Supporting Information

Zhang et al. 10.1073/pnas.1100999108

SI Text

Correlation Coefficient Classifier. A correlation coefficient classifier (1) was used for all analyses in this paper. The classifier works by learning a "classification vector" for each class k, which is simply the mean of all of the training examples from class k. To produce a class label prediction for a test point (which can be used to calculate the zero-one loss classification accuracy described below), the correlation coefficient was calculated between the test point and all of the classification vectors, and the class with the highest correlation was given as the class label. To create classification scores that were used when calculating the area under the receiver operating characteristic (ROC) classification measure, the correlation coefficient values between a test point and the classification vector for the class of interest were returned.

Measuring Decoding Accuracy. There are several different ways of measuring the accuracy of a decoding procedure. These methods include the zero-one loss function (2), normalized-rank (3, 4), ROC curve measures (5), and mutual information measures (6). In this paper, we use the zero-one loss measure, which is the most widely used decoding accuracy measure in neuroscience (1, 7, 8), and we also use an ROC curve-based measure which allows us to compare the decoding accuracy of the attended and nonattended stimuli in an unbiased way. In practice, we found all these measures to give very similar results (Fig. S7). Below we describe the two measures used in the paper in more detail.

Zero-one loss measure. The zero-one loss measure (which we refer to as the "classification accuracy") is one of the simplest decoding measures. This measure calculates the percentage of test examples that were correctly classified by the classifier. Because of its simplicity and wide use as a decoding accuracy measure, it is our preferred measure and we use it to report the results in Figs. 1*C* and 3. One shortcoming of this measure, however, is that it is not able to handle the case when multiple classes are present (as is the case when three objects are shown in our cluttered displays). The reason that this measure is problematic when multiple objects are correct is that conventional multiclass classifiers only return one predicted class for each test point. Thus, if three labels are correct, the classifier will only choose one of them, and on average one would expect that the classification accuracy for each object would be approximately one-third the classification accuracy when only a single object is present (if the classifier is choosing randomly between the three objects that are present). To deal with the case when multiple correct answers are possible (i.e., Fig. 1*B*), we thus used a multiclass area under the ROC-based measure that is described below.

Area under the ROC curve measure. ROC curves graph the proportion of positive examples correctly classified (true positive rate) as a function of the proportion of the negative examples incorrectly classified (the false positive rate) (5). To create such a function, a classifier must be used that returns a classification score that measures how likely a point is to belong to a particular class (rather than just returning a predicted label). To create an ROC curve for binary classification tasks, we take the classification scores for all "negative" class test points and sort them in descending order. We then measure the proportion of positive test-point scores that are greater than each successively smaller negative test-point score, which gives us an ROC curve, and the area under this curve (AUROC) gives us one overall number of classification accuracy. This AUROC measure is invariant to the ratio of the number of positive to negative test examples (5), which is an important property because we have different proportions of positive and negative test examples in the isolated-object decoding analyses versus the clutter decoding analyses [the measure used by Li et al. (9) did not have this property, and thus their results could have been biased because they had a different proportion of positive and negative test examples in their isolated versus cluttered decoding analyses]. To estimate the AUROC curves in the multiclass setting, we use the "class reference formulation" (5), which estimates the area under the ROC curve separately for each class k, with the positive examples being all test points that belong to class k and the negative test examples being all other points, and the final result being the average of all of the AUROC values from all classes. Because this method calculates AUROC separately for each class, it is possible to obtain a classification score for when multiple classes are present that is comparable to when only one class is present (e.g., a perfect classification score on one class does not preclude a perfect classification score on a second class in the case when multiple classes are present at the same time).

For the analyses in our paper, the classification scores for class k were the correlation coefficients between each test point and the mean vector of the training examples from class k. To apply this AUROC measure to our data for the isolated-object decoding analysis, we calculated the AUROC in step (*v*) of our decoding procedure (*Methods*). Because the test data were independent in each cross-validation split of the data, we pooled all of the classification scores from all test splits together to create one AUROC value for each bootstrap run, which is a method that is commonly used to obtain an average AUROC value over many splits of the data (5) (this pooling method had the benefit of giving 16 positive test examples and 176 negative examples when constructing the ROC curve, which presumably led to a more stable estimate than calculating the ROC area separately on each split of the data where there would only be 1 positive example and 15 negative examples from which to create a curve). When calculating the AUROC for the cluttered data, there was only one test set (i.e., no cross-validation was used), so no pooling was done. When calculating the AUROC for the attended object for class k, data from trials when class k was the attended object were used as positive scores and data from trials when object k was not shown were used as negative scores (data from trials when objects of class k were the nonattended object were not used in the analysis). Likewise, when calculating the AUROC for the nonattended objects, data from trials when the object k was the nonattended object were used to create the positive scores for the positive class and data from trials when object k was not shown were used to create the negative scores. We decided to also plot Fig. 2 using the AUROC measure rather than the zero-one loss measure to allow an easier comparison with Fig. 1*B*, although unlike Fig. 1*B* it would be possible to use the zero-one loss measure for this figure without introducing a bias.

