DMS task was performed. Our findings show that many neurons in aIT could provide sensory information sufficient for the classification of images when a 2AFC-DMS task was performed.

INTRODUCTION

Differences in neural responses to different complex images in inferotemporal cortex (IT) are thought to underlie the remarkable capacity of primates to recognize, identify, discriminate, and remember complex visual images. Neural response difference (selectivity) is found in IT for a vast array of different stimuli (Allred et al. 2005; Baylis and Driver 2001; Desimone et al. 1984; Freedman et al. 2003, 2006; Keysers et al. 2001; Kiani et al. 2007; Leopold et al. 2006; Schwartz et al. 1983; Sigala and Logothetis 2002; Vogels 1999; Zoccolan et al. 2005), but the dimensions over which this selectivity is arrayed are poorly understood. In some circumstances, selectivity in IT to complex images can be decoded to deduce which images were presented (Hung et al. 2005; Kiani et al. 2007), implying that selectivity could be used to discriminate or classify these stimuli. In addition, stimuli that are behaviorally difficult to discriminate can produce smaller response differences in IT than those that are easier to discriminate (Allred et al. 2005, 2007; Baylis and Driver 2001; Keysers et al. 2000; Kovacs et al. 1995; Op De Beeck et al. 2001).

Measurements of neural response and behavioral performance show that difference in response caused by altering stimuli are globally similar to changes in discriminating or classifying the same stimuli (Allred and Jagadeesh 2007; Freedman et al. 2003; Keysers et al. 2000; Kovacs et al. 1995; Leopold et al. 2006; Vogels and Orban 1994). Detailed comparisons between the discrimination capacity of the responses of individual neurons to specific stimuli and behavior in IT are rare, however (Allred and Jagadeesh 2007; Vogels and Orban 1994). The first aim in this report is to make quantitative comparisons between the discrimination capacity of individual neurons to the performance of the animal (Britten et al. 1992) by examining whether neural responses of single neurons in IT are sufficiently selective to provide the information necessary for completing a classification task with photographic images along arbitrary physical dimensions. Our first questions are whether cells sufficiently selective for classification exist, and if they do, what proportion of the population they form, and how that relevant population of cells might be selected. Furthermore, IT is susceptible to the influences of attention and task demands (Desimone 1996; Moran and Desimone 1985) and learning (Freedman et al. 2006; Koida and Komatsu 2007; Mruczek and Sheinberg 2007; Sigala and Logothetis 2002), therefore the relationship between neural response selectivity and the discrimination capacity of animals might depend on the task being performed with the images. So in the second part of the report, we examined the discrimination capacity of a separate population of neurons, recorded in response to the same images, but in sessions in which the monkeys merely viewed the stimuli.

We compare the selectivity of neurons to the monkey’s ability to discriminate the same stimuli as the discrimination becomes more and more demanding. We manipulated the discriminability of stimuli by interpolating (morphing) between two individual photographic images (Freedman et al. 2001–2003). We measured the difficulty of classifying a sample stimulus in a two-alternative forced-choice–delayed-to-match (2AFC-DMS) task in which the monkeys were required to classify each photographic image as one of two alternatives, the two photographic images from which each sample image was morphed. We recorded the activity of aIT neurons in sessions while measuring behavior in the 2AFC-DMS task or in sessions while a fixation task was presented. We asked whether neural responses of individual cells in aIT were sufficiently selective to provide the basis for the measured behavioral classification of those stimuli by comparing performance of neurons in aIT, calculated using an ideal observer analysis, ROC analysis, to performance in the classification...
task. During the performance 2AFC-DMS task, the neural performance of many cells, matched in both threshold and sensitivity to simultaneously collected behavioral performance. Average neural thresholds for individual images, across a population of selective cells, were correlated with average behavioral thresholds for the same images.

The relationships between neural and behavioral responses were not found, however, when behavior in the 2AFC-DMS task sessions was compared with neural responses collected during a fixation task performed during other sessions. These data suggest that the response differences in small populations of neurons in aIT cortex to photographic images were sufficient for classification of those images when classification of the image between two alternatives was performed. During subsequent sessions in which a fixation task was performed, the discriminatory power of a separately recorded population of neurons was lower than for those collected during the performance of the 2AFC-DMS task.

**METHODS**

**Subjects and surgery**

Two male adult monkeys (*Macaca mulatta*) weighing 5–8 kg were used in these experiments. Standard techniques were used for recording from awake-behaving primates (Allred et al. 2005): surgeries on each animal were performed (under gas anesthesia, isoflurane) to implant a head restraint, a cylinder to allow neural recording, and a scleral search coil to monitor eye position (Judge et al. 1980). The cylinder was implanted using stereotaxic measurements to access inferotemporal cortex (described in the following text). All animal handling, care, and surgical procedures were performed in accordance with guidelines established by the National Institutes of Health and approved by the Institutional Animal Care and Use Committee (IACUC) at the University of Washington.

**Recording procedures**

Single-unit recordings were made using standard techniques. On each day, an x-y stage for positioning and an electrode holder containing a sterile guide tube and tungsten microelectrode (Alpha-Omega, Nazareth, Israel) were attached to the top of the recording cylinder. The guide tube was lowered to ~15 mm above the expected location of aIT, using stereotaxic coordinates, and the electrode was moved using a microdrive (David Kopf Instruments, Tujunga, CA) and after passing through a preamp, filter, and amplifier, signals from the electrode were sorted on-line using the Alpha-Omega spike sorter. Responses of single aIT neurons were collected while monkeys viewed images and performed the 2AFC-DMS task or a fixation task. Coded spikes were stored on a PC at a rate of 1,000 Hz using CORTEX, a program for neural data collection and analysis developed at the National Institutes of Health (Bethesda, MD). On-line histograms were created to qualitatively judge selectivity, but for the purposes of this paper, all data analysis was performed post hoc on stored data. Eye movements were monitored and recorded (500 Hz) using a scleral search coil system (Fuchs and Robinson 1966; Judge et al. 1980). Materials for these procedures were obtained from Crist Instruments (Hagerstown, MD) and DNI (Newark, DE).

Neurons were selected using anatomical and physiological criteria. Structural MRI was used to guide placement of the recording chambers, which were centered on the following stereotaxic coordinates: monkey L: 17 mm lateral, 19 mm anterior; monkey G: 16 mm lateral, 18 mm anterior over the right hemisphere. Neural recordings were targeted near the center of the chamber (monkey L: 17 L, 17.5 A; G, 16 L, 17.5 A), in between the perirhinal sulcus and the anterior middle temporal sulcus. The sites might include cells from both TE and perirhinal cortex. However, the selection criterion for recording locations was the presence of cells that responded selectively to 1 of the 12 image pairs used in this study, and recording locations were altered until such selectivity was found. The experimenter usually found an apparently selective neuron (1 included in the population presented in this report) after sampling one to three sites in a session. On isolating an apparently selective neuron, the monkey began performing the 2AFC-DMS task with the stimulus set for which the isolated neuron was selective. We sampled locations within the chamber, moving 0.5–1 mm when appropriate selectivity for the images was not found for 2–3 days. When selective cells were found, the area was resampled until we could no longer isolate cells with selectivity for 1 of 12 pairs of images. The total area spanned was ~4-mm diam circle centered on the locations described in the preceding text, and there were no systematic differences in the locations sampled at any point in the study. Anatomical data are unavailable for the precise locations of recording sites because animals are still being used in other related experiments.

**Behavioral paradigm and data collection**

2AFC-DMS. The monkey performed the 2AFC-DMS task (Fig. 1A) with two different sample stimuli [chosen from 1 of 12 pairs of photographic images and the 9 morphed variants of those two samples (Fig. 2), see following text for description of morphing]. The target choices always consisted of the pair of photographic images, “choice array,” and the monkey was required to classify each of the morph

![A: Fixation, Stimulus, Delay, Saccade](https://via.placeholder.com/150)

**FIG. 1.** Task and behavioral performance. A: 2 alternative forced-choice—delayed-match-to-sample (2AFC-DMS) task with 2 possible sample stimuli appearing on different trials. Samples could be 1 of the pair of photographic images (which also served as the choice array) or morph variants as shown in B. One example image pair is shown, but the images used could be any of 11 other pairs (Fig. 2). B: morph exemplars that could appear as sample stimuli.
FIG. 2. Twelve image pairs used as sample stimuli in 2AFC-DMS task. The images to the right (A and B) are the 2 original images, and 4 levels of morph are shown for each image. The correct choice for the top row of images was A; the correct choice for the bottom row of images B. The 5 images in each row correspond to morph levels 1–5 for that image. The image numbers to the left correspond to the image number designation for that pair of images.
variants as one of the pair by making a saccade to the matching stimulus in the choice array based on the similarity of the sample to one of the two choices.

