Cognitive Science 32 (2008) 1323–1348

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1080/03640210802451588

Using fMRI to Test Models of Complex Cognition

John R. Anderson[a], Cameron S. Carter[b], Jon M. Fincham[a], Yulin Qin[a], Susan M. Ravizza[c], Miriam Rosenberg-Lee[a]

[a]*Psychology Department, Carnegie Mellon University*
[b]*Departments of Psychiatry and Psychology, University of California at Davis*
[c]*Psychology Department, Michigan State University*

Received 27 October 2007; received in revised form 9 June 2008; accepted 11 June 2008

Abstract

This article investigates the potential of fMRI to test assumptions about different components in models of complex cognitive tasks. If the components of a model can be associated with specific brain regions, one can make predictions for the temporal course of the BOLD response in these regions. An event-locked procedure is described for dealing with temporal variability and bringing model runs and individual data trials into alignment. Statistical methods for testing the model are described that deal with the scan-to-scan correlations in the errors of measurement of the BOLD signal. This approach is illustrated using a "sacrificial" ACT-R model that involves mapping 6 modules onto 6 brain regions in an experiment from Ravizza, Anderson, and Carter (in press) concerned with equation solving. The model's visual encoding predicted the BOLD response in the fusiform gyrus, its controlled retrieval predicted the BOLD response in the lateral inferior prefrontal cortex, and its subgoal setting predicted the BOLD response in the anterior cingulate cortex. On the other hand, its motor programming failed to predict anticipatory activation in the motor cortex, its representational changes failed to predicted the pattern of activity in the posterior parietal cortex, and its procedural component failed to predict an initial spike in caudate. The results illustrate the power of such data to direct the development of a theory of complex problem solving, both at the level of a specific task model as well as at the level of the cognitive architecture.

Keywords: Cognitive modeling; fMRI; Problem solving; ACT-R; Model evaluation

1. Introduction

Neural imaging has seen an amazing rise of prominence in cognitive science. It almost seems that unless one can see a process in the brain, one does not believe it exists. This growing emphasis on imaging evidence has a potentially negative consequence when combined with

Correspondence should be sent to John R. Anderson, Psychology Department, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: ja@cmu.edu

the focus on simple tasks in neural imaging. Because these tasks cannot reveal more complex processes, there is the danger that researchers will start to believe in a theory of the human mind that ignores its intellectual capacity. Although restricting some imaging techniques such as event-related potential (ERP) to brief tasks might be justified, this is not the case for functional magnetic resonance imaging (fMRI). The temporal resolution of fMRI, often a source of frustration in understanding the detail of simple acts of cognition, is well matched to the temporal grain size of complex cognition. The spatial resolution of fMRI is sufficient to allow identification of regions that have functional significance in the interpretation of complex cognition. This article takes a model of a modestly complex mathematical task and shows how fMRI can confirm aspects of the model, suggests how to improve other aspects, and indicates that certain aspects of the underlying theory might be fundamentally wrong. However, there are methodological hurdles to analyzing fMRI data from complex tasks and assessing the match of a model to the data. This article is concerned with these methodological issues.

fMRI data have been used many times to assess models of cognitive processes. Most of the analysis efforts have related different conditions of an experiment to different magnitudes of the blood oxygen level dependent (BOLD) response (e.g., Kable & Glimcher, 2007; Lange & Lappe, 2006), but we and others have made efforts to predict the time course of the BOLD response (e.g., Anderson, Qin, Jung, & Carter, 2007; Brown & Braver, 2005; Just & Varma, 2007). This article describes a methodology that can be used to test time-course predictions from a wide range of information-processing models. To illustrate the issues involved, we will use an experiment from one of our laboratories.

2. A study of equation solving

Ravizza, Anderson, and Carter (in press) reported a study of equation solving that was mainly focused on the neural correlates of errors. However, we just report an attempt to model the solution of correct trials. All equations involved performing two algebraic transformations but varied in terms of the complexity of the mental arithmetic required to solve the equations. In the *small-number condition,* the operands on the left side of the equation were numbers—for example, $(x + 4) \times 5 = 30$—which only required single-digit arithmetic (e.g., $30/5 = 6$, $6 - 4 = 2$). The *large-number condition* required at least one two-digit calculation—for example, $19 + (13 \times x) = 71$.

A trial involved twenty-two 2-second scans. A trial began with an alerting stimulus (an exclamation mark) presented for 1 second followed by an equation that stayed on the screen until the participant indicated he or she knew the answer or 39 seconds had elapsed. When participants felt that they had solved the equation, they pressed a key under the thumb of their right hand. When they had pressed the thumb key, a "counter" with a starting value of "00" was displayed in the center of the screen. Participants pressed a key under their index finger to increment the tens column and a key under their middle finger to increment the ones column. Once they had entered their answer, they pressed a key under their thumb to indicate that they were done. A fixation stimulus was then presented for the amount of time needed to ensure that the total duration of the trial was 44 seconds. The mean time from stimulus presentation

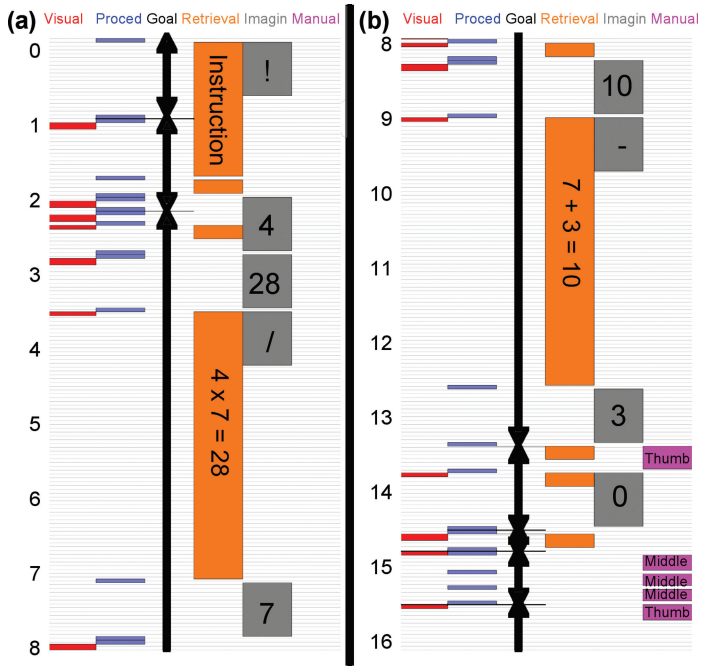


Fig. 1. Engagement of the visual, procedural, goal, retrieval, imaginal, and manual modules in the solution of the equation "28/(10 – x) = 4." Time is given in seconds. Lengths of boxes reflect approximate times the modules are engaged. Arrows indicate subtasks of activity controlled by the setting of a subgoal.

to the first thumb press was 13.7 seconds in the small-number condition and 20.8 seconds in the large-number condition.

We developed a model[1] in the ACT-R architecture (Anderson, 2007; Anderson et al., 2004) for this task. As the focus of this article is on the methodology for evaluating the match between cognitive models and fMRI data, the model is just a first-pass attempt and does not correspond in all aspects to the data. From the point of view of the article, it is a success that the methodology was able to identify the points of correspondence and the points of miscorrespondence. If we were more focused on getting the model right rather than providing an exposition of the methodology, we might have worked toward a better-fitting model and only reported that. However, the current model is basically a sacrificial lamb for purposes of illustrating the methodological issues.