Assessing Statistical Significance of the Decoding Results. Permutation tests were used to assess the significance of the results. To assess whether the decoding accuracy for the isolated objects was higher than the decoding accuracy for the three-object displays in Fig. 1*B*, we ran the decoding analyses simultaneously for the isolated-object results and the cluttered results (i.e., we trained a classifier on isolated-object data, and then tested both the isolated-object data and the cluttered data using the same classifier), and we saved all of the classification scores. We then calculated the real isolated-object AUROC value as described above and also created a null distribution that assumed there was no difference between the isolated-object results and the cluttered-display results. To create a single point from this null distribution, the same number of target present (absent) classification scores from isolated-object trials was randomly selected from the combination of target present (absent) classification scores from isolated-object and cluttered-display trials. The random selection of points was done separately for each class, and the same randomly selected points were used for all 50 bootstrap runs, thus emulating one full decoding analysis with randomly shuffled labels. The results were then averaged over all classes and all bootstrap runs to create one point in the null distribution that was equivalent to the real isolated-object decoding accuracy. This procedure was repeated 1,000 times to get a full null distribution, and the *P* value was found by assessing how many of the points in the null distribution were greater than the real isolated-object AUROC accuracy. The results showed that by 75 ms after stimulus array onset (±75 ms to account for the fact that a 150-ms bin was used), none of the points in the null distribution were greater than the real isolated-object AUROC accuracy (thus yielding a *P* value of less than 0.001), and this remained the case into the time period when the target object changed color. It is interesting to note that when the procedure was done using a null distribution that consisted of the isolated-object and attended-object classification scores (but not including the nonattended-object classification scores) the *P* value was greater than 0.01 at 475 ms after the cue onset and became greater than 0.10 at 525 ms after the cue onset, indicating that the isolated-object accuracy was becoming indistinguishable from the attended-object accuracy around those times.

To assess whether the decoding accuracies were higher for the attended stimulus compared with the nonattended stimuli (Fig. 1*B*), we ran a procedure in which we randomly shuffled the labels that indicated which stimulus was attended and which stimuli were nonattended when calculating the classification accuracy for the attended and nonattended stimuli. The procedure entailed randomly shuffling the attended and nonattended labels, and then running 10 bootstrap iterations of the cross-validation procedure using these shuffled labels to generate one sample from a null distribution for the attended stimuli and one sample from the null distribution for the nonattended stimuli (only 10 bootstrap iterations were used to save computation time). This procedure was then repeated 500 times to generate null distributions for the attended and nonattended stimuli which represented what would be the expected classification accuracy that would occur if there was no difference between the classification accuracies between the attended and nonattended stimuli. Time points were then found in which the real classification accuracy for the attended stimulus was higher than the classification accuracy for the attended stimulus's null distribution at a level of $P < 0.01$ (i.e., time points where five or fewer points from the attended null distribution were higher than the real attended stimuli decoding accuracy). Likewise, time points were found where the nonattended decoding accuracies were lower than the null distribution for the nonattended stimuli at $P < 0.01$. The results from this procedure showed that at 150 ms after the onset of the cue, both the attended decoding accuracies and the nonattended decoding accuracies dropped below the $P < 0.01$ level (and by the next time point, none of the values in the null distributions exceeded the values for the actual decoding accuracies until the end of the experiment, i.e., $P < 0.002$).

To assess whether information about the position of the attended stimulus in clutter was above chance (Fig. 1*C*), we ran a permutation test in which we randomly shuffled the attended location labels and ran the full clutter position decoding experiment. This procedure was done 200 times to obtain a null distribution, and the actual attended location position decoding accuracy was compared with this null distribution. Starting at 200 ± 75 ms after cue onset, none of the values in the null distribution were greater than the actual position decoding values (i.e., the *P* value was less than 0.005), and this lasted into the time period when stimuli began to undergo color changes.

Noise Correlation Analyses. Before analyzing whether noise correlations impact decoding performance, we first examined the level of noise correlations as a function of signal correlation for the 623 pairs of neurons in our dataset that had been recorded simultaneously, using the standard methods that have been used in previous studies (10, 11). Theoretical work has illustrated that if noise correlations are stronger for neurons that have higher signal correlations, then decoding accuracies could be lower in the presence of noise correlations (12). Similar to previous findings (10, 11), noise correlations in our dataset appear to increase with signal correlations [and were similar or perhaps a little stronger than previously reported noise correlations; these previous studies showed an increase in noise correlations of ∼0.10 (or possibly a little less) going from the least signal-correlated neurons to the most signal-correlated neurons, which is the range we tried to match when we added noise correlations to our data]. It should be noted, however, that the rise of noise correlations with signal correlations seen in our data could be due to correlated noise creating "fake" signal correlations (combined with limited sampling). Indeed, we noticed that if we calculated noise and signal correlations during the fixation period before the stimuli were shown, we saw the same trend of increasing noise correlations with increasing signal correlations *even when we used the labels for the upcoming stimuli that had not been shown yet* to calculate the signal correlations (Fig. S6*A Left*). Thus, the signal and noise correlation relationship in our data could very well be an artifact. Given that examining noise correlations was not the focus of this paper, we decided not to pursue this issue further.