An example image pair and associated trials are illustrated in Fig. 1A. Each trial began with the presentation of a fixation point (0.3°, red square). After the monkey acquired fixation (within a 4 or 5° diam window centered on the fixation spot), there was a variable delay before the onset of the sample, presented for 320 ms. After another variable delay period (400–700 ms), the choice array was presented. The two choice stimuli were located 5° to the left and either 5° up or down from the fixation point. Location of the two targets was randomized between the two positions, so the monkey could not determine the location of correct saccade before choice array onset. All images (both samples and choices) subtended 90 × 90 pixels, which was about 4° of visual angle at normal viewing distance. If the monkey’s gaze left the fixation window at any time before the onset of the choice array, the trial was discarded. After another variable delay period (mean: 567 ms, 412–1,212 ms), the fixation point turned off, providing the monkey’s cue that reward was available for making a saccade to the correct stimulus in the choice array. On trials where the monkey initiated a saccade before the cue, the saccade choice was recorded, then both target images turned off, and no reward was administered. When a saccade occurred after the cue, and a correct saccade was made, the monkey received a reward. Analysis of latency from choice array onset to saccade onset indicated that rather than using the cue to guide behavior on a trial-by-trial basis, both monkeys made their saccades at a stereotypical latency after the onset of the choice array, centered after the mean time of cue (mean cue of reward availability: 567 ms; mean latency: 882 ms); the monkeys did not learn the relevance of the cue on each trial and instead waited long enough to obtain reward on most (93%) of the correct trials. Monkeys were required to maintain fixation for 500 ms after arriving at their target for the trial to be counted as a valid choice. All trials in which the monkey made a selection from the choice array were included in this analysis; trials where the monkey aborted the trial are not included.

In a single session/block of trials, the monkey performed the 2AFC-DMS task with 11 possible sample images [1 pair of photographic images chosen from a set of 12 possible image pairs (Fig. 2, labeled A and B) and 9 morphed variants, described in the following text] presented randomly within a block of trials. On average, we obtained 17 ± 4.0 trials for each sample stimulus, ranging from a minimum of 5 trials to a maximum of 33 trials.

**FIXATION TASK (MORPH).** After recording sessions using the 2AFC-DMS task was completed, neural responses to the same set of stimuli (1 of 12 image pairs, and the morph variants of each) were examined in a fixation task. This task was identical to the 2AFC-DMS task except that the trial ended before the presentation of the choice stimuli, and the monkey was rewarded at that point for having maintained fixation throughout the trial.

**FIXATION TASK (SEARCH).** To isolate neurons for the experiments, monkeys performed a fixation task. This version of the fixation task was used only to search for cells, and insufficient data were collected to characterize the neural responses to the search images. The appropriate image pair was selected with this task before beginning the 2AFC-DMS task described in the preceding text. In this task, the monkey was rewarded for maintaining fixation (on the fixation point) while two successive, identical sample stimuli were presented for 300 ms, separated by a 300-ms interval. The stimuli were presented twice to assist in our ability to isolate and find cells (it is easier to hear the periodic burst of responses to successive presentations of an image), but this design was used only to search for cells and not during any data presented in the report. Stimulus sets for the passive fixation task consisted of the 12 pairs of photographs (Fig. 2, labeled A and B). Using this task, the experimenter chose 1 of the 12 pairs of photographic images (Fig. 2) to use in the 2AFC-DMS task and the morph fixation task described in the preceding text. The particular pair of images was chosen because qualitative assessment of the neurons suggested that the recorded neuron produced stronger responses to one image in the pair than the other.

**ANIMAL TRAINING.** The animals were trained using operant conditioning techniques with water as the reward for desired behavior. Each monkey was trained first to complete the 2AFC-DMS task with one sample image, which remained consistent throughout a block of trials. When the monkey was consistently picking that sample image from the choice array, the alternative sample image (the other image in the original pair) was presented as a sample image while the monkey learned to choose the matching image from the choice array. After the monkey performed well in this “reversal” training, we decreased the number of trials in a blocks in which one sample images was presented. Once blocks were sufficiently short (consisting of <10 trials), we randomized the sample presentations so either sample stimulus could appear any trial. When the monkey performed well with two sample stimuli, we introduced the morphed sample variants, first including the eight additional sample variants for which there was a correct response and finally introducing one additional sample variant that putatively consisted of equal parts of the pair of photographic images and for which choices made for that image were rewarded randomly. Monkeys were trained extensively with the 12 sets image pairs and their morphed variants before recording sessions began. Thus both the original images and the morphed variants were familiar to the monkeys before the neuronal data were collected (for both the 2AFC-DMS task and the morph fixation task).

**MORPHED IMAGES.** Each of the 12 pairs of images was morphed using MorphX (http://www.norrkross.com/software/morphx/MorphX.php), a freeware, open-source program for morphing between two photographic images. Details of the algorithm are available from the source code for the MorphX. A brief description of the morphing process is described here: first, the experimenter chose two images (image A and image B, Fig. 2); she then set control points on the two images to designate corresponding areas. MorphX then created a series of image frames warping the surface of image A to the control points on image B simultaneously altering image colors in image A to correspond to image colors in image B. Nine intermediate images in sequence between the two original images were created. All of the image pairs and the eight morph variants for which a correct answer was imposed are shown in Fig. 2. The process of morphing was complicated and depended on the setting of individual control points in the two images. For example, two identical images could be morphed using different control points; different control points would result in a different set of morph variants. We used only one set of morph variants (Fig. 2), created with one set of control points for each pair of photographic images. Multiple features in the image, including shape, texture, and color, change as the images are morphed and change between the different morphed variants; some features might change more or less rapidly than others. Different image pairs may have key features that change at different rates among the morph variants.

The morphing algorithm in MorphX was complicated and cannot be presumed to be linear (steps between individual frames were not necessarily identical). Nevertheless, we assigned a level to each morph variant, corresponding to the order of morph between the two original stimuli. There are 11 possible sample images for each pair (2 original + the 9 intermediate morph images). One image is putatively located at the midpoint between the two images in the choice pair; this image is assigned morph level (or strength) 0, designating that it was considered to be of equal similarity to the two original images from which it was morphed. In addition to this image, there were four additional morphed images for each of the original sample images. These four variants for each original image, correspond to morph levels 1–4; level 1 was deemed most distant from the original, level 4 the most similar. Finally, the two images in the choice pair (from which the variants were morphed) correspond to morph level 5.
REWARD CONTINGENCIES FOR MORPHED IMAGES. The morphed variants, along with the individual images in the choice pair were used as samples in the 2AFC-DMS task described in the preceding text. In every case, the monkey’s task was to classify the sample image as one image in the choice pair (the images from which the samples were morphed), by judging the similarity between the image presented during the sample period and the two available choices (the images from which the samples were morphed). The monkey was rewarded for a correct classification when the morphed sample was located closer in the morph sequence to the image. The monkey was rewarded randomly for the one morph variant that was putatively at the midpoint of the sequence. The data collected during the presentation of this ambiguous image are not included in this report because no behavioral response is objectively correct (performance for the ambiguous, randomly rewarded stimulus was 0.503, not significantly different from 0.50, $P > 0.40$). Across the sessions, for this ambiguous stimulus, the monkeys were also equally likely to choose either of the two images in the choice pair (proportion chose A, 49%, not significantly different from 50%, sign test, $P = 0.28$), showing that there was no overall bias in favor of the two images in each pair of images.

EFF AND INEFF STIMULI. In characterizing the neural and behavioral performance, morph level was presumed to correspond to the strength of the stimulus in the “preferred” (Eff) or “null” (Ineff) direction (assignment of the 2 images Eff and Ineff is described in the following text). Unlike conventional characteristics of stimulus strength [orientation (Vogels and Orban 1994), correlated noise (Britten et al. 1992; Uka and DeAngelis 2003)], the morph dimension was not quantitatively characterized in the physical stimulus space. Behavioral responses with these morphs, however, changed systematically as a function of morphing level (Fig. 1B), and behavioral responses can be well fit with Weibull functions (see following text) using the arbitrary morph “level” unit as the stimulus dimension, suggesting the feasibility of this approach in analyzing this data.