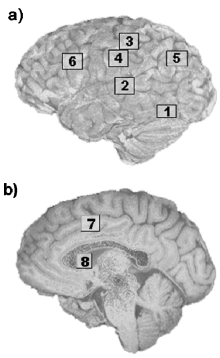
ACT–R assumes that cognition emerges through the interaction of a set of modules. Fig. 1 illustrates the activity of the six modules relevant to the performance of this task in the solution of the equation 28/(10 – x) = 4 (the small-number condition). It represents the 16 sec from the prompt through to the completion of the answer:

1. The visual module is active at many points throughout the performance of the task. For the first 9 seconds it is engaged in encoding parts of the equation as needed. At the end of the task (14–16 seconds) it is engaged in monitoring the counter display as the answer

is keyed out. Each of the small boxes in the visual column represents 85 msec (a fixed parameter of the ACT–R theory) for encoding an element from the screen.

2. The procedural module represents the firing of productions that unpack the logic involved in solving an equation of this form. Each of the small boxes in the procedural column represents 50 msec to fire a production (another fixed parameter of the ACT–R theory). These productions fire throughout the task, although there are two periods of quiescence while the arithmetic facts are being retrieved.
3. The goal module holds the various subgoals that control the different parts of task performance. The duration of these subgoals is indicated by the double-headed arrows: The initial subgoal controls retrieval of task instructions, the second subgoal controls selection of a strategy for solving the equation (there were a variety of equation forms that required different strategies), the third subgoal controls execution of the strategy, the fourth subgoal controls outputting the tens digit, the fifth subgoal controls outputting the ones digit, and a final subgoal bridges the interval until the next trial. These subgoal settings are important because a subgoal restricts the set of productions that can apply while it is in effect. Note that there is a single subgoal during the central period of equation solving. Although one could imagine setting subgoals during this phase, our observation is that the control structure of simple equation solving is flat with students just responding to the equation on the screen. Note that subgoal setting takes no time in ACT–R.
4. The retrieval module is occupied retrieving various pieces of information from declarative memory. The length of these retrievals is determined by a single latency scale parameter that is estimated to fit the latency data. The small boxes reflect retrieval of the values of specific numbers (4, 28, 7, 3, 0). These retrievals are quick because the facts are highly over-learned. The longer boxes involve retrieval of task instructions and two facts, one from the multiplication table and one from the addition table.
5. The imaginal module builds up representations of the arithmetic problems that need a solution—first "28 ÷ 7 = __," then "10 – 7 = __?," and then builds a representation of the answer in the form of separate tens and ones digits (e.g., "0" and "3") for keying. The duration of each of these updates to the problem representation takes a period of time that is also estimated to fit the latency data.
6. The manual module is activated to program the output of finger presses. The theory of the time for each finger press is taken from the EPIC theory (Meyer & Kieras, 1997).

As Fig. 1 illustrates, a model of the performance of such tasks makes many detailed assumptions about bits and pieces of information-processing activity. The behavioral data in such an experiment (in this case, latencies to solution) hardly provide justification for all the detailed assumptions about module activity in Fig. 1. Any complete model of how the task is performed will have to make assumptions about complex sets of processing steps. However, although the process is complicated, it does not necessarily follow that those complicated steps are anything like Fig. 1 in terms of the modules involved or sequences of operations. To get converging data to test the predicted patterns of activity in ACT–R modules, we have associated these modules with specific brain regions and developed a theory of how the activity of these modules relates to the fMRI responses in these regions. Fig. 2 illustrates the eight brain regions that have been mapped onto specific modules in ACT–R (from Anderson, 2007—the auditory and vocal



	X	Y	Z	Region
1. Visual	42	-61	-9	Fusiform
2. Aural	46	-22	9	Auditory
3. Manual	41	-20	50	Motor
4. Vocal	43	-14	33	Motor
5. Imaginal	23	-63	40	PPC
6. Declarative	43	23	24	LIPFC
7. Goal	7	10	39	ACC
8. Procedural	14	10	7	Caudate

Fig. 2. An illustration of the locations of the eight brain regions associated with ACT-R modules. In part (a) are the regions close to the surface of the cortex, and in part (b) are the regions deeper in the brain. The Talairach coordinates are for the left side, which are the areas reported in this research. Most of the regions are cubes 5 voxels long, 5 voxels wide, and 4 voxels high. The exceptions are the procedural (caudate), which is $4 \times 4 \times 4$; the goal (ACC), which is $5 \times 3 \times 4$; and the fusiform, which is $5 \times 5 \times 3$. A voxel is 3.125 mm long and wide and 3.2 mm high. PPC = posterior parietal cortex; LIPFC = lateral inferior prefrontal cortex; ACC = anterior cingulate cortex)

modules are not engaged in this experiment). Later, we describe how we predict the temporal course of activity in one of these regions from the temporal pattern of engagement of the associated module (as illustrated in Fig. 1).

3. Predicting the BOLD response

Fig. 3 illustrates how to predict the BOLD response for a region from the activity of a module. At the bottom of the figure is the activity of a hypothetical module—it is active for half a second starting at 1 second, for 1 second starting at 4 seconds, and for 1.5 seconds starting at 12 seconds. Each of these activities constitutes a metabolic demand on the region of the brain supporting this module and the figure indicates the BOLD responses reflecting each of these activities. Characteristic of the BOLD response, the signal is delayed and spread out over many seconds, reaching a peak about 4 to 5 seconds after the actual metabolic demand. As is typical (e.g., Glover, 1999), we represent the BOLD response to a demand that occurred t seconds ago by a gamma function:

$$H(t) = m(t/s)^a e^{-(t/s)}$$

The parameter m is the magnitude parameter and determines the height of the function, the parameter s is the scale parameter and determines the time scale, and the parameter a is the shape parameter and determines the narrowness of the function. In our work, we have fixed a to be 6, s to be 0.75 sec, and estimate the magnitude to fit the magnitude of response in a brain region.[2] The longer a module is active, the greater the metabolic expenditure, and the larger the BOLD response. This is represented in the height of the individual signals in Fig. 3. The cumulative BOLD response in a particular region is the sum of the individual

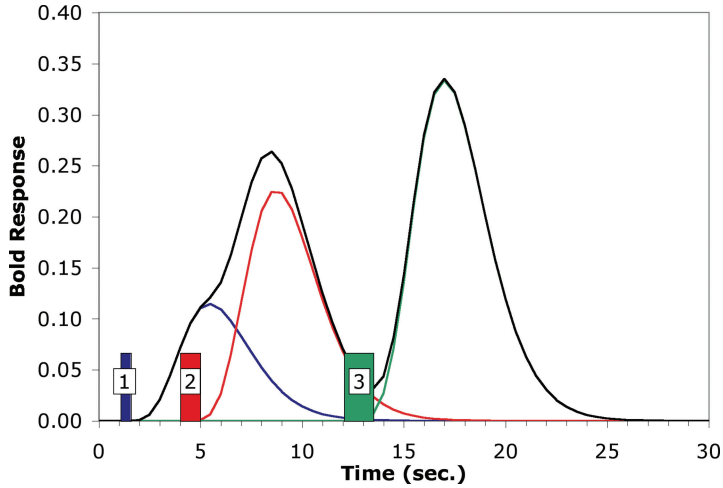


Fig. 3. Illustration of how three separate activities of a single module accumulate to produce a single BOLD response. The individual responses come from gamma functions with shape parameter a set to 6, the time scale parameter$s$ set to 0.75 seconds, and the magnitude parameter m set to 0.5.