To examine whether noise correlations had a large impact on decoding accuracy, we added noise correlations to our pseudopopulation vectors using two different methods. The first method, which we call "multiplicative uniform noise," involved multiplying our pseudopopulation vectors by random variables drawn from a uniform distribution over the interval [1 10]. The second method, which we call "additive Gaussian noise," involved generating random vectors using a multivariate Gaussian distribution that had zero mean and a covariance matrix that was based on neurons' signal correlations, and these randomly generated vectors were scaled by 0.001 and added to the pseudopopulation vectors (the uniform distribution interval of [1 10] and the scaling factor of 0.001 were chosen so that they would give rise to noise correlations that were of similar magnitude to those seen in the real data). As can be seen in Fig. S6*B*, both methods created noise correlations that increased with increasing signal correlations, and these noise correlations had a similar range of values as those seen in our data and in previous studies (10, 11). Our decoding procedure was then applied to pseudopopulations that had noise correlations induced by either multiplicative uniform noise (Fig. S6*C Left*) or by additive Gaussian noise (Fig. S6*C Right*) using a correlation coefficient classifier [new noise correlations were added each time new pseudopopulations were created; i.e., noise correlations were added before z-score normalization in step (*iii*) of our decoding procedure]. As can be seen, using populations with noise correlations led to only a slight decrease in the decoding accuracy, which suggests that classifier performance can be robust to at least some forms of noise correlations that are commonly seen in data.

Additional supplementary web material that contains related results can be found online (14).

1. Meyers EM, Freedman DJ, Kreiman G, Miller EK, Poggio T (2008) Dynamic population coding of category information in inferior temporal and prefrontal cortex. *J Neurophysiol* 100:1407–1419.
2. Duda R, Hart P, Stork D (2001) *Pattern Classification* (Wiley, New York).
3. Mitchell TM, et al. (2004) Learning to decode cognitive states from brain images. *Mach Learn* 57:145–175.
4. Meyers EM, Kreiman G (2011) Tutorial on Pattern Classification in Cell Recording. *Understanding Visual Population Codes*, eds Kriegeskorte N, Kreiman G (MIT Press, Cambridge, MA).
5. Fawcett T (2004) ROC graphs: Notes and practical considerations for researchers. *Mach Learn* 31:1–38.
6. Quian Quiroga R, Panzeri S (2009) Extracting information from neuronal populations: Information theory and decoding approaches. *Nat Rev Neurosci* 10:173–185.
7. Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005) Fast readout of object identity from macaque inferior temporal cortex. *Science* 310:863–866.
8. Quian Quiroga R, Snyder LH, Batista AP, Cui H, Andersen RA (2006) Movement intention is better predicted than attention in the posterior parietal cortex. *J Neurosci* 26: 3615–3620.
9. Li N, Cox DD, Zoccolan D, DiCarlo JJ (2009) What response properties do individual neurons need to underlie position and clutter "invariant" object recognition? *J Neurophysiol* 102:360–376.
10. Smith MA, Kohn A (2008) Spatial and temporal scales of neuronal correlation in primary visual cortex. *J Neurosci* 28:12591–12603.
11. Cohen MR, Maunsell JHR (2009) Attention improves performance primarily by reducing interneuronal correlations. *Nat Neurosci* 12:1594–1600.
12. Averbeck BB, Lee D (2006) Effects of noise correlations on information encoding and decoding. *J Neurophysiol* 95:3633–3644.
13. Crowe DA, Averbeck BB, Chafee MV (2010) Rapid sequences of population activity patterns dynamically encode task-critical spatial information in parietal cortex. *J Neurosci* 30:11640–11653.
14. Zhang Y, et al. Object decoding with attention in inferior temporal cortex. Available at http://cbcl.mit.edu/people/emeyers/pnas2011/index.html.

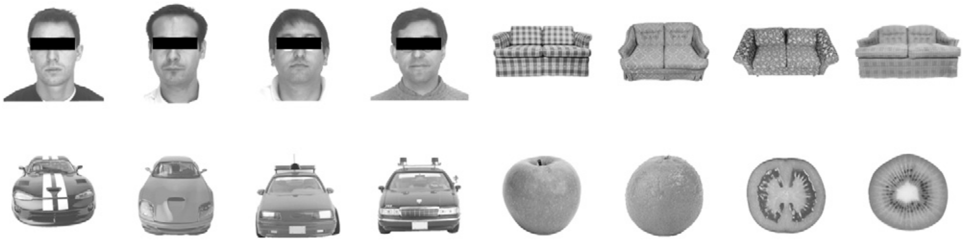

Fig. S1. Stimulus sets. The stimuli came from four categories (face, couch, car, or fruit). Two-thirds of the multiple-object trials consisted of all three images from the same category, and one-third of the trials consisted of images from different categories. For all analyses reported in this paper, all images were treated the same (i.e., the category of the stimuli was ignored).