DATA FITTING: BEHAVIORAL DATA ANALYSIS: PSYCHOMETRIC FUNCTIONS. To obtain a psychometric function from behavioral data, we fit the proportion correct as a function of morph level using the Weibull function (Eq. 1). Fitting the data decreases the noise associated with measurements with individual stimuli and allows more reliable comparisons between data sets

$$y = \gamma + (\lambda - \gamma) \times (1 - e^{-\left(\frac{x}{\lambda}\right)^\alpha})$$

where $y$ represents the proportion correct, gamma and lambda represent the floor and ceiling parameters, alpha represents the location parameter, and beta represents a value correlated with the slope. We fit the data using Eq. 1 using the maximum likelihood method described by Wichmann and Hill (2001) (psignifit toolbox, version 2.5.6, http://bootstrap-software.org/psignifit) to determine the best-fit parameters, constraining lambda to the range [0.9–1]. The minimum performance value is defined as 0.5 (gamma), and the ceiling performance is represented by lambda. Allowing lambda to range from 0.9 to 1, results in different fit values for a small set of data where ceiling performance was poor. In general, in the neural data (which were fit using the same equation, method described in the following text), these are cells that were unselective for the original stimuli. Therefore allowing a broader or narrower constraint for lambda (0.6–1 or 1–1) did not significantly change any of the main conclusions of the study. The steepness of the curve is reflected by the slope parameter (beta, larger values indicate steeper slopes), and the relative horizontal location of the curve is reflected by the location parameter (alpha).

Threshold for each fitted curve was defined as the morph level where performance reaches 0.75, defined in units of morph level. Threshold is reached at a lower morph level if the psychometric function is steep; therefore low-threshold values represent good performance, whereas higher-threshold values represent poor performance. The sensitivity of the monkey to changes in morph level was reflected in the measured slope at a performance level of 0.75, defined in units of performance increase/morph level. Similar calculations were made for neural performance as described in the following text.

Goodness of fit was evaluated by comparing the residuals between the data and the fit to Monte Carlo simulation of the expected distributions, and looking for “overdispersion” of the errors compared with that expected by chance (Wichman and Hill 2001).

PHYSIOLOGICAL DATA ANALYSIS. Neural responses to each of the samples was calculated by summing the spikes that occurred during the presentation of the sample stimulus, for each sample stimulus. The epoch used to calculate the sum was the sample duration, offset by a neural response latency of 75 ms (75–375 ms). Using these responses, the Eff stimulus was defined as the stimulus of the 2 that made up a pair) that evoked the bigger responses in that epoch (Eff > Ineff). Eff and Ineff correspond, respectively, to a preferred and null stimulus for the recorded cell. Either image in a pair could correspond to the Eff stimulus, depending on the response of the cell under study. Across the population of experiments, stimulus A was the preferred (Eff) stimulus in 69 experiments, and stimulus B was the preferred (Eff) stimulus in 61 experiments. The number of cells that preferred image A or B for each individual pair was also similar.

NEURAL PERFORMANCE. Neural performance was calculated using an ideal observer analysis, the ROC statistic. We defined the Eff and Ineff image based on the response to the original images in the 12 pairs; we paired each morph level (morph variant) of the Eff stimulus with the corresponding level of the Ineff stimulus (morph level 5, Eff compared with morph level 5, Ineff, morph level 4, Eff, compared with morph level 4, Ineff, and so on). We calculated the total number of spikes in the sample epoch (75–375 ms after stimulus onset) for each trial where an Eff image was presented and for each trial where the paired Ineff image (with the same putative difference from the original Ineff choice image) was presented. The two resulting spike distributions were compared with one another using the aROC statistic (area under the ROC curve). An ROC value was thus obtained for each morph level, corresponding to the neural performance at that morph level (Britten et al. 1992; Green and Swets 1966) using data from the presentation of both Eff and Ineff stimuli; this value was plotted as a function of morph level as was the behavioral performance. Implementing the aROC analysis on this data as described in the preceding text requires several theoretical assumptions about the data (Shadlen et al. 1996). The aROC statistic is a convenient summary of the selectivity of the neurons and is primarily used for that purpose; however, the assumptions are considered further in the discussion.

For each experiment, the neural performance was fit using the Weibull function shown in Eq. 1; the same function also used to fit behavioral data. Neural threshold was defined as the morph level at which the neuron reached a performance value of 75%, and other values were calculated as described in the fits to the behavioral data.

Neural and behavioral performance were compared by asking whether the best-fitting curve to the pooled data for neuron, and behavior was distinguishable from the best fitting individual curves for the behavioral and neural using the method and Matlab toolbox described by Wichmann and Hill (pfcmp) (Wichmann and Hill 2001).

Because the use of the aROC statistic calculating neural performance is not well established in aIT, we repeated the calculations of the aROC statistic with the threshold set to 0.75. This analysis confirmed that the use of the aROC statistic was valid, and that the aROC statistic could be used to compare neural and behavioral performance.

The sensitivity of the monkey to changes in morph level was reflected in the measured slope at a performance level of 0.75, defined in units of performance increase/morph level. Similar calculations were made for neural performance as described in the following text.

Goodness of fit was evaluated by comparing the residuals between the data and the fit to Monte Carlo simulation of the expected distributions, and looking for “overdispersion” of the errors compared with that expected by chance (Wichman and Hill 2001).
response was greater (Eff) or less than (Ineff) the criterion response level. The main conclusions of the paper are valid with this alternative calculation of neural performance.

Neural performance calculated on incorrect trials was not significantly different from neural performance on correct trials for any morph level \((P > 0.10)\), all morph levels, paired \(t\)-test. The power of this test depends on the number of correct and incorrect trials; at high morph levels the number of incorrect trials was low (mean = 2.2 trials, morph level = 5). At low morph levels, the number of incorrect trials was higher (mean incorrect trials = 12.1 trials, morph level = 1). Even at low morph levels the power from the small \(n\) is low. Therefore an inability to demonstrate a difference between neural performance on correct and incorrect trials could stem from a lack of sufficient power in the analysis. But because we were unable to demonstrate a difference on correct and incorrect trials, we pooled responses on correct and incorrect trials to make performance calculations.

**NEURONAL DATABASE. 2AFC-DMS task.** Selection of cells for inclusion in the population could influence the conclusions made in this report. We therefore detail the selection of cells for the neuronal database in this section. We attempted 163 recording sessions in which the monkey performed the DMS task. In 123 (75%) of them, we were able to successfully isolate cells that appeared to be selective for 1 of the 12 image pairs (using qualitative criterion). One or two sites were sampled before including a cell in the population. Thus the cells described in this report were common (found after sampling 1 or 2 sites, in 75% of attempted sessions). One hundred fifty-four experiments were collected in the 123 sessions (some sessions contain >1 cell; some cells were recorded with >1 image set). Cells were pruned from the data set for two behavioral criteria: performance <60% for all three of the easiest morph levels (7 sessions) or less than five trials of data collected for each sample stimulus (6 sessions); these experiments were excluded, resulting in 141 data sets. We then selected visually responsive cells (response during the sample epoch greater than the baseline response, \(P < 0.05\)) for further analysis. Eleven cells were excluded because of this criterion, resulting in a population of 130 experiments.

Finally, neuronal threshold or behavioral thresholds were in calculable (fits did not converge, or thresholds were much less than one (the lowest morph level tested) or much more than five (the highest morph level tested) for a subgroup of experiments. These experiments (7) are excluded from calculations and comparisons where parameters of the Weibull fit or thresholds are discussed.

**Morph fixation task.** We attempted 86 recording sessions in which the monkey performed the fixation task with the morph stimuli. In 63 (73%) of them, we were able to successfully isolate cells that appeared to be selective for 1 of the 12 image pairs. These 63 sessions resulted in 104 experiments (multiple cells were recorded in some sessions, and some cells tested with more than one stimulus set). Experiments were further pruned from the database based on the number of trials of data collected for each cell, resulting in 102 data sets. We then selected visually responsive cells for further analysis, resulting in a population of 84 experiments, which were further characterized.

**DIFFERENCES BETWEEN MORPH FIXATION TASK AND DMS TASK.** Neural responses in the morph fixation task were collected after the collection of data for the DMS task, and data were collected from a separate population of neurons. This experimental design raises the possibility that differences (other than the performance of the task) exist between the two sets of data. Because the experiment design allows for these differences in design producing differences seen in the two data sets collected in the two conditions, we list the possibilities here: recording methodology was as identical as possible: the same equipment was used; the same criteria were used for selecting cells, by the same experimenter. Because the fixation task data were obtained after the DMS task data, we cannot rule out the possibility that the animals’ cortex was changed during the interval between the two sets of recordings. However, these animals continue to participate in experiments, and during performance of related classification tasks, selectivity for photographic images can still be found with the same frequency as detailed in the 2AFC-DMS task here. No differences in eye movements could be detected between the two tasks; the duration and location of fixations were similar and frequency of saccades was similar. In both tasks, the 90% of fixation locations were located within a 1° window centered on the fixation point. The recording locations were similar in both tasks, and no systematic differences were detectable. In both animals, the same area of cortex was targeted and the recording depths were similar. No systematic differences in the minor changes in track locations and depths could be detected between the data collected in the two different sets of sessions.