BOLD responses driven by a module's activities. This can be calculated by convolving the hemodynamic response, $H(t)$, with a demand function, $D(t)$, which is 1 whenever the module is active and 0 otherwise:

$$B(t) = \int_0^t D(x)H(t-x)dx$$

The dark line in Fig. 3 shows the predicted response. Note that when two module activities are close together their effects basically sum, as in the case of Activities 1 and 2; whereas when they are farther apart they separate, as in the case of Activities 2 and 3.

One complication concerns a module like the goal module, which is only engaged at a moment in time when the subgoal is being changed. In the case of such modules with point activity, we predicted the hemodynamic function at time t by summing the hemodynamic effects of the n points of activity before t:

$$B(t) = \sum_{i=1}^{n} H(t - t_i)$$

where t_i is the time of the ith activity.

A similar convolution methodology is frequently used in software packages for fMRI data analysis that convolve the condition structure of trials in an experiment with a hemodynamic response to produce a condition-sensitive pattern of activity. This pattern is regressed against brain activity to find which regions are sensitive to the conditions (e.g., Friston, 2003). Our application is finer grained (breaking up a single trial in a number of module activities rather than describing a whole trial as a single event) and is used for confirmatory purposes rather than exploratory purposes. Because the computations illustrated in Fig. 3 are implemented as part of the ACT-R simulation program, any ACT-R model implemented by any researcher

automatically makes predictions for the regions associated with the modules. This has brought an unprecedented level of potential constraint to theorizing in ACT-R. The same approach can be used by any theory that makes predictions about the temporal duration of mental activities.

4. Temporal variability: The methodological challenge

The model and participants displayed a wide range of solution times. Fig. 4 shows the distribution of times to the first thumb press for large and small problems for both the participants and the model.[3] All the durations in the model are fixed, except the times for retrievals and the times for imaginal updates. The imaginal update latencies were 0.7 seconds in Fig. 1, but on individual trials they varied uniformly from 0 to 1.5 seconds. The retrieval times are scaled by a latency factor, which was also 0.7 in Fig. 1, but this varied from 0 to 1.2 seconds on individual trials. With these two latency ranges, the model does a good job of matching the latency distributions ($r = .958$).

The challenge is how to discern the predicted patterns in the imaging data when there is such variability in solution times. In order to discern reliable patterns, fMRI data requires aggregating across multiple trials, but how do we aggregate data from a 10-second trial with data from a 30-second trial and compare this to the predictions of a model in which trials are equally variable? This article contrasts two methods for averaging scans from different trials: *onset-locked averaging* and *event-locked averaging*. The data for either averaging method are obtained by taking all the scans for a trial, calculating percentage change of each scan from the baseline defined as the first scan within a respective trial, and linearly detrending the data so that first and last scans are both 0. In this experiment, with sequences of 22 values per trial, the middle 20 values are non-zero. The difference between the two averaging methods turns on how these 20 middle scans are aligned and averaged.

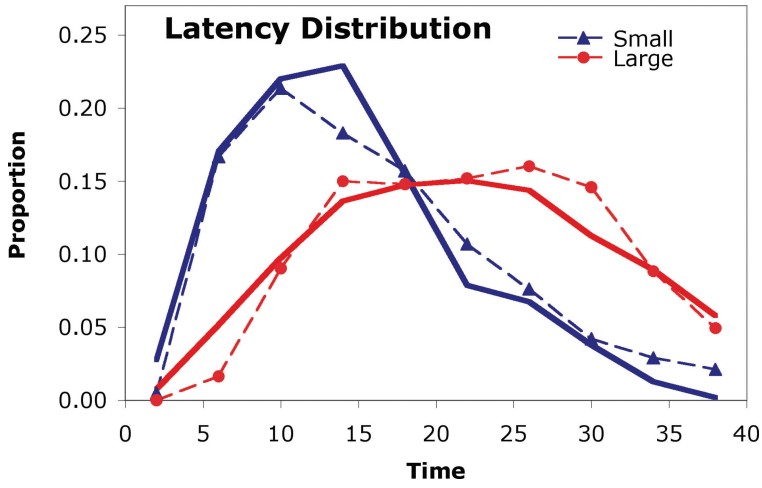


Fig. 4. The distribution times to solve equations in the large-number condition and the small-number condition. Dotted lines connect data, and solid lines are the fits of the ACT-R model (Ravizza, Anderson, & Carter, in press).

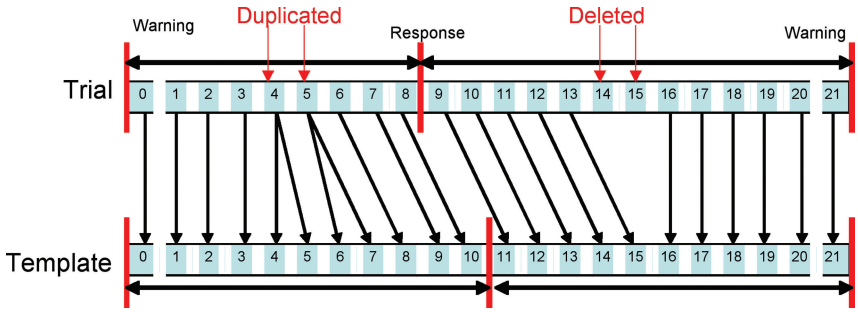


Fig. 5. An illustration of event-locked averaging showing how the scans from a trial are assigned to a template for averaging with other trials. In this case, the time to response is shorter than the average time for the large-number equation condition.

The onset-locked approach just averages the 20 scans independent of when the response occurred. The event-locked analysis attempts to anchor the data around all of the observable events. This experiment has three observable events—presentation of the problem, response generation, and the end of the 22 scans for the trial. These three events define two intervals—one defined by the period of working on the problem and the other defined by the period of inactivity between problems. The event-locked procedure attempts to align scans within an interval according to their position relative to the beginning or end of the interval, whichever is closer. Fig. 5 illustrates how this is done in the case where the first interval is shorter than average and, consequently, the second interval is longer than average. As can be seen, the first interval in the trial has to be expanded to fit the average length, and the second interval has to be shrunk to match the average length.

The following defines the procedure, illustrated in Fig. 5, for taking an interval of n scans and deriving a scan sequence of the mean length m for purposes of averaging:

1. If n is greater than or equal to m, create a sequence of length m by taking $m/2$ scans from the beginning and $m/2$ from the end. If m is odd, select one more from the beginning. This means just deleting the $n - m$ scans in the middle.
2. If n is less than m, create a beginning sequence of length $m/2$ by taking the first $n/2$ scans and duplicating the last scan in this first $n/2$ until the length is $m/2$. Construct backwards from the end similarly. If either n or m is odd, the extra scan is from the beginning.

This creates scan sequences that preserve the temporal structure of the beginning and end of the subsequences and just represent the approximate average activity in their middle. It can be considered a combination of stimulus locking and response locking, which are standard ways of aligning fMRI data. Although the experiment in this article has just two intervals, this method can be applied to experiments that have many more event-defined intervals (e.g., Anderson & Qin, 2008; Rosenberg-Lee, Lovett, & Anderson, submitted).

