Supplemental Movies
Movie 1. Target Detection for Subject 1 (GP 6)
Movie 2. Target Detection for Subject 2 (GP 1)
Movie 3. Target Detection for Subject 3 (GP 7)
Movie 4. Target Detection for Subject 4 (GP 5)
Movie 5. Working memory (1-back) Target Detection for Subject 1 (GP 6)

These movies depict the spatio-temporal topographies of the real-time HG AA to the target (above) and nontarget (below) stimuli during the 0-back task for each subject (Movies 1-4). Movie 5 shows the same for the 1-back condition for Subject 1. These provide more information than the 4 single time frames chosen in panel E of Figure 1. The volume rendered MRIs here are the patients’ individual brains with electrode co-registration.

Supplemental Methods
In the following, we describe the background and details of the procedure for statistical assessment of the high-gamma analytic amplitude (AA) results. There are two statistical tests to perform on the data and it is important to keep them conceptually distinct. First, there is the single-condition question, where it is asked if the AA at some latency is significantly different from the baseline AAs for the same condition. Second, there is the two-condition or comparison question, where it is asked if the AA at some latency is significantly different from the AA at the same latency for another condition. In the present study, the two conditions are targets and non-targets, which are two categories of stimulus-response trials occurring within the same overall block of data. We begin by describing in detail the single-condition test, and describe the two-condition test further below. In both discussions the issue of minimizing multiple-comparisons is emphasized, given the large number of time samples (original sampling rate 2003 Hz) and space samples (64-128 electrodes). We take an overall mass univariate approach, as is typically done for fMRI (Frackowiak 2004) whereby each space-time pixel (or space-time-frequency voxel, in a full time-frequency analysis) is tested separately and the entire mass submitted to a multiple-testing correction procedure.

Preliminaries: the single-trial stack

Both tests begin with a “stack” of the single-trial AAs, which is a matrix of size \( N \times S \) that we call \( T_{\mu V} \). The \( T \) is a capital “tau” for “trials”, and it is in units of \( \mu V \) (the subscript will be changed after normalizing by the baseline mean). \( N \) is the total number of single-trials after artifact rejections, \( N = N_t + N_n \), where \( N_t \) and \( N_n \) are the number of trials in conditions 1 and 2, respectively. In our task, the number of non-targets (\( N_n \)) was roughly 80 and the number of targets (\( N_t \)) was roughly 30. For convenience, the single-trials are ordered such that the first \( N_t \) rows of \( T_{\mu V} \) are the single-trials of condition 1, and the last \( N_n \) rows of \( T_{\mu V} \) are the single-trials of condition 2.

The number of columns in \( T_{\mu V} \) is the number of samples, \( S \), in the epoch taken around each event. These samples are taken at a lower sample rate (250 Hz) than the original recordings (2003 Hz) for computational efficiency. Recall from the main article that the AA time-series for a given electrode is of the same total length and sample rate as the original signal. But the original sampling rate is not the real temporal resolution of the AA. By definition, the AA contains modulations of the envelope of a carrier signal such that the modulations are slower than the carrier frequency. Thus, if the carrier frequency is 100 Hz, then the AA modulations occur no faster than 100 Hz. This makes intuitive sense, because it is practically meaningless to discuss, say, a 200 Hz modulation of a 100 Hz carrier wave. In practice, there is little meaningful modulation at even half the carrier frequency, and since we average across frequencies from 70-160 Hz, the modulations become even slower (averaging is always a low-pass filter). Thus, by the Nyquist sampling theorem, we needn’t sample the AA faster than about double the center frequency of the filter. We chose a resampled timeline at 250 Hz around each event (stimulus). Also recall that the target trials are realigned to a resampled timeline at 250 Hz such that the interval from 350 ms post-stimulus to 100 ms pre-response is expanded/contracted according to the median RT. The non-realigned single-trials are shown in the SST plots (e.g. Fig. 1C) in z-score units, but for all other analyses we use the single-trial stack made after realignment of the target trials. We use an epoch from \(-250\) to either \(1200\) or \(1300\) ms, relative to stimulus onset, so at 250 Hz this gives \( S = 363 \) or \( S = 388 \). The number of samples in the baseline \((-250\) to \(0\) ms) is \( S_0 = 63 \).