Differences in the stimulus experience in the two tasks might lead to different adaptation effects and differences in the sequence of the data collection for the two tasks might lead to long-term adaptation effects. Both of these possibilities are addressed by looking for within and across session trends in the fixation and DMS task data.

**DIFFERENCE INDEX (NEURONAL ANALYSIS).** For both the morph fixation and the 2AFC-DMS tasks, some analyses are described for subsets of the original population, which were chosen based on a difference index (Britten et al. 1992). The difference index is defined in Eq. 2.

\[
d_i = 1 - \frac{(I/E)}{\text{where } I = \text{Ineff-baseline and } E = \text{Eff-baseline}}
\]

A cell with a much smaller response to the Ineff stimulus than the Eff (relative to baseline) will produce an \(d_i\) close to 1. A cell with nearly equal responses to the Ineff and Eff stimulus will produce a \(d_i\) close to 0. The index can be less than zero or greater than one, depending on interactions between the response to the sample stimulus and during the baseline period. The baseline response was calculated over all trials of sample presentation using the 300-ms epoch before the onset of the sample stimulus (−300 to 0 ms).

**RESULTS**

**Psychophysical performance with morphed stimuli**

Behavioral performance in the 2AFC-DMS task (Fig. 1A) with morphed stimuli was surprisingly predictable. The morphed images themselves change along many different dimensions; as one stimulus was morphed into the other, multiple features change at different rates. Nevertheless after training in the task, monkeys’ performance changed systematically as a function of the morph level, corresponding to the ordinal placement of the images between the two extremes from which they were morphed (Fig. 1B). Behavior with eight morphed samples and the original photographs is shown in Fig. 3: four samples were similar to one of the pair of original images while another four were similar to the other image as shown in Fig. 2. The monkey was rewarded for making the choice of the similar image for these eight images. Figure 3 illustrates the performance in the task for each monkey’s behavior, pooled across all sessions as a function of morph level (samples of morph level 5 are identical to the choice stimuli; morph level 1 is most dissimilar). The monkeys performed well when one of the two original photographs was presented as the sample stimulus, morph level 5, \(>90\%\) correct. With increasing dissimilarity between the morphed sample stimulus and the choice stimuli, the monkeys produced fewer correct choices (morph level 2). Even with the most difficult samples (morph level 1, see METHODS) the monkeys produced performance above
characterized (see METHODS), but behavior as a function of morph performance/monkey A chance (Fig. 3, 0.65, $\chi^2$ test, different from chance = 0.5, $P \ll 0.001$). The morph dimension cannot be quantitatively characterized (see METHODS), but behavior as a function of morph level could be. The performance as a function of morph level can be fit with a Weibull function, using the morph level as the stimulus parameter. The Weibull function in Eq. 1 fit the data well, and the measured threshold (the morph level at which 75% performance is reached) is 1.54 for monkey A and 1.64 for monkey B, a hypothetical stimulus that would be placed between the first and second morph level; this low-threshold demonstrates the strong performance in the task. The quality of the Weibull function fit, assessed by looking for “overdispersion” of the data points compared with the fit values (see METHODS, psignifit toolbox, $P > 0.05$) (Wichmann and Hill 2001), could be rejected for behavior in only 10 of 130 experiments (7.6% of sessions). The data were well fit by the Weibull function; therefore the threshold and slope parameters were a convenient summary of the performance and could be used to compare behavioral and neural performance.

Neural selectivity for the images in the choice pairs (during 2AFC-DMS Task)

Selectivity for the 12 pairs of photographs (Fig. 2) was found frequently in inferotemporal cortex. We attempted 163 recording sessions in this experiment. In 123 (75%) of these sessions, after sampling one to three sites, 130 visually responsive cells were found in sessions where the monkey worked well and long enough to assess behavior with the stimuli. We selected image pairs for further study in the 2AFC-DMS task from a limited set of 12 image pairs (image pairs, Fig. 2) based on qualitative assessment of different responses to one of the pair of 12 images, using the search fixation task to identify the appropriate stimulus; the assignment of the two images in the pair as Eff (preferred) and Ineff (null) was made post hoc based on the mean response during the sample epoch (75–375 ms after stimulus onset). To quantify the selectivity for the Eff and Ineff stimulus in a pair, we calculated a difference index (Eq. 1) (Britten et al. 1992). The distribution of difference indices is shown for the 130 experiments (Fig. 4, 0.69 ± 0.43). Some cells included in the population were mildly selective for the images used in the session (■, $di < 0.5$), whereas many others were strongly selective for the image pair used in the session (□, $di > 0.5$). Cells were equally likely to respond better to either of the two images in the pairs of images shown in Fig. 2; for 69 experiments the image A was the Eff, or preferred image, whereas for 61 experiments image B was the Eff, or preferred image.

Neural selectivity for morph variants of sample image (during 2AFC-DMS task)

Neural responses during the 2AFC-DMS task to morphed sample variants was also surprisingly predictable. Two examples of selective cells ($di = 0.75$ and $di = 1.31$) collected in two different experiments while the monkey performed the 2AFC-DMS task are shown in Fig. 5, $A–D$. Figure 5A shows the response of a single cell to the pictures of the bird and fish (Fig. 5A, panel 5) and the responses to interpolated, or morphed versions of these individual stimuli (Fig. 5A, panels 1–4). The cell responded better to the presentation of the picture of the bird (Eff, red line) than to the presentation of the picture of the fish (Ineff, blue line) (Fig. 5A, panel 5). As the Eff image was morphed toward the Ineff image, the mean responses to the two sets of images converged; the mean response to the Eff image (and variants) decreased and the response to the Ineff image (and variants) increased as the stimuli were morphed to the middle image (Fig. 5A, panels 1–5). Change in response as a function of morph level is summarized by the mean spike count in the 75- to 375-ms time window after the onset of the sample stimulus (Fig. 5B). In this cell, the mean response in the sample epoch to the Eff sample changed roughly linearly as a function of the morph level (Fig. 5B). Figure 5C shows the response of another example cell; this cell was selective for the pair of stimuli containing the bird and the tree. The cell responded better to the response of the image of the bird (Eff, red, Fig. 5C) than to the picture of the tree (Ineff, blue, Fig. 5C). Again, as the sample resembled more and more the midpoint between the

FIG. 3. Behavioral performance for 2 monkeys averaged across all sessions for which neural data were collected. The abscissa shows the level of the morph stimulus, in arbitrary morph units; the morph level reflects deviation from the original photographs in the choice array to which the sample must be compared; samples of morph level 5 are identical to the target stimuli; other stimuli are ranked by the degree of similarity to those originals. The data were fit with a psychometric function (Eq. 1). The threshold across all sessions is 1.64 for monkey A (- - -) and 1.54 for monkey B (---, ◦); this threshold corresponds to a morph stimulus in between the 1st and 2nd levels where the monkeys attains a performance level of 75%, marked by arrows. The stimuli corresponded to a morph stimulus in between the 1st and 2nd levels where the performance as a function of morph level could be. The performance as a function of morph level can be fit with a Weibull function, using the morph level as the stimulus parameter. The Weibull function in Eq. 1 fit the data well, and the measured threshold (the morph level at which 75% performance is reached) is 1.54 for monkey A and 1.64 for monkey B, a hypothetical stimulus that would be placed between the first and second morph level; this low-threshold demonstrates the strong performance in the task. The quality of the Weibull function fit, assessed by looking for “overdispersion” of the data points compared with the fit values (see METHODS, psignifit toolbox, $P > 0.05$) (Wichmann and Hill 2001), could be rejected for behavior in only 10 of 130 experiments (7.6% of sessions). The data were well fit by the Weibull function; therefore the threshold and slope parameters were a convenient summary of the performance and could be used to compare behavioral and neural performance.

Neural selectivity for the images in the choice pairs (during 2AFC-DMS Task)

Selectivity for the 12 pairs of photographs (Fig. 2) was found frequently in inferotemporal cortex. We attempted 163 recording sessions in this experiment. In 123 (75%) of these sessions, after sampling one to three sites, 130 visually responsive cells were found in sessions where the monkey worked well and long enough to assess behavior with the stimuli. We selected image pairs for further study in the 2AFC-DMS task from a limited set of 12 image pairs (image pairs, Fig. 2) based on qualitative assessment of different responses to one of the pair of 12 images, using the search fixation task to identify the appropriate stimulus; the assignment of the two images in the pair as Eff (preferred) and Ineff (null) was made post hoc based on the mean response during the sample epoch (75–375 ms after stimulus onset). To quantify the selectivity for the Eff and Ineff stimulus in a pair, we calculated a difference index (Eq. 1) (Britten et al. 1992). The distribution of difference indices is shown for the 130 experiments (Fig. 4, 0.69 ± 0.43). Some cells included in the population were mildly selective for the images used in the session (■, $di < 0.5$), whereas many others were strongly selective for the image pair used in the session (□, $di > 0.5$). Cells were equally likely to respond better to either of the two images in the pairs of images shown in Fig. 2; for 69 experiments the image A was the Eff, or preferred image, whereas for 61 experiments image B was the Eff, or preferred image.