Fig. 6 illustrates the results of applying onset-locked averaging and event-locked averaging[4] to the six predefined regions corresponding to ACT–R modules. Event-locked averaging does succeed in extracting more of the temporal structure of activity in a region. For instance, in the

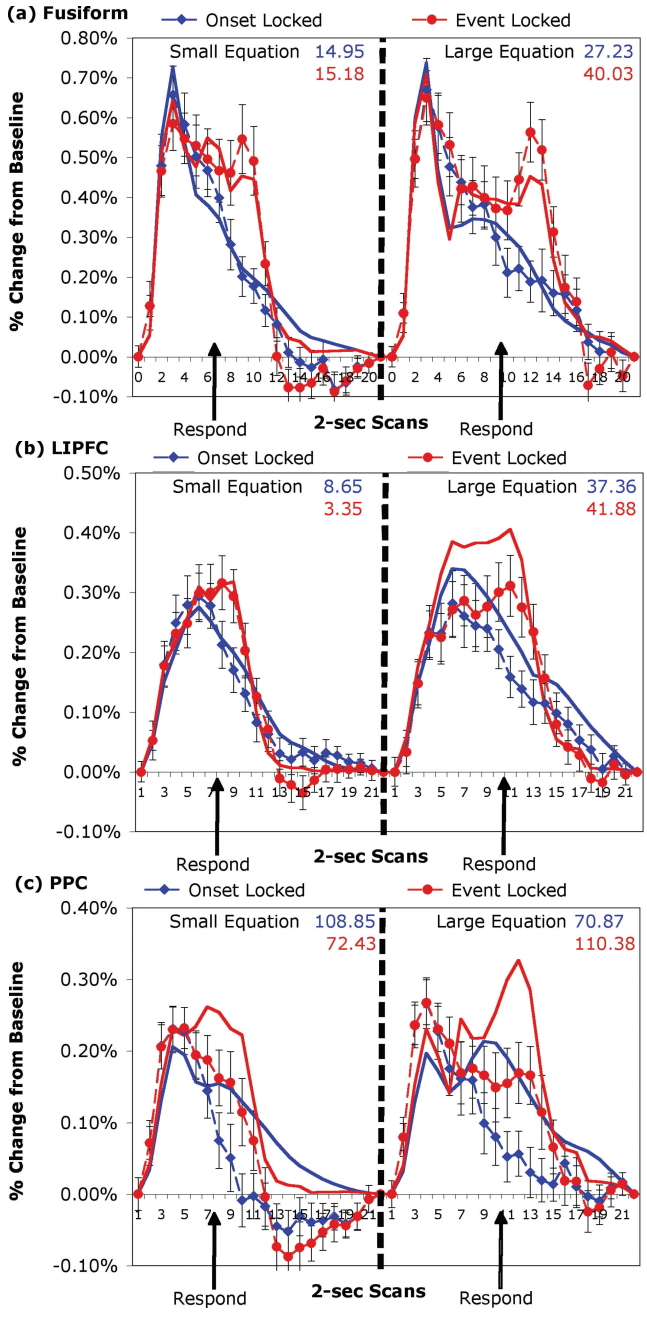


Fig. 6. Mean percentage change in BOLD signal for the six predefined regions of the experiment for which ACT–R makes predictions. These are all left hemisphere regions (see Fig. 2). The mean BOLD responses for onset-locked averaging (blue–darker) and event-locked (red–lighter) averaging are displayed along with their standard errors. The solid curve gives the predictions of the ACT–R model. The numbers in the panels give the SSSEPs for the two types of averaging. LIPFC = lateral inferior prefrontal cortex; PPC = posterior parietal cortex; ACC = anterior cingulated cortex. (*Continued*)

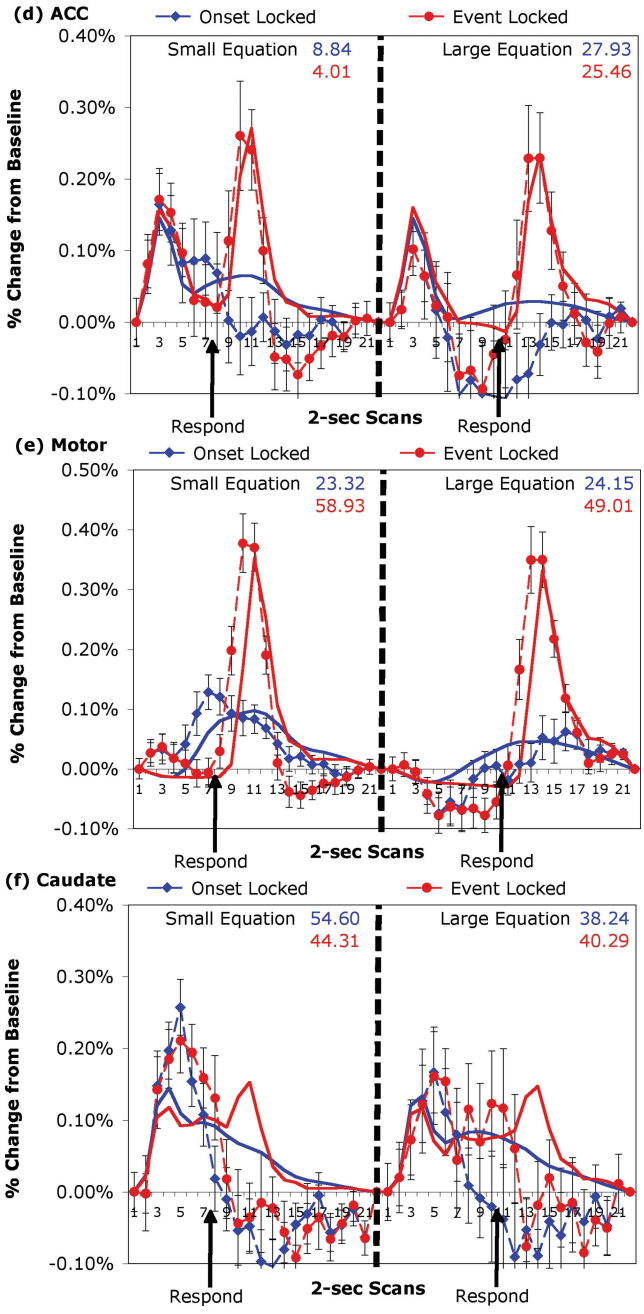


Fig. 6. Mean percentage change in BOLD signal for the six predefined regions of the experiment for which ACT–R makes predictions. These are all left hemisphere regions (see Fig. 2). The mean BOLD responses for onset-locked averaging (blue–darker) and event-locked (red–lighter) averaging are displayed along with their standard errors. The solid curve gives the predictions of the ACT–R model. The numbers in the panels give the SSSEPs for the two types of averaging. LIPFC = lateral inferior prefrontal cortex; PPC = posterior parietal cortex; ACC = anterior cingulated cortex.

anterior cingulate cortex (ACC; Fig. 6d), it identifies responses synched both to problem onset and response delivery, whereas the second rise associated with the response is almost lost in the onset-locked method due to averaging trials with different points of response generation. We can also calculate onset-locked averages or event-locked averages for the trials from the model. Fig. 6 compares the model predictions with the data under the same two averaging schemes. These predictions require estimation of a magnitude parameter for the hemodynamic response. These magnitude parameters were estimated to minimize a measure of deviation that we describe later, called Sum of Squared Standard Errors of Prediction (SSSEP).

5. Assessing the match between model and data

If the model is an accurate description of the task and the module-region mappings are valid, then both the onset-locked averages and the event-locked averages should match between model and data. Although the event-locked averages provide a more meaningful representation of the data and the model predictions, it is not as obvious, which will provide a tougher test of whether model and data correspond.