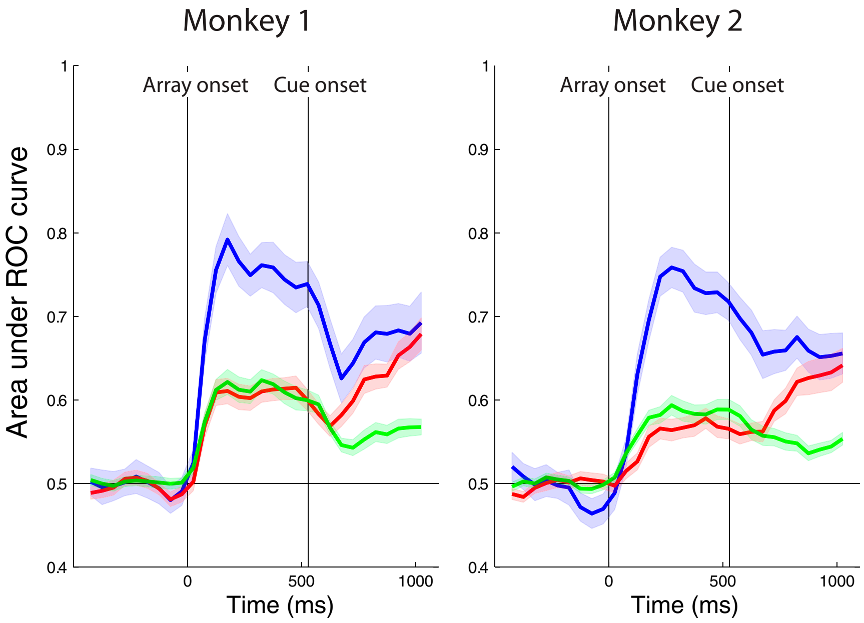

Fig. S2. Graphs showing decoding results for both monkeys separately. The two monkeys show a similar pattern of reduced decoding accuracy when multiple objects are presented (red and green traces) compared with when only a single object is presented (blue trace) before the onset of the attentional cue. After the onset of the attentional cue (indicated by the black vertical line at ~500 ms), information about the attended object increased (red trace), whereas information about the nonattended objects decreased (green trace) for both monkeys. Because the results were similar for both monkeys, we combined data from both monkeys for all other analyses. The findings are in agreement with Agram et al. (1), who also showed decreased performance when multiple objects were present.