Neural selectivity for morph variants of sample image (during 2AFC-DMS task)

Neural responses during the 2AFC-DMS task to morphed sample variants was also surprisingly predictable. Two examples of selective cells ($di = 0.75$ and $di = 1.31$) collected in two different experiments while the monkey performed the 2AFC-DMS task are shown in Fig. 5, $A–D$. Figure 5A shows the response of a single cell to the pictures of the bird and fish (Fig. 5A, panel 5) and the responses to interpolated, or morphed versions of these individual stimuli (Fig. 5A, panels 1–4). The cell responded better to the presentation of the picture of the bird (Eff, red line) than to the presentation of the picture of the fish (Ineff, blue line) (Fig. 5A, panel 5). As the Eff image was morphed toward the Ineff image, the mean responses to the two sets of images converged; the mean response to the Eff image (and variants) decreased and the response to the Ineff image (and variants) increased as the stimuli were morphed to the middle image (Fig. 5A, panels 1–5). Change in response as a function of morph level is summarized by the mean spike count in the 75- to 375-ms time window after the onset of the sample stimulus (Fig. 5B). In this cell, the mean response in the sample epoch to the Eff sample changed roughly linearly as a function of the morph level (Fig. 5B). Figure 5C shows the response of another example cell; this cell was selective for the pair of stimuli containing the bird and the tree. The cell responded better to the response of the image of the bird (Eff, red, Fig. 5C) than to the picture of the tree (Ineff, blue, Fig. 5C). Again, as the sample resembled more and more the midpoint between the
two extremes (decreasing morph level), the responses of the neuron converged to an intermediate value (Fig. 5, panel 1).

Comparison of neural and behavioral performance

The goal of collecting these data was to compare the behavioral performance of the monkey to the information available in the neurons for the classification task. This goal requires translating or decoding the neural responses into neural performance and finding neural thresholds. We used the aROC statistic to translate neural response into neural performance (see METHODS for additional details) (Allred and Jagadeesh 2007; Britten et al. 1992; Green and Swets 1966). The biological plausibility of this decoding algorithm for responses to arbitrary pairs of visual images is debatable. Nevertheless, the aROC statistic, applied to these neural data, provides a convenient characterization of the performance of individual neurons, which can be compared with simultaneously collected behavior.

For the two example cells in Fig. 5, neural performance matched behavioral performance collected at the same time. Neural performance (aROC statistic, neurometric curves) is shown for the two example cells from Fig. 5 [Fig. 6A (blue line and points)] and can be compared with simultaneously collected behavior (red line and points, example cells left top and left bottom). The best-fitting curves for neural and behavioral performance for these neurons are also shown (Weibull function, Eq. 1). The neural performance matched the behavioral performance for these example neurons as reflected in the similarity of the threshold and slope values derived from the fits (neuron/behavior threshold: 7A, top left: 2.18/2.33 6A, bottom left: 2.53/2.53; slope: 6A, top left: 0.26/0.28 6A, bottom left: 0.13/0.10); furthermore, the best-fitting curves for the behavioral and neural data are statistically indistinguishable from the common best fitting curve ($P > 0.50$) (Wichmann and Hill 2001). Many neurometric functions calculated from neural responses had thresholds and slopes similar to the simultaneously collected psychometric (behavioral) curves. Similar patterns of responses for behavior and neural performance are shown for four more example neurons (Fig. 6A, middle and right). Again the curves describing the behavioral and neural performance in each example were statistically indistinguishable (Wichmann and Hill 2001), and thresholds and slopes are similar.

For many cells in aIT, neural and behavioral performance was similar (Fig. 6, B–D). Figure 6, B–D, shows the comparisons between the Weibull fits for neural and behavioral performance for different parameters in the Weibull fit. Figure 6B shows the ratio of neural to behavioral thresholds. Cells with neural thresholds similar to behavioral thresholds have ratios near one. Across the population of cells, the ratios were greater than one, indicating that on average, neural thresholds were higher than behavioral thresholds (Fig. 6B, all bars, $P < 0.0001$, sign test). However, when only the population of selective cells (Fig. 4, di > 0.5, gray bars) was considered, the ratios were not significantly different from one (Fig. 6B, di > 0.5, gray bars, n: 79, 61% of the population, $P = 0.1147$, sign test). Cells strongly selective for the choice pair, from which sample images were morphed, reflected the behavioral capacity of the animal with these images. Interestingly, the ratio of neural to behavioral thresholds across the whole population of cells in aIT appears bimodal: there were two peaks, one near one, and the other near 10 (meaning a 10-fold higher threshold for neural responses than for behavior).
Two other parameters for the Weibull fits to neural and behavioral performance also matched one another for selective cells. Slopes for the fits to behavioral and neural performance were also similar (Fig. 6C). Across the population of cells, the slope ratios were centered near one (though, significantly different from 1, P < 0.0172); sub-selecting the selective cells (gray bars) brought the mean slope ratios closer to one (not significantly different from 1, P = 0.074). The alpha parameter of the Weibull fit (which reflects the shift in the function along the x axis) is shown for the population in Fig. 6D. Across the population of cells, the alpha ratios were greater than one (P = 0.0001, sign test); when selective cells were selected, the ratios were distributed around one (gray bars, di > 0.5, P = 0.8221, sign test). These data show that many cells in the population produced neural performance similar to behavior in the 2AFC-DMS task.

Behavioral and neural thresholds in individual sessions are compared directly in Fig. 7A, in which neural and behavioral thresholds are plotted versus each other. Neural thresholds were higher than behavioral thresholds across the entire population (P < 0.0001, sign test) and especially across the subpopulation of less selective cells (di < 0.5, shown in blue, Fig. 7A, P < 0.0001, sign test) but not across the subpopulation of selective cells (di > 0.5, shown in red, Fig. 7A, P = 0.092, sign test). Neurons exhibited higher neural thresholds than behavioral thresholds when the cells recorded in the session were weakly selective, but similar neural and behavioral thresholds when the cells were selective for the choice pair. Furthermore, for many cells, neural and behavioral performance was indistinguishable (P > 0.05, solid red circles, compare best-fitting common curve to best fitting individual curves) (Wichman and Hill 2001): of the population of 130 cells, 70 (54%) neurons produced fits indistinguishable from the behavior. Of the subpopulation of 79 cells with di > 0.5, 51 (64%) produced fits statistically indistinguishable from the behavior.
in those same sessions. Across the whole population of cells, the mean of neural thresholds was higher than behavioral thresholds (which hovered around 1.5 or between the lowest 2 morph levels, \( P = 0.0001 \), sign test). Average neural threshold was higher than behavioral thresholds until cells with \( di > 0.4 \) were selected \( [ P = 0.2242, \text{sign test}, n \text{ cells} = 82 \text{ (53%)}] \). Average neural and behavioral thresholds remained indistinguishable until the very most selective cells \((n = 28, 18\% \text{ of population, } di > 1.0 \text{ sign test, } n = 28, P = 0.0357)\) are selected. For this small population, the neural thresholds are lower than behavioral thresholds, indicating that the neurons performed better than the monkey. The proportion of cells included by selecting different levels of selectivity using the difference index is shown in Fig. 7C. The average threshold of cells consisting of \(~50\%\) of the most selective cells (based on responses to the choice pair) was comparable to behavior with the same stimuli, during performance of a 2AFC-DMS task. The average threshold of cells consisting of \(~20\%\) of the most selective cells reflected better neural than behavioral performance.

The session by session correlation between neural and behavioral thresholds was positive for the subset of selective cells \( (di > 0.5) \) and becomes progressively larger as more selective cells are chosen (Fig. 7D). The correlation between neuron and behavior across sessions became significant when the cells with \( di > 0.8 \) \((n = 42/130, 32\% \text{ of the population})\) cells were singled out \( (r = 0.27, P = 0.045, \text{Fig. 7D})\) and remained selective as more selective cells are included (peaking at 0.53, \( n = 12 \)). The significant, and increasing, correlation suggests that session by session correlation between performance and neural response was present, for the most selective cells in the population, but that correlation was weakened by including the less selective cells in the population.