Because we work with amplitude and not power (squaring the amplitude to obtain power is a nonlinear transform, as is conversion to dB units), and because averaging across trials is a purely linear process, we can normalize by the baseline mean either before or after averaging. Also, conversion to % units \((x_{0\%} = 100 \cdot x/\mu - 100)\) is a purely linear transform that only
relabels the ordinate, and therefore can be done on the single-trials. This conversion centers the distribution on 0 %, making the definition of the null hypothesis below easier, so we will do this conversion on the single-trials for convenience. We normalize each single-trial (row in \( T_{µν} \)) by the appropriate baseline mean (\( µ_{b1} \) for condition 1 and \( µ_{b2} \) for condition 2), and convert to % units, to give the stack of normalized single-trials, \( T_{γb} \):

\[
µ_{b1} = \frac{1}{N_1 S_b} \sum_n \sum_s T_{µν}(n, s), \text{ for } 1 \leq n \leq N_1, \text{ and } 1 \leq s \leq S_b,
\]

\[
µ_{b2} = \frac{1}{N_2 S_b} \sum_n \sum_s T_{µν}(n, s), \text{ for } N_1 + 1 \leq n \leq N, \text{ and } 1 \leq s \leq S_b.
\]

\[
T_{γb}(n, s) = \frac{100}{µ_{b1}} T_{µν}(n, s) - 100, \text{ for } 1 \leq n \leq N_1, \text{ and } 1 \leq s \leq S.
\]

\[
T_{γb}(n, s) = \frac{100}{µ_{b2}} T_{µν}(n, s) - 100, \text{ for } N_1 + 1 \leq n \leq N, \text{ and } 1 \leq s \leq S.
\]

Averaging across trials gives the event-related AA, \( EAA \), which is of size 1-by-\( S \). The \( EAA \) for \( γ_{high} \) is the basic activation time-series used in most results (e.g., Fig. 1D-E, Suppl. Movies, etc.). For conditions 1 and 2 respectively:

\[
EAA_1(s) = \frac{1}{N_1} \sum_n T_{γb}(n, s), \text{ for } 1 \leq n \leq N_1, \text{ and}
\]

\[
EAA_2(s) = \frac{1}{N_2} \sum_n T_{γb}(n, s), \text{ for } N_1 + 1 \leq n \leq N.
\]

The values in \( EAA \) are in % units, where -100 % corresponds to an AA of 0 \( µV \), 0 % corresponds to the mean AA of the baseline, 100 % to double the baseline mean, etc.

**Single-condition analyses**

The single-condition statistical question asks if the \( EAA \) at some latency is significantly different from the baseline AAs for the same condition. For simplicity, we only discuss the test for condition 1, with the test for condition 2 being identical. We take a nonparametric approach that does not assume knowledge of any underlying distribution, and obtain \( P \) values from bootstrap confidence intervals. Although there have been several alternative schools of statistical inference (Fisher’s hypothetico-deductive approach; the Neyman-Pearson decision-theoretic approach; the Bayesian approach; etc.) and sometimes acrimonious debate between them, all schools and authors have argued in favor of confidence intervals of some sort. To obtain bootstrap confidence intervals, a distribution is built up around the observed value, in this case \( EAA(s) \). This same distribution can be used to assign a significance level (Efron and Tibshirani 1993) based on the percentile position of the null value within this distribution. The details of our test are given below, but we begin by outlining the use of a range null hypothesis instead of a point null hypothesis.

Shortly after the introduction of the null hypothesis significance test (NHST) by Fisher (1935), a number of criticisms were raised and a healthy debate ensued in the general statistical and psychological literature. This debate is generally unknown in the psychophysiology and neuroimaging literature, so we give a brief review. The main criticism that concerns us here is that a point null hypothesis, such as \( H_0: EAA(s) = 0.000 \), is always technically false (Meehl 1967) and therefore becomes too easy to reject as the precision of the test or the number of observations increases. This point was first raised by Berkson (1938), who noted that “as a matter of observation, when the numbers in the data are quite large, the \( P \)‘s tend to come out small.” The first attempt at mathematical formalism was given by Hodges and Lehman (1954), who distinguished “statistical significance” from “material significance” and attempted to get at the latter by using a range null hypothesis. This criticism was echoed by subsequent workers (Bulmer 1957; Nunnally 1960), and Kish (1959) aptly called the null hypothesis of exactly zero difference as “the nullest of null hypotheses.” Binder (1963) first used the term “point null hypothesis” and suggested that we “recognize a more or less broad indifference zone about the null hypothesis consisting of values which are essentially equivalent to the null hypothesis for our present theory or practice.” An excellent early introduction to this issue, to which the interested reader is referred, is Bakan (1966), who distinguished “sharp” and “loose” null hypotheses. This issue has by no means been forgotten and has generated repeated debate in the statistics and psychology literatures (Morrison and Henkel 1970; Serlin and Lapsley 1985; Cohen 1994; Hertwig and Todd 2000). The specific advice that we follow is the “good-enough” principle of Serlin and Lapsley (1985; 1993). They suggest that “a good-enough belt of width \( Δ \) must also be included in the prediction, so that a value \( 0 ± Δ \) is predicted,” and they show that this avoids the above mentioned paradox...
As a matter of experience with several EEG/ECoG data sets, it is found that the point null hypothesis, \( H_0: ERAA_1(s) = 0 \), erroneously rejects \( H_0 \) for many baseline latencies, i.e., it is far too liberal. It also fails an intuitive requirement that a noisier baseline should make it more difficult to obtain significance, because we would test our confidence limits against precisely 0 \%, regardless of the baseline variance around 0 \%. Based on all of these considerations, we use the following range null hypothesis:

\[
H_0 : \ ERAA_1(s) - \Delta = 0 \quad \text{for } s: ERAA_1(s) > \Delta.
\]

\[
H_0 : \ ERAA_1(s) + \Delta = 0 \quad \text{for } s: ERAA_1(s) < -\Delta,
\]

where \( \Delta \) is taken as the bootstrap standard error of the baseline means. \( \Delta \) will be larger for noisier baselines and hence make it more difficult to reject \( H_0 \). In our data set, a typical value for \( \Delta \) is \(-3-4 \%)\), so a typical null hypothesis is that \( ERAA_1(s) \) falls within the range 0 ± 3 \%. Our assigned significance level (P-value) will be, not the probability that \( ERAA_1(s) \) is different from the baseline mean, but the probability that it exhibits greater than 3 \% difference from the baseline mean.

This also gives an important means of reducing the multiple-testing problem without making any \textit{a priori} assumptions about the distribution of the data or about specific latencies of interest (LOIs): only those latencies exceeding the standard error of the baseline are tested for significance relative to the baseline. These might be thought of as empirical LOIs whose identification depends only on the same assumption that went into the null hypothesis. Namely, departures less than the “good-enough” belt, \( \Delta \), are not \textit{materially} significant and therefore cannot achieve \textit{statistical} significance by an adequate procedure. By definition of the range null hypothesis, these latencies cannot reach significance and so needn’t be tested at all, reducing computational time and the multiple-testing problem.

With this formulation, the hypothesis testing uses the bootstrap procedure as described in section 15.4 of Effron and Tibshirani (1993). Their example involves testing an observed mean, \( \hat{\theta} \), against the null hypothesis that it equals 0. A bootstrap distribution is built around \( \hat{\theta} \), and the assigned significance level (ASL) is the position of 0 within the bootstrap distribution. Our inclusion of \( \Delta \) prevents this from being a point null hypothesis. They use the improved BC\(_a\) (“bias-corrected and accelerated”) method (see also their section 14.3), rather than the percentile of the distribution directly. This bootstrap procedure is described next in detail, where \( \hat{\theta} \) is taken as the value of \( ERAA_1(s) - \Delta \) or \( ERAA_1(s) + \Delta \).

The end result of the bootstrap procedure at each latency is a “raw” (uncorrected) P-value. The raw P-value is called by Effron the “assigned significance level”, \( \text{ASL}_{BC_a} \). The subscript \( BC_a \) indicates the bias-corrected and accelerated method, and the hat indicates that the \( \text{ASL} \) is estimated from a particular sample and bootstrap distribution. The bootstrap distribution around \( \hat{\theta} \) is built up from \( B = 10000 \) bootstrap samples from the single-trials in \( T_{\text{no}} \). There are \( N_1 \) single-trials and so each replication includes \( N_1 \) single-trials drawn at random with replacement from the original set. In terms of Matlab code, the set of trials for each bootstrap replication is defined by: \( \text{ceil(rand}([1 \ N_1])) \) \* \( N_1 \). This gives a vector of size 1-by-\( N_1 \) containing random integers between 1 and \( N_1 \). For each bootstrap iteration, \( b \), \( \hat{\theta} \) is computed and the set of \( B \) bootstrap replications of \( \hat{\theta} \) is \( \hat{\theta}^*(b) \), also called loosely the “bootstrap distribution”. By the ordinary percentile method, the \( \text{ASL} \) is the proportion of bootstrap replications less than 0 (or greater than 0 for negative values).