1. Agam Y, et al. (2010) Robust selectivity to two object images in human visual cortex. *Current Biology* 20:872–879.

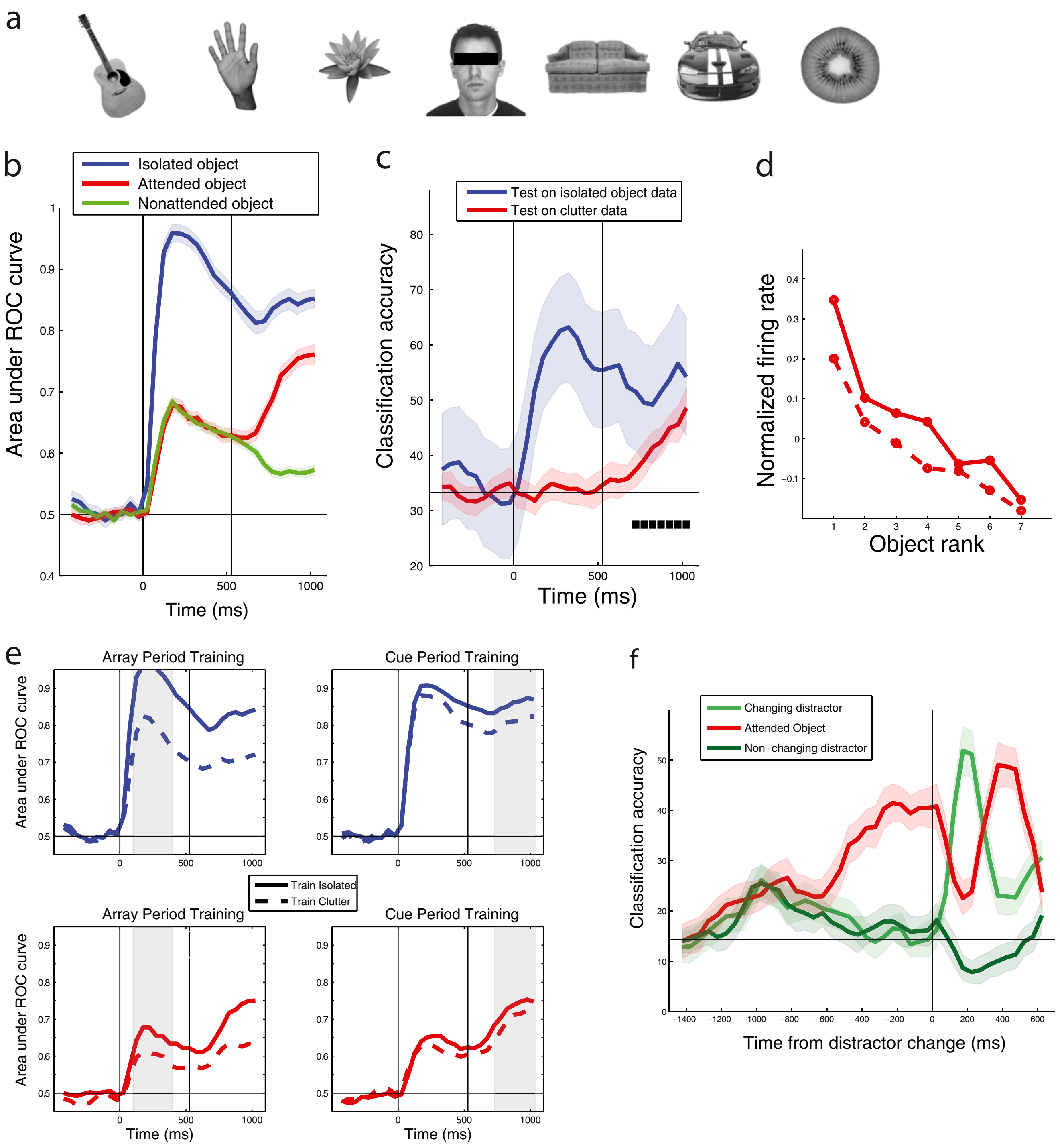

Fig. S3. Replication of the results on a second stimulus set. (*A*) We replicated the experiment on the second monkey using a new stimulus set that consisted of seven unique objects (*B–D*). The decoding results were similar to those with 16 stimuli, except the decoding accuracy for the attended object did not reach the accuracy seen for the isolated object. (*E*) Replication of the results in Fig. 2 using the seven-stimulus set. (*F*) Replication of the results in Fig. 3 using the seven-stimulus set. Chance performance is 1/7, or 14.29%.