The average neural performance (as opposed to thresholds) at each morph level also matched the behavior for the most selective cells (Fig. 7E). In these three panels, the mean behavior and mean neural performance for the different morphed sample variants (morph levels 1–5) are plotted against each other for populations of progressively more selective cells. When all the cells in the population were included, behavior was better than the performance of the neurons (Fig. 7E, \( di > 0, n = 129, \text{left, } P < 0.0001, \text{all morph levels} \)). When a selective population of neurons was chosen, however, both the neural and behavioral performances were comparable to each other for each of the stimuli, the original, and the morphed variants of the sample stimuli (Fig. 7E, right, \( di > 0.8, n = 51, P = 0.10, \text{all morph levels} \)).

The average neural thresholds for the two monkeys were not significantly different \( (\text{monkey } G: \text{mean threshold } 3.86; \text{monkey } L: \text{mean threshold } 2.73, \text{not significantly different, rank sum, } P = 0.1553) \).

**Psychometric and neurometric variability across different stimuli**

Is the neural selectivity for individual pairs of stimuli related to the differing ability of the animal to discriminate different visual stimuli? Some of the variability in behavioral performance across different sessions (Fig. 7A) resulted from the different stimulus pairs used during different sessions. Some images were easier to discriminate than others, and showed both steeper slopes and lower thresholds. For instance, the

---

**FIG. 7.** Comparison of the neuronal and behavioral performance. **A:** behavioral threshold on each session plotted against neural threshold on each session. Solid circles indicate experiments in which neuronal and psychophysical thresholds were statistically indistinguishable. Open circles designate experiments where the 2 values were significantly different. Red circles indicate data drawn from the population of selective cells \( (di > 0.5) \). Blue circles indicate data drawn from the population of less selective cells \( (di < 0.5) \). The diagonal line indicates the line of equivalent psychophysical and behavioral thresholds. **B:** geometric mean of neural threshold as a function of cell selectivity \( (di; (lue squares)) \) and geometric mean behavioral threshold for the same subgroup of sessions (red circles). Error bars are SE. **C:** proportion of cells greater than a critical di value (the proportion of cells shown here corresponds to the equivalent di criterion in B). **D:** correlation (Pearson R) between behavioral and neural thresholds as a function of selectivity; open circles \( P > 0.05 \); closed circles \( P < 0.05 \). **E:** mean behavioral performance vs. mean neural performance for different subgroups of cells: difference index left to right panels \( (di < 0, di > 0.5, di > 0.8) \).
example cells shown in Figs. 5A and 6A, the image pair shown in Fig. 6A (left, top) produced a steeper psychometric function than the one shown in Fig. 6A (left, bottom). In these example neurons, the steeper slope shown in Fig. 6 (left, top vs. bottom) for behavior is reflected in the steeper slope for the neural response (slopes at 0.75 performance, neuron/behavior: top, 0.26/0.28; bottom, 0.13/0.10). Is the superior behavioral performance (lower threshold) in the task for a given stimulus pair reflected in the neural performance for the same stimulus pair? To address this question, we calculated the geometric mean of the threshold across the populations of selective cells (di > 0.8) for each stimulus pair. For this population of selective cells, the mean behavioral threshold was correlated with the mean neural threshold for the same stimuli (Fig. 8A, n = 51, r = 0.76, P = 0.01). The variability in behavioral performance with different stimuli was reflected in the variability of neural performance with the same stimuli. In addition, the neural and behavioral thresholds across the stimulus pairs were statistically indistinguishable (P = 0.07). The correlation between neural and behavioral thresholds over different images depended on choosing experiments in which cells selective for the original images were recorded (di > 0.8). The behavioral and neural thresholds averaged across images were only weakly correlated when all cells were included (di > 0.0, r = 0.397, P = 0.227) or when selective cells were included (di > 0.5, r = 0.20, P = 0.55; Fig. 8B).

**Neural performance during fixation task**

Neural performance of selective cells measured when the monkey performs the 2AFC-DMS task was similar to the behavioral performance of the monkey with the same sets of images (Figs. 5–9). Are neural responses similarly selective in sessions where no specific task was performed? If, for example, the agreement between neural and behavioral performance resulted from the selection of stimuli and cells and the calculation applied, neural performance should be similarly high for all cells similar to the behavioral performance of the monkey (Fig. 10A, black triangles, fixation task, black squares, 2AFC-DMS task). Selective cells collected during the 2AFC-DMS task had thresholds similar to the behavioral performance of the monkey (Fig. 10A, ■, di > 0.4, similar to behavior, P = 0.2242), but even the most selective cells during the fixation task had higher thresholds than the mean threshold in the 2AFC-DMS task (Fig. 10A, ▲, P < 0.0001, compared with mean behavioral threshold in the 2AFC-DMS task = 1.5, - - -). In addition, the mean thresholds for individual stimuli of selective cells in the fixation task (di > 0.8) were not comparable to the mean threshold of the behavior with those same stimuli (measured separately; Fig. 10B, behavioral threshold for each stimulus pair in 2AFC-DMS task vs. neural threshold in fixation task, P < 0.0001). Behavioral performance in the task at all morph levels was also higher than the performance of even the most selective cells collected during sessions in which the fixation task was performed (Fig. 10C). For all populations of cells (Fig. 10C, left, middle, P < 0.0001, all morph levels) and for even the most selective cells, the behavior is better than the neural performance (Fig. 10C, di > 0.8, right, all morph levels,
In addition, the mean performance at each morph level for any subgroup of selective cells was lower during fixation task sessions than for neurons collected during the 2AFC-DMS task sessions (compare neural performance in fixation task sessions than for neurons collected during the level for any subgroup of selective cells was lower during

sessions, \( n \) sessions, by selecting a much smaller subset of neurons from each other (\( \text{P} < 0.001 \)). In addition, the aROG selectivity at the highest morph level is equated for the two populations of cells collected in the two types of sessions, by selecting a much smaller subset of neurons from the fixation task sessions (\( n = 15, 18\% \) of cells, fixation task sessions, \( n = 79, 61\% \) of cells DMS task sessions) neural performance at morph levels 2–4 are significantly different from each other (\( \text{P} < 0.05 \)). This pattern of responses suggests that even the most selective cells collected during the fixation task showed more rapid degradation of response when the images were morphed than the selective cells collected during the classification task.

The average neural thresholds (across the sample population for the 2 monkeys in the fixation task) were not significantly different (\( \text{monkey G: mean threshold 6.51, monkey L: mean threshold 6.70, not significantly different, rank sum, \( \text{P} = 0.9736 \)).

Neural responses in the fixation task were collected after the data collected in the 2AFC-DMS task, and thus monkeys were more familiar with the stimuli during the fixation task than during the 2AFC-DMS task. Neural responses in IT can decrease with repeated presentations of stimuli result in lower neural responses to those same stimuli, in the majority of neurons (Li et al. 1993; Sawamura et al. 2006). Could thresholds be higher in the fixation task because the responses in the fixation task represent more familiar, and thus more highly adapted responses to the choice images and their morphed variants? If adaptation was the reason for the difference in neural response between the fixation tasks and the behavioral task, we should see evidence of adaptation’s effect on neural responses during the months of data collection for the 2AFC-DMS tasks. To test this prediction, we divided the data from each monkey into two groups, an early set of sessions (the first 1/3 of the 2AFC-DMS data) and a late set of sessions (the last 1/3 of the 2AFC-DMS data). The comparison between these two sets of sessions, in distribution of selectivity, thresholds, and neural performance is shown in Fig. 11. The proportions of selective cells in the first months of recording and the last months of recording are similar (Wilcoxon rank sum test, \( \text{P} = 0.7100, \text{Fig. 11A} \)). Furthermore, the mean thresholds early and late in the session are similar (Fig. 11B, \( \text{P} > 0.10 \), except for 2 values at \( \text{di} > 1.0 \) and \( \text{di} > 1.1 \)). Interestingly, this small population of cells showed a significant difference in thresholds (\( n = 10, \text{P} < 0.05 \)) in early versus late sessions. This difference suggests the possibility that the thresholds are lower when the monkey is most familiar with the task (the opposite direction from that expected from adaptation). These data undermine a significant role for long-term adaptation in the threshold differences between the two sets of sessions (fixation versus 2AFC-DMS task) and raises the possibility that the thresholds of the most selective cells were lower after more extensive familiarization with the stimuli during

\[
\begin{align*}
\text{FIG. 10. Comparison of the neuronal thresholds in} & \text{ fixation task and 2AFC-DMS task. A: geometric mean of neural threshold as a function of cell selectivity for fixation task (\( \cdot \)), mean of neural thresholds as a function of cell selectivity for 2AFC-DMS task (\( \cdot \), as in Fig. 7B) and mean behavioral threshold across all sessions (---). Error bars are standard errors of the mean. B: correlation between geometric mean behavioral (measured in DMS task) and geometric mean neural thresholds for different images (measured in fixation task). C: mean behavioral performance (in 2AFC-DMS task, as in Fig. 7E) vs. mean neural performance in fixation task for different subgroups of cells: difference index left to right panels (\( \text{di} > 0, \text{di} > 0.5, \text{di} > 0.8 \)).}
\end{align*}
\]

\[
\begin{align*}
\text{FIG. 11. Comparisons of neural responses during early sessions and later} & \text{ sessions during performance of 2AFC-DMS task. A: proportion of cells greater than a criterion di value. \( \cdot \), early sessions; ---, late sessions. B: mean thresholds as a function of criterion di value. \( \odot \), early sessions; \( \triangle \), late sessions. Values at \( \text{di} > 1.0 \) and \( \text{di} > 1.1 \) are statistically different (\( \text{P} < 0.05 \)), all others are statistically similar (\( \text{P} > 0.05 \)). C: neural performance in early sessions vs. late sessions for different subpopulations of cells chosen based on their selectivity. Right to left: di > 0.0, di > 0.5, di > 0.8.}
\end{align*}
\]
the performance of the task. Finally, the neural performance (as opposed to threshold) in initial and later sessions was statistically indistinguishable, for all subgroups of cells (Fig. 11C, \( P > 0.10 \), all morph levels and cell subgroups). Thus long-term adaptation is unlikely to explain the differences in neural performance in the cell populations in the fixation versus DMS task sessions.