Table 1 shows two ways previously used (e.g., Anderson, 2005) to assess the match between model and data. Part (a) of the table shows the correlations between the predictions of the model for the various regions and the actual data—either onset-locked averaging or event-locked averaging. The correlations tend to be somewhat weaker for onset-locked because there is less variability in the onset-locked data and, therefore, less signal with which to correlate. The variance of the mean BOLD responses averaged 70% greater in the case of event locking (whereas there is no difference in the standard error of the means). In addition to the issue of differences in variability, it is hard to know how to assess correlations because it is not clear what constitutes a good enough correlation.

Table 1
Two measures to assess correspondence between predictions and data

	Visual Fusiform	Procedural Caudate	Goal ACC	Retrieval LIPFC	Imaginal PPC	Manual Motor	Average
			(a) Correlations				
Onset Locked	0.956	0.633	0.662	0.956	0.725	0.712	0.738
Event Locked	0.958	0.472	0.892	0.969	0.871	0.826	0.806
Average	0.956	0.535	0.752	0.965	0.833	0.814	0.780
			(b) Sum of Squared Standard Errors of Prediction				
Onset Locked	72.18	126.78	37.09	32.13	187.38	58.93	87.59
Event Locked	87.41	110.91	43.15	29.33	130.00	165.60	94.28
Average	79.79	118.85	40.12	30.73	158.69	112.26	92.89

Note. ACC = anterior cingulated cortex; LIPFC = lateral inferior prefrontal cortex; PPC = posterior parietal cortex.

For these reasons, we use a measure of deviation from prediction that is scaled by the error in measurement and is calculated as:

$$\sum_{i=1} \frac{(\bar{X}_i - \hat{X}_i)^2}{S^2_{\bar{X}_i}}$$

where the summation is over the non-zero points along a BOLD curve (20 for each problem type in this experiment). This measure divides the squared deviations between observed means and predicted means by the squared standard error of the means. These SSSEP quantities are reproduced in Part (b) of the table. With 40 non-zero points and a single magnitude parameter estimated to minimize this quantity, the expected value of these sums is 39 under the null hypothesis that the model captures all the systematic variability in the data. Although these sums will vary by chance around their expected value, sums substantially larger than this expected value are a sign of a problem in the match between model and data.

6. Significance tests

To accept or reject the matches between region and modules, one would need to determine a critical value, such that any SSSEP value in Table 1b greater than the critical value would constitute a significant deviation. As each of the terms in the SSSEP sum above approximates a squared normal standard deviate and one parameter is estimated, one might treat these quantities as chi-squares distributed with 39 *df*. However, these sums are not distributed as chi-squares because the errors in measurement of individual points on a BOLD response are not independent. There are strong correlations between error for adjacent scans because of various factors including scanner variability, movement, changes in blood flow, and general oxygen level of the blood (Friston, 2007). In the Appendix, we show that the terms in the sum are correlated as a first-order autoregressive process and show that the correlation for our data is approximately .70.

Kotz and Adams (1964) described an alternative to a chi-square distribution for calculating the critical values for such correlated sums. Kotz and Adams examined the case of a sum of gamma variables (and a chi-square is a special case of a gamma distribution) that have such a correlational structure. They noted that the distribution of a sum of n such correlated gamma variables is closely approximated by another gamma distribution with index α_n and scale factor β_n. In the case where the individual terms are chi-square distributed and have correlation r, these gamma parameters can be calculated as

$$\alpha_n = n/2S(n, r)$$
$$\beta_n = 2 * S(n, r)$$
$$S(n, r) = 1 + \frac{2r}{1 - r}\left(1 - \frac{1 - r^n}{n(1 - r)}\right).$$

Because there are 20 non-zero terms being predicted for each curve in Fig. 6 and the correlation is.7, this implies the SSSEP for a curve should be a gamma with the parameters $\alpha_n = 2.05$ and $\beta_n = 9.78$.

There are a couple of complications to extending the analysis above to assessing the deviations of the predicted means from the observed means. First, it is necessary to assess the simultaneous fit to two sequences of BOLD responses in a region, that for the small-number equations and that for the large-number equations. Because the two sets of squared deviations are independent, one can treat their sums as a gamma distribution with parameters twice the index and the same scale (i.e., $\alpha = 2\alpha_n = 4.10$ and $\beta = \beta_n = 9.78$). A value greater than 77.1 would occur from such a distribution with probability .05 and so one could use this as the critical value for assessing the measures in Table 1b. However, the second complication is that this value does not reflect the fact that a magnitude parameter was estimated to minimize the measure of deviation.

This article takes an alternative approach that deals with the issue of parameter estimation. We estimated the magnitude parameter to minimize the SSSEP for the first 10 non-zero points in the small-number condition. Only the first 10 points are used because the small-number equations tend to show a post-problem undershoot at the end. The hemodynamic function we use cannot produce an undershoot.[5] The undershoot is minimal in the case of the large-number equations because they take longer. A magnitude parameter estimated for all 20 points in the small-number equations would be too small.

The fit to the large-number equations reflects a parameter-free prediction that can be assessed with a gamma distribution with $\alpha = 2.05$ and $\beta_n = 9.78$ (the Kotz & Adams, 1964, approximation). An SSSEP greater than 47.2 would be significant. The SSSEPs for the large equations are reported in the right panel of Fig. 6 and exceed the critical value for the posterior parietal cortex (PPC; Part c) and the motor cortex under event locking (Part e). We also report in the left panels of the figure the SSSEPs for the first 10 points of the small-number equations. Using the same Kotz and Adams formulas, this should be distributed as a gamma variate with $\alpha = 1.20$ and $\beta_n = 8.31$ for which the .05 critical value would be 28.1. However, this choice of a critical value does not take into account the fact that a magnitude parameter was estimated to get the fit. Nonetheless, an SSSEP for any small-number fit exceeding 28.1 clearly reflects a significant deviation. By these measures, there are significant deviations in the model's predictions again for the PPC (Part c) and the motor cortex under event locking (Part e). Also, there are significant deviations of the small-number fit for the caudate (Part f).

7. Evaluation of alternative models

The preceding statistical analysis tests whether we have a "perfect model" of the data by assessing whether the measurement of deviation exceeds a critical value. Although it provides some useful information, this approach has a number of undesirable features. First, testing for critical values rewards noise in the data. If the measurement of the data is noisy, it is unlikely that models with serious problems will reach the critical value. Second, with enough precision in measurement, any model will exceed a critical value because there will always be some small and, perhaps unimportant, deviation. With sufficiently large sample sizes (i.e., precision

of measurement) this approach will always reject the ACT-R model or any model because some tiny detail is sure to be wrong. On this view, it might be more worthwhile to compare ACT-R to alternative models and ask which does the best job of accounting for the data. As we noted, there have been numerous efforts to compare models against imaging data, but they have tended not to deal with the time courses of the BOLD responses observed, and none deal with this specific task. Therefore, comparing ACT-R to alternative models is problematic. Instead, we tried alternate mappings of the modules onto brain regions in order to provide a comparison with the mapping of the ACT-R model.