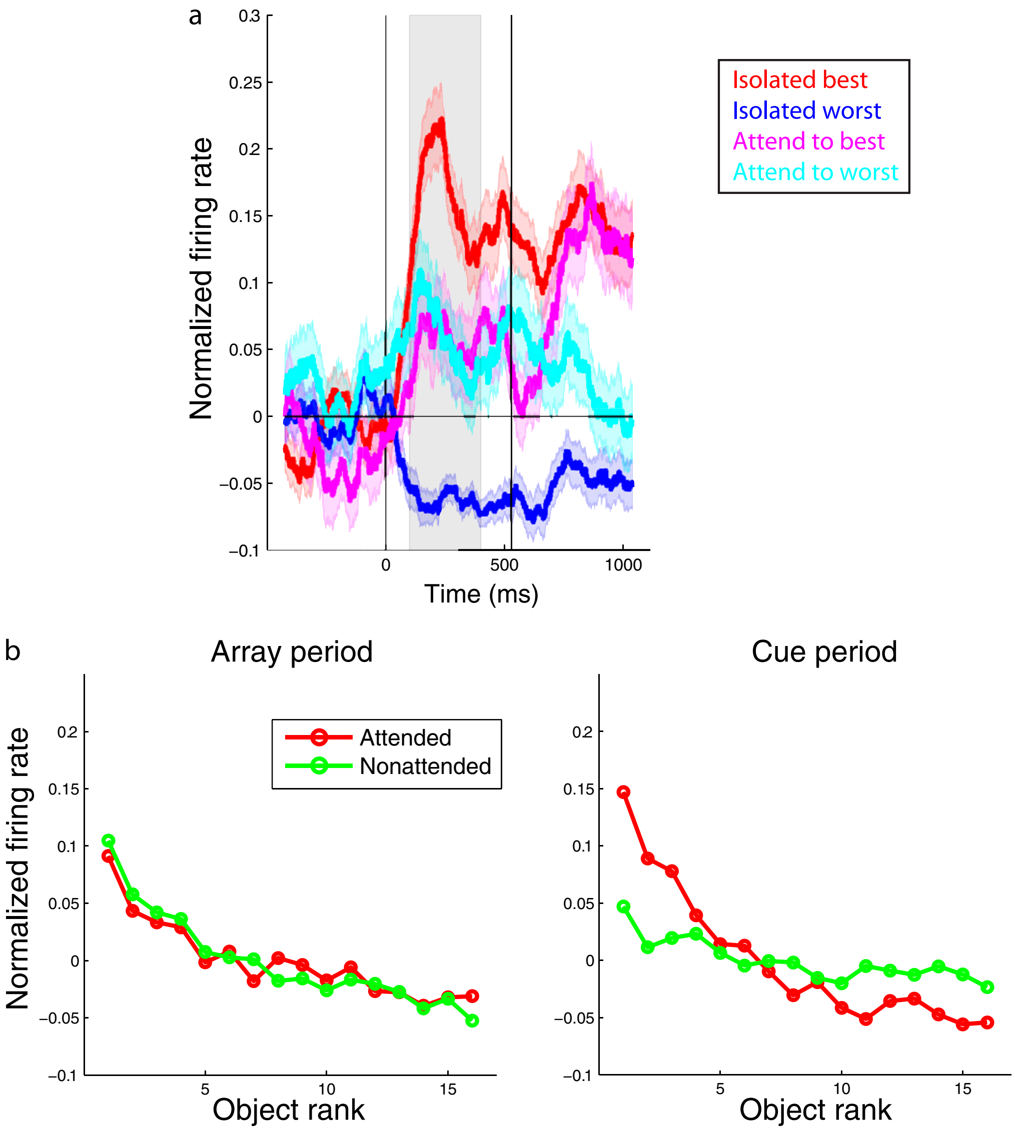

Fig. S4. Effects of attention on firing rates averaged across the population of cells. (*A*) Z-score–normalized firing rates to the "best stimulus" (the stimulus that elicited the highest firing rate) and the "worst stimulus" (the stimulus that elicited the lowest firing rate) on isolated-object trials (red and blue traces), and on three object trials (magenta and cyan traces). The best and worst stimuli were found using data on isolated-object trials using a time period from 100 to 400 ms after stimulus onset (gray shaded region). To highlight the attention-related effects (and ignore the stimulus-based effects) in the three-object data, only the three-object trials that had both the best and worst stimuli were used in this analysis. To correct for selection biases on the isolated-object results, we randomly shuffled the labels of the stimuli, found the best and worst stimuli on these shuffled data, and subtracted these randomly shuffled best to worst firing rates from the best and worst firing rates obtained from the real stimuli. The results were averaged over all neurons and the colored shaded regions are one SEM. (*B*) Z-score–normalized firing rates for the stimulus that was cued on the three-object trials, sorted from the best to worst stimulus as determined on isolated-object trials. Before attention was deployed the ranking seen on isolated-object trials was preserved (*Left*), and after attention was deployed this ranking was accentuated for the attended stimulus and largely abolished for the nonattended stimulus.

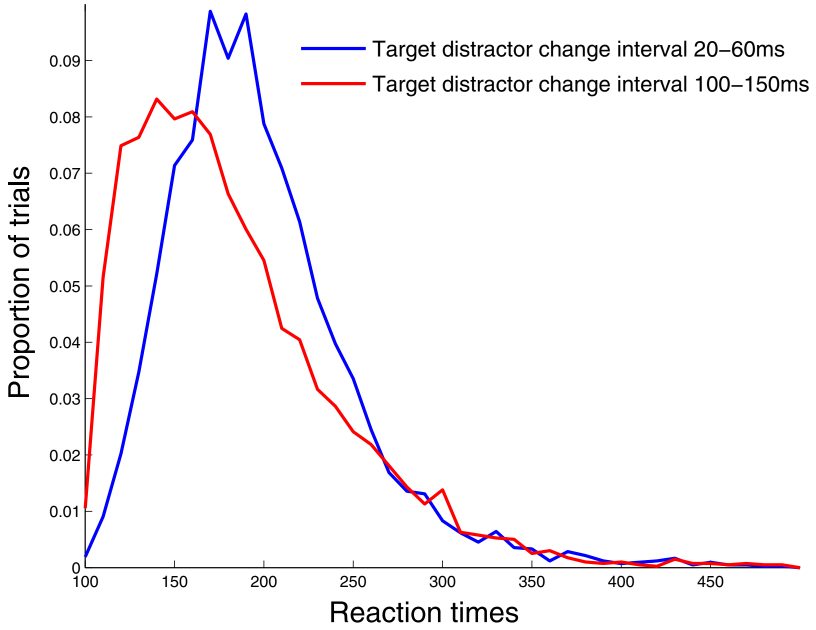

Fig. S5. Reaction times were slower when the target changed color soon after the distractor changed color. The distribution of reaction times on trials when the time difference between the target and distractor color change was 20–60 ms (blue trace) was compared with when the time difference was 100–150 ms (red trace). As can be seen, the distributions were shifted to longer reaction times when the time between the target and distractor changes was short. This increase in reaction time could be related to the fact that changes in distractor saliency caused information in inferior temporal cortex to be dominated by properties related to the distractor immediately following the distractor color change.