In addition to the confound of the order of the sessions, the timing and stimulus presentation in the fixation and 2AFC-DMS tasks were different. The fixation task proceeded more rapidly (because of the absence of the choice portion of the task) and the DMS task had more presentations of the original images (which appeared as both samples and choices). The difference in stimulus experience might have influenced adaptation that occurs over the course of the recording session. To test this possibility, we compared the neural performance calculated on the first half of each session to neural performance calculated on the second half of each session for both the DMS task and the fixation task. There were no detectable differences between the neural performance in the first or second half of the task for either the DMS task or the fixation task (neural performance in two halves of session, \( P > 0.10 \) for all morph levels and for both tasks, and for all selectivity levels, Fig. 12, A and B).

The aROC statistic obscures the mean spike responses measured in response to the sample stimulus in both tasks. The response levels can be seen in Fig. 13, which shows the mean responses as a function of morph level across populations of selective cells for the DMS task sessions (Fig. 13A) and for the fixation task sessions (Fig. 13B). Neural responses across the population were greater for Eff images (□) compared to Ineff images (○) collected in the DMS task sessions compared with fixation task sessions (\( P < 0.05 \) all morph levels except morph level = 1, \( P < 0.05 \)). Responses to Ineff images (○) were not significantly different in the two sets of sessions. We also examined the variability in neural response in the two different tasks, for identical stimuli, by comparing the slope of the variance versus mean relationship for each stimulus in the recorded neural populations in the two different sessions. The variance is roughly equal to the mean for both populations in both the fixation session and the DMS sessions (variance = 1.4 \times\) mean for fixation data, variance = 1.3 \times\) mean for DMS task data), comparable to that seen in other studies (Erickson et al. 2000).

**Discussion**

In this study, the decoded performance of many individual neurons matched behavioral performance in a 2AFC-DMS task in sensitivity and threshold when the monkey performed a 2AFC-DMS task with the images. In addition, thresholds for performance of small populations of neurons selective for individual images were correlated with threshold for behavioral performance with those same images, showing that the selectivity in aIT for different visual images reflected the animals’ ability to discriminate these images. These characteristics were not found in a separate population of neurons recorded in subsequent sessions in which the monkey performed a fixation task. In the following text, we discuss the relationship between this study and others in which behavior and neural responses to difficult to discriminate stimuli have been measured and compared, the plausibility of the calculations underlying the aROC statistic in reading out neural responses into performance, and the potential interpretations of differences between the data collected during sessions in which the 2AFC-DMS task was performed and during sessions in which the fixation task was performed.

**Relationship to other quantitative comparisons between neural and behavioral performance**

The selective cells the thresholds of which match behavioral thresholds found in this report were not a rare: cells consisting of >50% of the sampled population match the behavior of the monkey (Fig. 7A); this is a significant finding of the study. It is not unexpected that the occasional neuron might match behav-

![FIG. 12. Comparison of neural performance between the 1st half of a recording session and the 2nd half of a recording session. A: during 2AFC-DMS task. B: during fixation task.](http://jn.physiology.org/)}
ioral performance. But in this study, cells were often found that matched the performance of the monkey in classifying stimuli across an arbitrary stimulus dimensions even though these cells were found while searching with a small and incomplete stimulus set (12 possible image pairs). These cells, could, further, be identified by their degree of selectivity for the choice stimuli. It is, however, the case, that selectivity for these stimuli was less prevalent than, for example, direction selectivity in MT (Britten et al. 1992), where almost no cells have a $d_i$ less than one. One potential explanation of this difference is the inadequacy of the stimuli used to search for selective stimuli. In MT, all directions can be systematically sampled. It is impossible, however, to sample all potential stimuli in IT. Therefore some cells entered the population even if they were not particularly selective for any of the 12 image pairs tested.

In another version of the 2AFC-DMS task, manipulating the duration of stimulus presentation, the average neural responses of a population of highly selective cells are also comparable to behavioral performance in a 2AFC-DMS task (Allred and Jagadeesh 2007). When sample duration is manipulated, however, even the relatively less selective neurons in aIT in the sample perform as well as the monkey at short stimulus durations, while at long stimulus durations, those same less-selective neurons perform less well than the monkey. Finally, when monkeys perform a classification task with color patches, many cells in IT are highly selective for color patches (Koida and Komatsu 2007).

In several other studies, in contrast, only a few rarely found cells can match the performance of the monkey (Freedman et al. 2003; Muhammad et al. 2006; Vogels and Orban 1994). Differences across studies in the frequency of cells with particular properties could be attributed to differences in selection of cells for inclusion in the study. We note that the cells included in this report were selected using an inclusive criterion. Cells to include in the sample were found frequently in recording sessions after sampling a few sites where neurons could be isolated (see METHODS). Therefore in our hands, the cells described in this study were not rare or difficult to find. However, it is possible that the cells are rare within IT cortex itself, but that our methods of identification consistently found relevant clusters of cells (Erickson et al. 2000).

Another potential difference between this study and others is that the monkeys were well trained in a 2AFC-DMS task with a limited set of photographic stimuli before the recording session began. Because the monkeys were very well trained with these pairs of images, learning may have played a role in the likelihood of finding selectivity for these stimuli and the sensitivity to the morphed images (Erickson et al. 2000; Freedman et al. 2006; Peissig and Sheinberg 2007; Sigala et al. 2002). Finally, the task used in this report, a 2AFC-DMS task, posed a limited question that must be assessed about each image: the monkey was required to classify each image as one of two a closely related variant in contrast to more complex categorization tasks used in some other studies (Freedman et al. 2003).

Plausibility of aROC statistic for translating neural responses into neural performance

The sufficiency of the neural responses in aIT for performing the 2AFC-DMS task depends on the plausibility of the calculation used to read out the neural responses into performance. The aROC statistic was used to calculate neural performance, which was then compared with the monkey’s behavior. The

**FIG. 13.** Response to sample stimulus in 2AFC-DMS or fixation task. Mean response across cells as a function of morph level. Error bars are SE across cells Eff images (○); Ineff images (□). A: 2AFC-DMS task. B: fixation task.
biological plausibility of implementing the ROC calculation depends on a number of assumptions about the availability of information in neurons other than the one the neural performance of which is being calculated (Shadlen et al. 1996). Morphing is a complicated method of changing the image that alters the image along many different dimensions, and the biological plausibility of the aROC statistic would not be predicted on theoretical grounds with these kinds of stimuli. Nevertheless, across the population of selective cells, the aROC calculation produces neural performance quite similar to behavioral performance (Fig. 6) when the monkey performs the 2AFC-DMS task. To use aROC calculation as more than a computational trick, however, aIT neurons would have to meet several assumptions: first, one must assume that for each recorded neuron, there existed an “anti-neuron” elsewhere in aIT cortex with identical but opposite selectivity. In this experiment, that means a neuron that responds better to stimulus A in a pair than to stimulus B has a corresponding anti-neuron somewhere in the brain that responds better to B than to A; second, the anti-neuron must respond to Ineff sample in the same way that the recorded neuron responded to the Eff sample and vice versa. Third, on each trial, the observer must be able to select and compare the responses of populations of neurons and anti-neurons. Fourth, the responses of the relevant populations of neurons and anti-neurons must be approximated by the distribution of responses to repeated presentations of the Eff and Ineff stimuli. The performance value calculated from the data, with these assumptions, represents the proportion of trials that the ideal observer would correctly identify the effective sample by comparing responses from the neuron and anti-neuron spike count distributions; given these assumptions, the calculation could plausibly be implemented in the brain. These conditions seem implausible in a naïve context (horses and giraffes are not “opposites” in any conventional image space), but experience-dependent modification of neural responses in IT (Peissig and Sheinberg 2007) might produce them, a question for further study.