This analysis uses the combined SSSEP for the small-number and large-number conditions as a measure of how well a module mapping predicted a brain region. As noted earlier, under the null hypothesis, this sum should be distributed as a gamma with $\alpha = 4.10$ and $\beta_n = 9.78$. Table 2 gives sums of deviations for various models. The first six rows in Table 2 give the fit of each module to each region. The seventh row gives SSSEPs for the best linear combination of all six modules.[6] The eighth row gives the value of the sum for the pair of modules that gave the best fit. These best pairs of regions were the same for either the onset-locked averages or the event-locked averages. The weightings of the best-fitting pairs for the event-locked averages are:

1. Fusiform: .070Visual + .002Retrieval.
2. LIPFC: .012Visual + .004Retrieval.
3. PPC: .035Visual − .005Manual.

Table 2
SSSEPs for each different mapping of modules onto brain region

	Fusiform	LIPFC	PPC	ACC	Motor	Caudate
(a) Onset-Locked						
1. Visual	72.18	75.65	100.13	46.96	118.56	111.23
2. Retrieval	158.31	32.13	192.61	68.45	111.68	135.37
3. Imaginal	167.66	53.03	197.56	65.11	109.07	139.08
4. Goal	362.99	271.12	220.88	37.09	106.77	146.76
5. Manual	820.93	460.23	507.72	74.60	58.93	145.75
6. Procedural	96.86	57.98	136.89	50.20	110.19	126.79
7. All	30.32	30.32	30.32	30.32	37.91	35.36
8. Best Pair	59.33	30.30	46.43	26.83	58.90	68.82
(b) Event-Locked						
1. Visual	87.41	100.00	100.74	80.53	448.93	82.81
2. Retrieval	210.83	29.33	130.27	116.71	463.00	72.79
3. Imaginal	134.66	32.22	123.14	95.77	449.39	83.34
4. Goal	517.12	483.48	390.16	43.15	321.74	149.41
5. Manual	1055.59	683.88	669.96	120.79	165.60	150.22
6. Procedural	138.04	185.15	180.08	64.97	403.46	110.91
7. All	56.04	30.32	44.36	35.81	93.44	45.99
8. Best Pair	74.85	30.34	70.36	42.89	159.58	49.03

Note. LIPFC = lateral inferior prefrontal cortex; PPC = posterior parietal cortex; ACC = anterior cingulated cortex.

4. ACC: .024Goal − .001Manual.
5. Motor: .0003Imaginal + .013Manual.
6. Caudate: .00003Visual + .002Retrieval.

Of course, the weighted sum of multiple modules will do better than any single module in the sum. These different mappings can be compared using the Bayesian Information Criterion (BIC) for statistical inference (Raftery, 1995; Wagenmakers, 2007). The BIC statistic for a model is defined as:

$$BIC = -2\ln(L) + k\ln(n),$$

where L is the likelihood of the data under the model, k is the number of parameters being estimated, and n is the number of observations. Smaller BIC values represent more plausible models. The BIC score will penalize models that use more than one module because they must estimate a weight parameter for each module. As Wagenmakers described, the Bayes Factor (BF) can be approximately calculated from the difference, ΔBIC, in the BIC scores for two models:

$$BF = e^{\Delta BIC/2}$$

The Bayes Factor can be interpreted as how many times more likely the model is with the smaller BIC score.

Table 3
BIC values for the different models of brain regions with best models for a region in italics

	Fusiform	LIPFC	PPC	ACC	Motor	Caudate
(a) Onset-Locked						
1. Visual	*15.25*	15.75	19.59	12.17	22.74	21.47
2. Retrieval	30.00	*11.15*	36.60	14.73	21.54	25.75
3. Imaginal	31.78	12.80	37.57	14.28	21.10	26.42
4. Goal	71.48	52.36	42.19	*11.39*	20.70	27.84
5. Manual	170.74	92.16	102.37	15.60	*13.49*	27.65
6. Procedural	19.05	13.38	26.02	12.50	21.29	24.20
7. All	29.54	29.54	29.54	29.54	29.88	29.73
8. Best Pair	17.23	14.78	*15.81*	14.75	17.18	*18.47*
(b) Event-Locked						
1. Visual	*17.53*	19.57	19.69	16.47	89.74	16.82
2. Retrieval	40.19	*11.07*	24.82	22.42	92.75	*15.34*
3. Imaginal	25.62	11.15	23.55	18.87	89.84	16.90
4. Goal	104.40	97.15	77.22	*11.83*	62.83	28.33
5. Manual	222.63	140.65	137.61	23.13	*31.38*	28.48
6. Procedural	26.23	35.14	34.16	14.26	80.05	21.41
7. All	31.59	29.54	30.38	29.76	36.94	30.53
8. Best Pair	19.32	14.78	*18.68*	15.50	33.93	16.06

Note. LIPFC = lateral inferior prefrontal cortex; PPC = posterior parietal cortex; ACC = anterior cingulated cortex.

Table 3 presents the BIC values for the various models whose sums of deviations are presented in Table 2—remember smaller values indicate better models. Models that involve a weighted average of all modules (see line 7 of Table 3) suffer a 5ln(40) penalty relative to the single-module models because they involve five more parameters. Clearly, the data never justify such a complex model. The best pair suffers a ln(40) penalty because it involves an extra parameter (see line 8). In some cases, the best pair model is better than any single module.

One might argue that the best-pair model should suffer an additional $2\ln(15) = 5.42$ penalty because one is just cherry-picking the best combination of the 15 possible pairs and, therefore, any pair should have prior probability of 1/15. One might also argue that, if one is just looking for the single module that offers the best competition to the ACT-R module in fitting a region, it ought to suffer a $2\ln(5) = 3.22$ penalty because there are five other modules to choose from and the prior probability that any will be the best is one fifth (see the discussion of hierarchical Bayesian modeling in Shiffrin, Lee, Wagenmakers, & Kim, 2008) Whether one adds on such penalties or not, it does not change the conclusion from Table 3 that the ACT–R modules offer the best models for four of the regions—visual-for-fusiform, retrieval-for-LIPFC, goal-for-ACC, and manual-for-motor. It also does not impact the conclusion that the ACT–R modules do not provide good models for the other two regions (PPC and caudate).

The BIC values in Part (a) of Table 3 for the onset-locked averages have an average value of 29.8 ($SD = 27.9$). The BIC values in Part (b) for the event-locked averages have an average of 42.8 ($SD = 42.2$). In 37 cases, the BIC value was larger in Part (b) by an average amount of 18.61, whereas in only 11 cases was the BIC value larger in Part (a) by an average of 5.77. This supports the conclusion that event-locked averages both produce a representation of the data that is a stronger test of a model and offers greater discrimination among models.

8. General discussion

The article has focused more on methodology than substantive conclusions about the ACT–R modules and their associated brain regions. However, such methodology is only useful to the extent that it yields substantive conclusions and so we review below what has been learned on the substantive issues. There were three ACT-R modules that provided best fits to their associated regions with no significant deviations:

1. Visual module and the fusiform gyrus: Many brain regions are involved in visual processing, and many of these show activity related to the task structure of experiments. For instance, Anderson and Qin (2008) found temporal-occipital regions that became active whenever something new appeared on the screen. Rosenberg-Lee et al. found regions in the left and right lingual gyrus that were differentially active depending on whether the bulk of the problem was presented in the left or right visual field. However, these other regions are not sensitive to the goals of the participant but respond when some change occurs in the visual field. In contrast, the fusiform responds during periods of time where the visual stimulus does not change but the participant is busy accessing information from the visual field on an as-needed basis. The ACT-R visual module encodes attended

information in the visual field and shows this sensitivity to processing goals.[7] Thus, its predictions match the BOLD response in this region and not the other regions.