\[
\text{ASL}_{\text{no}} = \#\{\hat{\theta}^*(b) < 0\}/B, \text{ where the } % \text{ subscript indicates the ordinary percentile method.}
\]

However, \( \text{ASL}_{\text{no}} \) values of 0 or 1 are not allowed, and are replaced by \( 1 \) by \( B \) and \( \frac{B-1}{B} \), respectively. To complete the procedure, we have only to introduce the modification based on the \( BC_a \) method. The formula for \( \text{ASL}_{BC_a} \) is found on p. 216 of Effron and Tibshirani (1993), but note that this page contains several errors. Consulting the online errata, the correct formulas are:

\[
\text{ASL}_{BC_a} = \Phi \left(\frac{w_0 - \hat{\theta}_0}{1 + \phi(w_0 - \hat{\theta}_0)} - \hat{\theta}_0\right), \text{ where } \Phi(\cdot) \text{ is the standard normal cumulative distribution function.}
\]

\[
w_0 = \Phi^{-1}(\text{ASL}_{\text{no}}), \text{ where } \Phi^{-1}(\cdot) \text{ is the inverse of the standard normal cumulative distribution function.}
\]

\[
\hat{\theta}_0 = \Phi^{-1}(\#\{\hat{\theta}^*(b) < \hat{\theta}\}/B), \text{ where } \hat{\theta}_0 \text{ is the bias-correction constant.}
\]
\[
\hat{a} = \frac{\sum (\hat{\theta}(j) - \bar{\theta}(n))^3}{6\left[\sum (\hat{\theta}(j) - \bar{\theta}(n))^2\right]^{3/2}}, \text{ for } 1 \leq n \leq N_1, \text{ where } \hat{a} \text{ is the acceleration constant.}
\]

The terms in the rather complicated expression for \( \hat{a} \) are explained on p. 186 of Efron and Tibshirani (1993). The formula is based on the jackknife values of the statistic \( \hat{\theta} \). \( \hat{\theta}(n) \) is the value of the statistic computed with the \( n \)th observation (single-trial) deleted, and:

\[
\hat{\theta}(1) = \sum_n \hat{\theta}(n)/N_1, \text{ for } 1 \leq n \leq N_1.
\]

Because of the simple (linear) computation of the test statistic \( \hat{\theta} \), for our case \( \hat{\theta}(1) = \hat{\theta} \), which simplifies and speeds up the calculation of \( \hat{a} \). The full formula is given in case one were to use a more cumbersome test statistics, such as dB power. These formulas and the BC\(_a\) method are discussed further in by Efron and Tibshirani (1993) and DiCiccio and Efron (1996). This does not change the A\(\hat{S}L\) drastically from the ordinary percentile method, but once understood it is fast to compute and improves the accuracy of the estimate.

Because we test the decreases/increases of AA separately against \( \pm \Delta \), basing the direction tested on the obtained AA, we double the A\(\hat{S}L\) as required for a two-tailed test (Efron and Tibshirani 1993; Good and Hardin 2003). This two-tailed A\(\hat{S}L\) is the raw \( P \)-value. After this raw \( P \)-value has been computed at each electrode, latency and condition separately, the full set of raw \( P \)-values is submitted together for multiple-testing correction as described further below.

**Two-condition statistical test**

The **two-condition** or **comparison** question asks if the values at a given electrode and latency are significantly different between conditions 1 and 2 (Non-targets vs. Targets) – that is, if the value of \( ERAA_1(s) \) significantly different from \( ERAA_2(s) \) at a given \( s \).

The logic of the permutation test is that we have lost the labels that assign single-trials to condition 1 or 2. On each replication random labels are assigned to the single-trials, which is the same as sampling without replacement. That is, all \( N \) single-trials are sampled on each replication, but with random assignment of labels such that \( N_1 \) single-trials are assigned to condition 1 and \( N_2 \) single-trials are assigned to condition 2. Each such sample without replacement is a permutation, and the calculation of the test statistic for all possible permutations would constitute the exact test. However, there are \( N! \) possible permutations, so for even moderately large \( N \) this is not feasible and we instead perform a Monte Carlo permutation test. The “Monte Carlo” refers to the fact that a random subset of all possible permutations is used to estimate the exact test. For each of \( B = 10000 \) permutations, the test statistic \( \hat{\theta} \) is recomputed from \( T_n^\prime \) and the set of \( B \) replications is \( \hat{\theta}^\ast(b) \). The test statistic for the current study is:

\[
\hat{\theta} = ERAA_2(s) - ERAA_1(s), \text{ for a given latency } s.
\]