Differences in selectivity during the 2AFC-DMS task sessions versus during fixation task sessions

A separate population of neurons recorded while presenting the same stimuli in sessions in which the monkey performed a fixation task were less selective than those found in sessions while the monkey performed a DMS task. Cells were less likely to be selective for the choice pair during a fixation task, and even when they were selective for the choice pair, the neural were much lower than behavioral thresholds (measured separately during the DMS task). This finding is compatible with some others studies that show influences of task performance and learning on the responses of IT neurons. In Leopold et al. (2006), more neurons appear to be selective for different face stimuli when a monkey performs a classification task: more cells respond to all of the face stimuli in an untrained monkey than in one who has been trained to perform classification. In Koida and Komatsu (2006), responses of populations of neural responses are more selective for the classification dimension when monkeys perform a classification task with colored patches. Response can occur earlier in a recognition task compared with a fixation task (Sary et al. 2006). Training with stimuli seems to increase the selective response to trained stimuli in ERPs/VEPs (Peissig and Sheinberg 2007), presenting the possibility that these cells might be easier to find in some conditions. In addition, perirhinal cortex, which may overlap with the region in which these responses were recorded, are influenced by the task being performed (Lehky and Tanaka 2007).

The neural data collected during the fixation data must be interpreted with some caution, however, because performance of the DMS versus fixation task is not the only difference between the two data sets. The population of cells was different and was collected in a set of sessions following the set of sessions in which the monkeys performed the task. We chose to examine responses in separate sessions because cueing an animal to not classify a stimulus might be similar to asking a person to not think about an elephant; mere cueing might be insufficient to prevent perceptual classification of the images during viewing (Koida and Komatsu 2006) and the processing of the image during fixation subsequently uncontrolled (Gilbert and Sigman 2007). The design imposes limits on the interpretation of the data, however. Because the population of cells was different in the two tasks and because the data in the fixation task was collected after the data in the 2AFC-DMS task, several important confounds exist. First, although we tried our best to apply the same criterion in selecting cells for inclusion into the population in the two tasks, biases in the inclusion of cells could contribute to the generally lower selectivity seen during the fixation task. In addition, because the data in the fixation task were collected after the performance of the 2AFC-DMS task, stimuli were more familiar to the monkey during the fixation task. Familiarity is known to produce adaptation in neural responses in IT (Li et al. 1993; Sawamura et al. 2006). Adaptation resulting from long term familiarity is unlikely, however, to explain the difference between fixation data and the DMS data. The stimuli were already familiar before the collection of the fixation data and neural responses showed no evidence that performance decreased (threshold increased) over the course of the 6 mo of data collection (Fig. 11). The two tasks also result in different rates of stimulus presentation and differences in the intratrial experience with the stimuli: in the fixation task, the sample stimuli are presented with shorter intertrial intervals because the choice portion of the task is removed, while in the DMS task, the two choice stimuli are presented more frequently because they are presented both as choices and as samples. This difference raises the possibility that short-term adaptation could play out differently in the two sets of data. We were unable, however, to detect any trends over the course of the session in either data set, and the differences between the two data sets remained when only the first half of the data sets were compared with behavior.

Finding no evidence for these identified confounds, furthermore, does not rule out other unidentified differences between the two data sets. This worry cannot be summarized dismissed because other examinations of the effect of task on neural responses have shown similar selectivity in data collected during fixation and during performance of another task (Hung et al. 2005; Suzuki et al. 2006). Finally, even when attributing the difference between the two data sets to the difference in the task performed, these data do not inform us about the changes in the responses of individual neurons between the two tasks because they did not follow a single neuron in the two condi-
tions (Koida and Komatsu 2006). The experimental design allows for the possibility that two completely different populations of neurons were being identified and recorded from during the two different tasks. Furthermore, the design of this task does not rule out the possibility that the increase in neural response selectivity could be general: a generalized increased gain of neural responses that enhances selectivity to all stimuli in IT cortex and possibly even in other cortices (Serences and Boynton 2007).

However, even while being careful not to interpret the fixation data too broadly, the data collected during the fixation task can be used to argue against a trivial explanation of the match seen between the neural and behavioral thresholds for selective cells: the possibility that the match is a computational artifact that results from the selectivity for the two images in the choice pair and the fitting algorithms used to calculate threshold. Using the same computational techniques to assess threshold (aROC statistic, fitting the data), on average, selective cells recorded during fixation did not match the measured threshold of the monkey.

**Stimulus variations and neural and behavioral thresholds**

The discriminability of stimuli (measured by threshold of neural performance) by the aIT neural population that were selective for that stimuli, and recording during performance of the DMS task was correlated with performance in the DMS task. This correlation could stem from physical stimulus differences among the different morph variants of different images. Better performance with a particular set of morphed image variants might have reflected larger physical differences between different morph variants, compared with those in another stimulus pair. Thus both the behavior and neural responses might reflect the different changes in the physical characteristics of the image. The connection between responses differences in IT and the discriminability of physical differences in stimuli, though assumed, is not quantitatively described: how much neural response difference corresponds to a particular difference in behavior? The relationship between the mean selectivity in aIT for particular morph variants and behavior (Fig. 8) suggests that the underlying selectivity of aIT cortex for these images might support the behavioral capacity with the images.

**Sufficiency versus necessity of aIT neural responses**

These data suggest that the neural responses of selective cells in aIT were sufficient to provide a basis for classifying the images, while the task is performed and assuming that the assumptions for the read out are met. Interpreting the meaning of this match is more difficult than interpreting similar matches seen between the discrimination capacity of individual neurons and behavior in studies where the stimulus is well defined and controlled (Britten et al. 1992; Parker and Newsome 1998; Uka and DeAngelis 2004). In those studies, only one parameter is meaningful for performance of the task (direction of motion, disparity) and that parameter is well defined and controlled. In this data, morphing produces a multi-dimensional change in the stimulus, any of which the monkey could use to drive his performance. The variability in behavior could diminish any relationship seen between the neural responses and the behavior. In addition, because multiple features of an image change through the morphing process, the monkey could rely on a feature change in the image that is correlated with the feature actually controlling the response of the neuron. Therefore data do not show that the neural responses are necessary or even used even while they are sufficient. Furthermore, if no match is found between behavior and neural performance capacity, the monkey may be using a feature that is uncorrelated with the feature controlling the response of the neuron.

The data also cannot show that the same characteristics of the image control both the neural responses and behavior. The images used in this study differ across lots of different features and physical dimensions; the stimulus differences may be represented widely throughout the visual system. The monkey might have relied on characteristics of the image that are widely represented throughout the brain; many other neurons might also approximate the performance of the monkey. The monkey could rely on features for which the neuron is selective on some trials but not others, resulting in variability in the relationship between the neuron and the behavior over different time scales. In spite of these possibilities, these data show that many neurons in aIT provided sufficient information to calculate performance of the monkey in a classification task with morphed photographic images and that the relevant neurons could be found by selecting neurons that are selective for the choice pair. This data provides support for the idea that aIT neurons provide sufficient evidence for the classification of these stimuli, although the data do not address the possibility that many other neurons also provide sufficient evidence nor do they show that aIT neurons were necessary for completing the task.

A causal role for this population of selective cells would be supported by microstimulation and lesion studies: if those methods were able to selectively target the relevant population of cells, enhancing their selectivity, or eliminating it should improve or impair performance. In a categorical face/object discrimination task, microstimulation in IT can bias monkeys’ choice of face stimuli (Afraz et al. 2006). In that study, microstimulation is able to target a population of face selective cells, biasing behavior in favor faces over other stimuli. The efficacy of microstimulation depends on the selectivity of the microstimulation site for the stimuli being classified (faces vs. objects) (Afraz et al. 2006). In addition, lesions can provide support for a causal role. Lesions of aIT and close by perirhinal cortex concur with an important role for this cortex in classification and identification tasks with altered versions of learned exemplars (Buckley and Gaffan 1998; Murray and Bussey 1999). The data, in summary, are compatible with a view of IT cortex in which the selectivity of neurons is arrayed across the dimensions that are perceptually important (Koida and Komatsu 2007).

**Acknowledgments**

We thank J. Skiver Thompson and K. M. Ahl for technical help and M. N. Shadlen, N. Brunet, and A. Akrami for comments on an earlier version of this manuscript.

**Grants**

This research was supported by the Sloan Foundation, the McKnight Foundation, the Whitehall Foundation, and the National Institutes of Health National Center for Research Resources.
REFERENCES