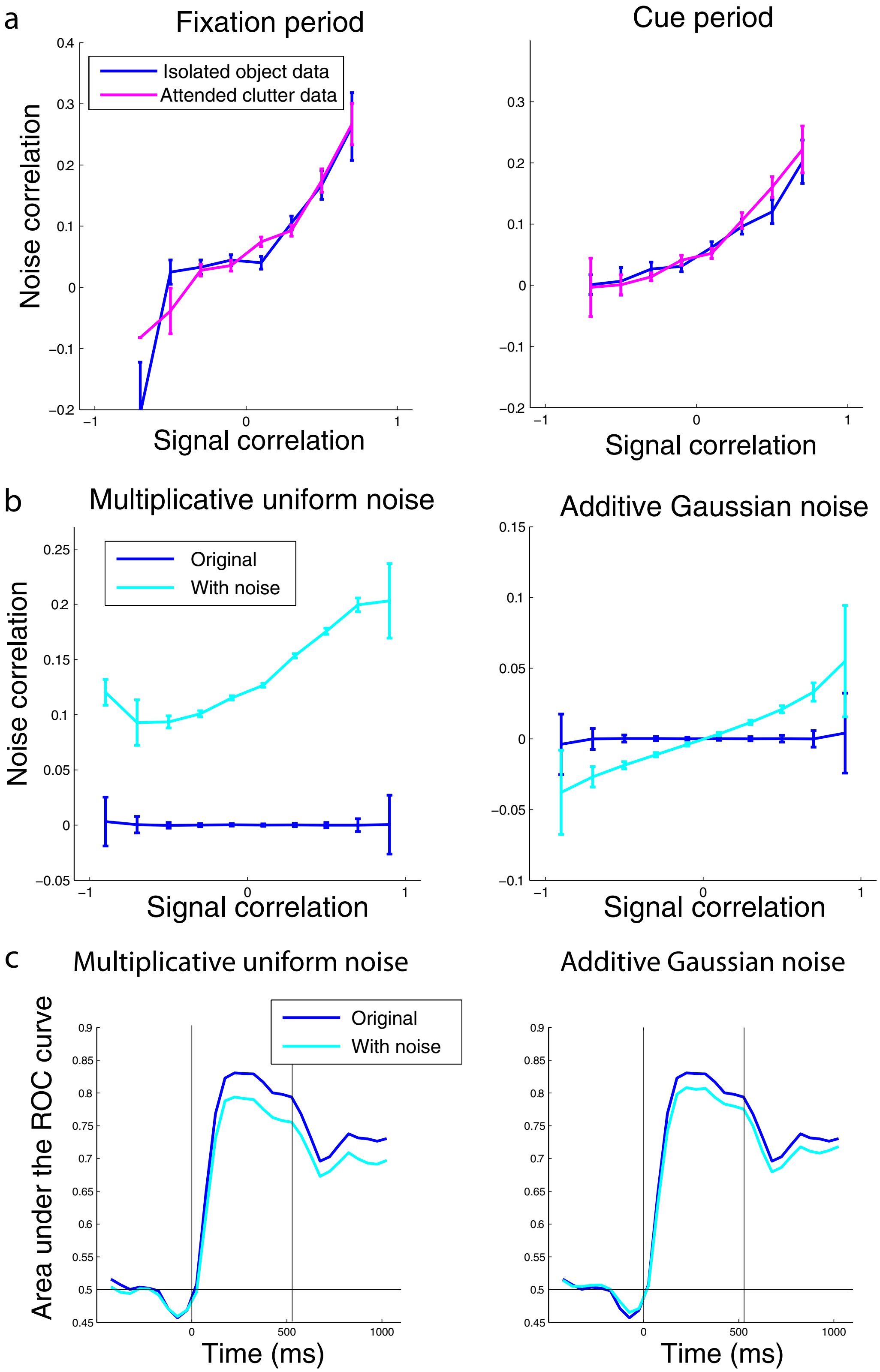

Fig. S6. Adding noise correlations had only a slight impact on population decoding accuracy. (*A*) Noise correlations as a function of signal correlations for the isolated-object trials (blue traces) and for the attended object in the three-object trials (magenta traces) for the fixation (*Left*) and cue periods (*Right*). The signal correlations during the fixation period were calculated using the upcoming stimuli, and are therefore essentially meaningless—so the fact that noise correlations still increase with signal correlations is an artifact at least for the baseline period (see *SI Text* for more details). (*B*) To add noise correlations to our pseudopopulation data, we either multiplied our pseudopopulation with uniform random variables in the range of [1 10] (*Left*) or added scaled zero-mean Gaussian noise vector using the signal correlations as the covariance matrix (*Right*). The noise correlations we created were of similar magnitude to those seen

Legend continued on following page

in our data and to those seen in the previous literature (10, 11). Results shown here are from the cue period, although similar noise correlations were created for data from all time periods. (*C*) Decoding accuracies on pseudopopulations that included either the multiplicative uniform noise (*Left*, cyan trace) or the additive Gaussian noise (*Right*, cyan trace) were only slightly lower than the decoding accuracy on the original pseudopopulations that did not include noise correlations (blue traces). This tentatively suggests that noise correlations might not have a large impact on decoding accuracies (at least for the classifiers used). However, more thorough analysis using data that were recorded simultaneously is needed before any strong conclusions can be drawn.