The null hypothesis of the permutation test is that the distributions of conditions 1 and 2 are equal (see p. 202 of Efron and Tibshirani 1993):

\[H_0 : F_1 = F_2, \text{ where } F \text{ stands for a distribution.}\]

The set of permutation replications \( \hat{\theta}^\ast(b) \) gives the distribution of the test statistic expected under \( H_0 \). Note the logical difference from the bootstrap test. The bootstrap distribution is built up around \( \hat{\theta} \) and the A\(\hat{S}L\) is the position of 0 (or, more generally, of \( H_0 \)) in the distribution. The permutation distribution is built up around 0 (or more generally, around \( H_0 \)) and the A\(\hat{S}L\) is the position of \( \hat{\theta} \) within the distribution:

\[
\hat{\theta}^\ast(b) \text{ if } |\hat{\theta}| < \hat{\theta}^\ast(b) \text{ and } B.
\]

The absolute values are used in the above formula to make the test two-tailed (see p. 212 of Efron and Tibshirani 1993). That is, we are equally interested in situations where \( ERAA_2(s) \) is greater or where \( ERAA_1(s) \) is greater, and at a given latency we have no a priori hypothesis about which condition is greater (even if we did have some biased interest or
predictions, a two-tailed exploratory approach is probably still preferable so as to allow the opposite result an equal chance of success).

The \( A_{\text{FDR}} \) values constitute the raw \( P \)-values, and the final \( P \)-values are obtained by submitting the whole mass to a correction procedure (FDR, see below). To reduce the multiple-testing problem, we can eliminate a number of latencies from consideration even without any \textit{a priori} assumptions about the data or latencies of interest. Namely, in the present study we are not interested in comparing \( E_{\text{MAR}}(s) \) to \( E_{\text{MAR}}(s) \) if neither of them is individually significant at \( s \). Thus, we only test latencies where at least one condition tested positive (\( P < 0.05 \)) by the above single-condition test. This reduces the number of tests submitted to the multiple-testing correction procedure.

**Correcting for multiple-comparisons**

Regardless of how the \textit{raw} \( P \)-values are obtained, we have tested a great number of space-time points and some procedure is required to minimize the number of false positives. The end result we seek is a set of \textit{corrected} \( P \)-values that are larger than the \textit{raw} \( P \)-values, and are therefore more conservative in rejecting the null hypothesis. However, we do not want the correction to be overly conservative (i.e., Bonferroni correction), which would give too many false negatives and therefore lack in statistical power. Because we have taken an overall mass univariate approach, where each space-time point is tested separately and then the whole mass of \textit{raw} \( P \)-values submitted to correction, our multiple-testing problem is very similar to that of fMRI statistics where each voxel is tested separately and then corrected \textit{en masse}. Our chosen procedure of correction, the false discovery rate (FDR), was originally identified in a survey of the fMRI statistics literature, where it was introduced by Genovese et al. (2002) and in the review by Nichols and Hayasaka (2003). The term “false discovery” was introduced in the statistical literature by Sorõ (1989), since null hypotheses that are rejected are “discoveries” and incorrect rejections (Type I errors) are “false discoveries”. The FDR and a method for its control were introduced in the now-classic work of Benjamini and Hochberg (1995). The FDR method was only more recently used in the EEG literature (Durka et al. 2004; Edwards et al. 2005) Here, we summarize the main features and advantages of this procedure for exploratory research.

The FDR is ideal for situations where one does not want to begin with too many \textit{a priori} assumptions about particular regions or latencies of interest (ROIs, LOIs). Such \textit{a priori} assumptions restrict the number of regions and latencies tested, which helps the multiple-testing problem, but note that it also restricts the number of potential discoveries. If one only tests the frontal lobe because that is the only ROI by some preconceived theory, then the experiment cannot make a discovery outside of the frontal lobe. This is like a reduction in statistical power (i.e. prone to false negatives), but is even worse because the whole experiment lacks the power to make a potentially meaningful discovery. Increasing the number of ROIs increases the possibility of making novel discoveries and allows the data to speak for itself. For example, even if the location of the activation had been predicted, the extent of the focus may not have been. We seek to \textit{image} the brain, i.e. to look at all of our data and to sample as widely and densely in space and time as possible. But at the same time we want to filter out those pixels that are the most likely to represent spurious or unreliable activations. The FDR, which is less conservative than Bonferroni correction and more suited to exploratory research, better allows us to increase the number of regions and latencies examined while still maintaining statistical control.