2. Retrieval module and the LIPFC: A great deal of research has found that the LIPFC is involved in retrieval (e.g., Buckner, Kelley, & Petersen, 1999; Cabeza, Dolcos, Graham, & Nyberg, 2002; Fletcher & Henson, 2001; Wagner et al., 2001), but there are different views on its exact function. Thompson-Schill and colleagues (Thompson-Schill, D'Esposito, Aguirre, & Farah, 1997; Thompson-Schill, D'Esposito, & Kan, 1999; Thompson-Schill et al., 1998) provided evidence in support of the *selection hypothesis*. According to the selection hypothesis, this region is not involved in the retrieval attempt *per se,* but rather the selection of information among competing alternatives after retrieval (e.g., OPEN vs. CLOSE). On the other hand, Wagner and colleagues (Badre & Wagner, 2002; Wagner et al., 2001; and later Martin & Cheng, 2006) argued that Thompson-Schill et al.'s (1997) data could be explained in terms of associative strength rather than selection (the *controlled retrieval hypothesis*). According to Wagner and colleagues Badre & Wagner, 2002; Wagner et al., 2001; and later Martin & Cheng, 2006), the LIPFC was responsible for controlled retrieval of information. Selection among competing alternatives is just one factor that influences this process. The ACT-R position can be seen as a version of the controlled retrieval hypothesis. The experiment does not discriminate between the two hypotheses. Other, more discriminative experiments from our laboratory have shown that this region does conform to ACT–R predictions about fan or associative interference (Sohn, Goode, Stenger, Carter, & Anderson, 2003, 2005), retention delay (Anderson et al., 2008), and repetition (Danker, Gunn, & Anderson, in press). Also, the Ravizza et al. experiment did find that the BOLD response was weaker on trials where errors occurred. This may be interpreted as demonstrating that errors increase when weaker signals drive the retrieval process.
3. Goal module and the ACC: There is a diverse set of views about the function of the ACC. Some have postulated that it is involved in controlling cognition, consistent with the ACT-R view that ties it to the goal module. For instance, Posner and Dehaene (1994) described the ACC as "involved in the attentional recruitment and control of brain areas to perform complex tasks" (p. 76). D'Esposito et al. (1995) have identified it with Baddeley's (1986) central executive, and Posner and DiGirolamo (1998) have related it to Norman and Shallice's (1986) Supervisory Activating System. Another theory relates it to error detection. This is supported by the error-related negativity (ERN) in event-related potentias that has been observed when errors are made in speeded response tasks (e.g., Falkenstein, Hohnbein, & Hoorman, 1995; Gehring, Goss, Coles, Meyer, & Donchin, 1993). On the other hand, this area responds more strongly in many tasks that do not involve errors. It has been argued (Botvinick, Braver, Carter, Barch, & Cohen, 2001; Carter et al., 2000; MacDonald, Cohen, Stenger, & Carter, 2000) that the ACC activity reflects response conflict and that error trials are just a special case of this. In a variation on the response conflict view, Brown and Braver (2005) suggested that the ACC might contribute to preparation and control simply by reflecting learned task-specific error likelihoods, which can be used by other regions to adjust control. The results from this experiment are not consistent with simplistic interpretations of any of these theories. A BOLD response occurs in the ACC on trials without errors and well in advance of any

motor response. Moreover, the variation in levels of ACC activation within single trials indicates that it cannot just be responding to the overall error rate for that trial. Finally, the patterns of activation emerge well in advance of any response, and so they cannot reflect simple response conflict. On the other hand, these results are consistent with a more general conflict view that assumes conflict occurs whenever there is ambiguity about what to do next, whether that next thing is a response or a choice for a path of internal cognition. Subgoals in ACT-R serve the function of determining which direction cognition takes at these points of ambiguity. Different regions of the ACC appear to reflect response conflict versus internal conflict. For instance, Van Veen and Carter (2005) found evidence that response conflict was handled by a more anterior and ventral region rather than a region that responded to what they called *semantic conflict*. The center of their semantic conflict region is quite close to the ACT-R ACC region (indeed, the ACT-R region is slightly further posterior than their semantic conflict region).

The relative success of the ACT-R model in addressing the patterns of the BOLD responses in these three regions shows the potential of a cognitive architecture like ACT-R to bring some order to the extremely rich and sometimes confusing literature of imaging results. However, the module predictions did not work out for other regions, and these failures of prediction can be informative:

1. Manual module and the motor cortex: Although the manual module clearly gave the best fit to the motor module, there were significant deviations between prediction and data. The BOLD response rises too early in the motor region, reflecting activity 2 seconds prior to movement execution. This suggests anticipatory motor activity—something found in some of our other experiments (e.g., Anderson & Qin, 2008) when trials are relatively long. It is worth noting that this is a case where event-locked averaging uncovers a discrepancy largely missed by the onset-locked averaging.
2. Imaginal module and the PPC: There is considerable evidence that the PPC serves a representational role like that of ACT–R's imaginal module. Research has found that this region is engaged in verbal encoding (Clark & Wagner, 2003; Davachi, Maril, & Wagner, 2001), mental rotation (Alivisatos & Petrides, 1997; Carpenter, Just, Keller, Eddy, & Thulborn, 1999; Heil, 2002; Richter, Ugurbil, Georgopoulos, & Kim, 1997; Zacks, Ollinger, Sheridan, & Tversky, 2002), and visual-spatial strategies in a variety of contexts (Dehaene et al., 2003; Reichle, Carpenter, & Just, 2000; Sohn et al., 2004). Nonetheless, the imaginal module misfits the PPC activity in two places. It slightly under-predicts the response at problem onset and, more seriously, it over-predicts the magnitude of response associated with response generation. As illustrated in Fig. 1, the ACT-R model distributed its encoding of the equation over the problem only taking in information as needed. If the model had encoded more of the information at the onset, there would have been a larger spike early in the trial and less activation later. Just as the motor data suggested that the model include anticipatory motor programming, this may be another instance where the imaging data suggest how to revise the ACT-R model.
3. Procedural module and the caudate: There is evidence (e.g., Amos, 2000; Ashby & Waldron, 2000; Frank, Loughry, & O'Reilly, 2001; Poldrack, Prabakharan, Seger, & Gabrieli, 1999; Wise, Murray, & Gerfen, 1996) that the basal ganglia play a role in

action selection, and the caudate is a region of the basal ganglia that has been targeted in a number of imaging studies. Both the small- and large-number equations show an early peak that is not predicted. We have found this early peak in caudate activity in a number of experiments that stretched out over significant periods of time (e.g., Anderson & Qin, 2008; Kao, Douglass, Fincham, & Anderson; Rosenberg-Lee et al., in press). It is not something we were able to detect in experiments with shorter trials (e.g., Anderson, 2005; Anderson et al., 2007). A number of issues are intertwined in understanding the caudate misfits. One has to do with the proper understanding of baseline. Although we do not model it, participants are probably still engaged in thought during the rest periods and so the procedural module would not be inactive. The BOLD response should rise from baseline defined by the rest period only to the degree that production firing is greater during task performance.[8] A second complication is that the right measure of effort might not be number of productions firing but rather the amount of work the productions do in transferring information among modules. Stocco and Anderson (2008) developed a successful model of caudate activity in an equation-solving task using this more complex measure. Another issue is the assumption that caudate only reflects procedural activity. This single function assumption is almost certainly wrong. This region of the caudate is known to also reflect reward activities (e.g., Breiter, Aharon, Kahneman, Dale, & Shizgal, 2001; Delgado, Locke, Stenger, & Fiez, 2003) and to respond to eye movements (e.g., Gerardin et al., 2003). The initial spikes that are causing the misfit in this experiment may reflect the increased work that the initial productions do, they may reflect the reward-related activity, or they may reflect eye movements. Increased caudate firing at episode boundaries has also been found in rats (Barnes, Kubota, Hu, Jin, & Graybiel, 2005) and in monkeys (Fujii & Graybiel, 2005). Whatever is involved, it seems quite ubiquitous and may reflect important aspects of the overall organization of cognition and behavior.