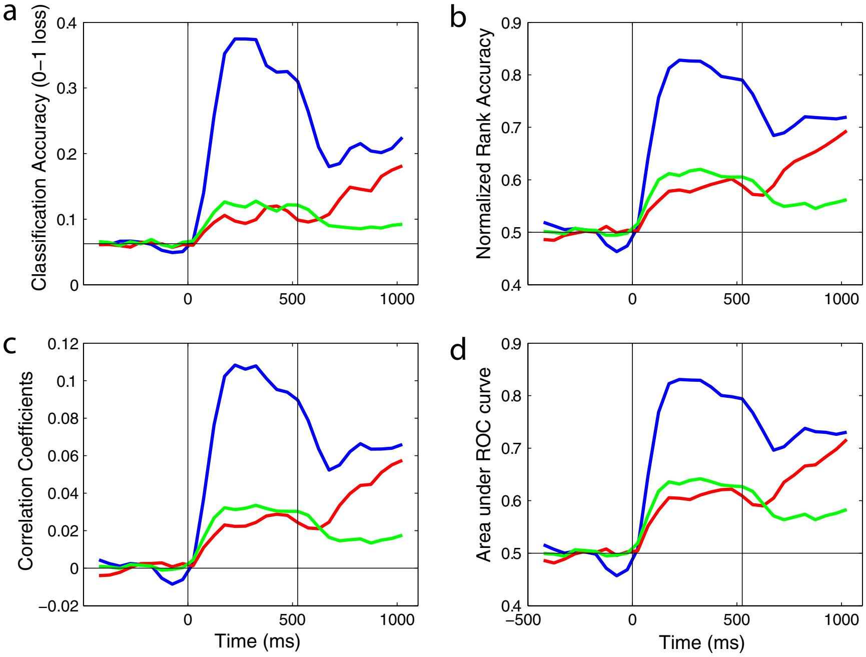

Fig. S7. Different measures of decoding accuracy yielded similar results. All results in this figure are for decoding the isolated object or the attended and nonattended objects (i.e., the same decoding problem as in Fig. 1*B*). (*A*) Zero-one loss results (results from the nonattended objects have been divided by 2 to equalize chance performance because there are two correct classes on each cluttered display). This measure is the most widely used decoding accuracy measure; however, biases can arise when using it to decode multiple objects from data from a single trial (see *SI Text* on measuring decoding accuracy). (*B*) Normalized-rank decoding accuracy (4). This measure can also be biased when decoding multiple objects from a single trial. (*C*) Average correlation coefficient values. These correlation coefficient values (which are the same as the classification scores that went into creating the AUROC measure) are unbiased in the multiple-object setting; however, they do not give a real measure of how often objects will be incorrectly decoded but instead just assess the similarity of the population activity on a given trial to the different class mean vectors (i.e., one could have equal average correlation coefficient values under different conditions but still be better at decoding under one condition due to less trial-by-trial variability in these values). (*D*) AUROC results (same as Fig. 1*B*). This method gives a decoding measure that allows one to compare the accuracy for multiple-object predictions from a single trial in an unbiased way.

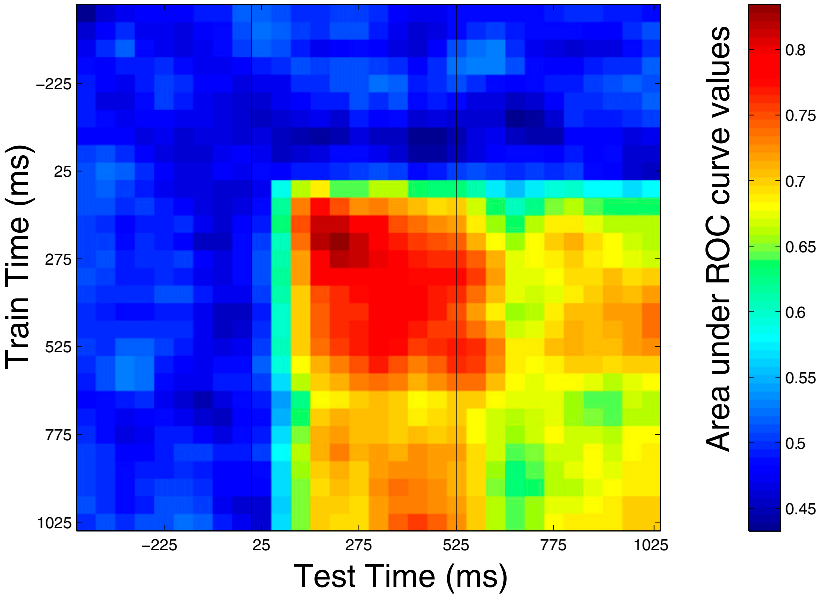

Fig. S8. The neural code for the identity of the stimuli is largely stationary over the time course of a trial. Previous work (1, 13) has found that more abstract/memory-related information is coded by a dynamic population code (i.e., different patterns of neural activity contain information about the same variables at different points in time in an experiment), whereas more visual-based information is contained in a static code (i.e., the same pattern of neural activity codes for an object at all time points in a trial). To test whether a dynamic or static code was present in this experiment, we trained the classifier at one point in time (using 150 ms of data) and then tested the classifier at either the same or a different time period. The results suggest that the code for the identity of the stimuli in this experiment was largely static (as indicated by the fact that there is not a strong diagonal of high decoding accuracy in the figure), which is consistent with previous findings. Based on the fact that the neural code for identity information was static (and that the highest decoding accuracies occurred when the stimuli were first shown), we trained the classifier using 500 ms of data from when the stimuli were first shown and tested at all other time points (although similar results were obtained when training with sliding 150-ms bins).