Let the number of hypotheses tested within a given set (“family”) of multiple comparisons be denoted \( m \). The Bonferroni method seeks to control the family-wise error (FWE) rate. This is the probability that \textit{any} of the \( m \) hypotheses is falsely declared significant, i.e. that the family of tests contains even a single error. The Bonferroni correction to the raw \( P \)-value is: \( P/m \). One might ask, what are the bounds of the “family”? This question leads to one of the major practical advantages of the FDR over the Bonferroni method. If one has run an experiment and statistical tests on \( n \) subjects, but now an eighth subject is added, then all \( n \) subjects need to be recorrected by the Bonferroni approach. It has even been semi-facetiously discussed in the statistics literature that the Bonferroni philosophy requires you to go back and recorrect \( P \)-values from prior experiments with each new experiment. And journals may have to correct, at the end of each year, for a “year-wise error rate”; and so on. However, the FDR is stable on sufficiently large subsets of data such that if an FDR level of \( \alpha \) has been used on each subset than the total FDR level expected over subsets remains at \( \alpha \). This means that each experiment and each patient can be tested and corrected as they arise, without going back to old studies or previous patients to reevaluate statistical conclusions.

The FDR approach controls the proportion of false positives, rather than the FWE as in the Bonferroni approach. The logic of the FDR and its mathematical formulation have already been well described (Benjamini and Hochberg 1995; Nichols and Hayasaka 2003) and are not repeated here. In using the FDR, the expected proportion of false positives is controlled at a specified level, \( \alpha \). We use a level of \( \alpha = 0.05 \), meaning that roughly 5% of the \( P \)-values declared significant (i.e. 5% of the “discoveries”) are expected to be false. This is implemented by correcting the raw \( P \)-values such that a threshold of \( P < 0.05 \)
is used to control at the FDR level $\alpha = 0.05$. We caution that these $P$-values, like all $P$-values in time-frequency analyses that we know of in the published EEG literature, should be interpreted as estimates, not as literal probabilities that are strictly either significant or not significant. The total procedure should be interpreted as indicating which data points are reasonably certain to exceed chance fluctuations expected in the baseline. For the topographies (e.g., Fig. 1E), we use a threshold of $\alpha = 0.1$, because we do not want to artificially truncate the extent of and smooth fall-off of the activation foci, while also providing some level of statistical assurance.

So far we have described the statistics as if each latency, $s$, is tested separately. This would have been a valid approach, but in order to further reduce the multiple-comparisons problem, and because we don’t desire hypothesis testing resolution less than 20 ms, we have instead done the test on a set of non-overlapping 20 ms windows. At a sampling rate of 250 Hz, this puts 5 samples in each 20 ms window. Thus, only every 5th sample is tested, and it is first replaced by the mean of the 5 samples within the window (the above formulas can be thought of as the case when the window size is 1 sample, or if we had down-sampled to 50 Hz instead of 250 Hz). Only these mean-window samples are submitted to bootstrap or permutation procedures, and to the FDR procedure for correction. After correction, the $P$-value for this mean-window sample is assigned to all 5 samples within the window. The statistical results are plotted (e.g., Fig. 1D) with a dot centered on each window, with a diameter of ~20 ms, to represent the fact that the tests were done on 20 ms windows. It is seen in these plots that this gives more than adequate time-resolution for hypothesis testing in the present study, but reduces the number of hypothesis tests by one-fifth.

To summarize, we have used three ways to reduce the problem of multiple-testing:
1) We test only at a time-resolution desired for the current study; testing every 20 ms was deemed more than sufficient.
2) We test only at latencies that exceed the standard error ($\Delta$) of the baseline (where no pre-stimulus baseline exists, alternate methods of determining such a range are possible but not described here since they are not required in the current study). For the two-condition test, we only test latencies where the difference exceeded the range of differences seen in the baseline and only latencies where at least one of the two conditions was individually significant.
3) The FDR was used for multiple-testing correction, rather than the overly-conservative Bonferroni approach.

References


Cohen J (1994). The earth is round ($p < 0.05$). American psychologist, 49(12): 997-1003.