8.1. Theoretical methodology

Although the results do offer some insight on the function of different brain regions, this article has been mainly about the methodology that can lead to these insights. It has focused on three methodological developments that have enabled fMRI data to be used to evaluate models of complex problem solving:

1. Event-locked averaging has been developed for dealing with the temporal variability in complex problem solving and revealing the structure of the BOLD response in the imaging data. This offers a much more meaningful representation of the data and a substantially more demanding test of any model's behavior. It can be applied in cases with many event-defined intervals.
2. A methodology has been developed for using an information-processing theory like ACT-R to generate predictions for activity in specific brain regions. Although we have used ACT-R, this same methodology can be used to relate many information-processing theories to fMRI data. The methodology provides a way of taking the typical assumption

of such models, that various components are working at different points in time, and mapping these demand functions onto predictions for the BOLD response.
3. A statistical method has been developed for assessing model fits that deal with the issue of non-independence of scans.

The results were not uniformly positive for the ACT–R model presented in this article, and one should also be prepared for such harsh judgments about other models. Such negative results demonstrate that particular models are, in fact, falsifiable using the current methodological approach. If one's goal is to improve a cognitive architecture and the models developed within it, such negative results are to be welcomed, as they point to directions for improvement.

Notes

1. This model can be downloaded from the models link at the ACT-R Web site (act-r.psy.cmu.edu) under the title of this article.
2. The setting of the parameters *a* and *s* comes from another experiment (Kao, Douglass, Fincham, & Anderson, in press) where participants pressed a finger to a visual signal. The parameters were set to fit the response in the motor area. Handwerker, Ollinger, and D'Esposito (2004) presented data indicating that the variability in the BOLD response across regions is small relative to the variability across participants.
3. The data in this and other figures are based on 618 observations for small-number equations and 487 observations for large-number equations collected over 16 participants, whereas the model numbers are based on 10,000 Monte Carlo trials for both small- and large-number equations.
4. We have considered a number of other ways to "warp" the data besides event locking. However, to provide focus to this article, we have chosen to just compare event locking with onset locking.
5. A negative undershoot at the end of the BOLD response has been frequently observed, and it is sometimes modeled as the difference of two gammas (e.g., Friston et al., 1998; Glover, 1999), but we have found that this does not extend well to modeling long processes like in these experiments; therefore, we consider proper treatment of the undershoot still an open issue.
6. Many of these have the value 30.32 because this is the value for which the gamma has maximum density.
7. This is not to deny that there is attentional modulation in other visual areas.
8. However, we should note the caudate is not typically found as part of the "default network" in resting state studies (e.g., Raichle & Snyder, 2007).

Acknowledgments

This research was supported by National Science Foundation Award REC-0087396. We would like to thank Jennifer Ferris for her comments on the article.

References

Alivisatos, B., & Petrides, M. (1997). Functional activation of the human brain during mental rotation. *Neuropsychologia, 35*, 111–118.

Amos, A. (2000). A computational model of information processing in the frontal cortex and basal ganglia. *Journal of Cognitive Neuroscience, 12*, 505–519.

Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science, 29,* 313–342.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of mind. *Psychological Review, 111*, 1036–1060.

Anderson, J. R., Byrne, D., Fincham, J. M., & Gunn, P. (2008). Role of prefrontal and parietal cortices in associative learning. *Cerebral Cortex, 18*, 904–914.

Anderson, J. R., & Qin, Y. (2008). Using brain imaging to extract the structure of complex events at the rational time band. *Journal of Cognitive Neuroscience*, 1624–1636.

Anderson, J. R., Qin, Y., Jung, K.-J., & Carter, C. S. (2007). Information-processing modules and their relative modality specificity. *Cognitive Psychology, 54,* 185–217.

Ashby, F. G., & Waldron, E. M. (2000). The neuropsychological bases of category learning. *Current Directions in Psychological Science, 9*, 10–14.

Baddeley, A. D. (1986). *Working memory*. New York: Oxford University Press.

Badre, D., &Wagner, A. D. (2002). Semantic retrieval, mnemonic control, and prefrontal cortex. *Behavioral and Cognitive Neuroscience Reviews, 1*, 206–218.

Barnes, T., Kubota, Y., Hu, D., Jin, D. Z., & Graybiel, A. M. (2005). Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. *Nature, 437*, 1158–1161.

Botvinick, M. M., Braver, T. S., Carter, C. S., Barch, D. M., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*, 624–652.

Breiter, H. C., Aharon, I., Kahneman, D., Dale, A., & Shizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron, 30*, 619–639.

Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science, 307,* 1118–1121.

Buckner, R. L., Kelley, W. M., & Petersen, S. E. (1999). Frontal cortex contributes to human memory formation. *Nature Neuroscience, 2,* 311–314.

Bullmore, E., Brammer, M., Williams, S., Rabe-Hesketh, S., Janot, N., David, A., et al. (1996). Statistical methods of estimation and inference for functional MR image analysis. *Magnetic Resonance in Medicine, 35,* 261–277.

Cabeza, R., Dolcos, F., Graham, R., & Nyberg, L. (2002). Similarities and differences in the neural correlates of episodic memory retrieval and working memory. *Neuroimage, 16,* 317–330.

Carpenter, P. A., Just, M. A., Keller, T. A., Eddy, W., & Thulborn, K. (1999). Graded function activation in the visuospatial system with the amount of task demand. *Journal of Cognitive Neuroscience, 11,* 9–24.

Carter C. S., MacDonald A. M., Botvinick, M., Ross, L. L., Stenger, V. A., Noll, D., et al. (2000). Parsing executive processes: Strategic versus evaluative functions of the anterior cingulate cortex. *Proceedings of the National Academy of Sciences USA, 97,* 1944–1948.

Clark, D., & Wagner, A. D. (2003). Assembling and encoding word representations: fMRI subsequent memory effects implicate a role for phonological control. *Neuropsychologia, 41,* 304–317.

Danker, J. F., Gunn, P., & Anderson, J. R. (in press). A rational analysis of memory predicts left prefrontal activation during controlled retrieval. *Cerebral Cortex.*

Davachi, L., Maril, A., & Wagner, A. D. (2001). When keeping in mind supports later bringing to mind: Neural markers of phonological rehearsal predict subsequent remembering. *Journal of Cognitive Neuroscience, 13,* 1059–1070.

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology, 20,* 487–506.

Delgado, M. R., Locke, H. M., Stenger, V. A., & Fiez, J. A. (2003). Dorsal striatum responses to reward and punishment: Effects of valence and magnitude manipulations. *Cognitive, Affective, & Behavioral Neuroscience, 3,* 27–38.

D'Esposito, M., Piazza, M., Detre, J. A., Alsop, D. C., Shin, R. K., Atlas, S., et al. (1995). The neural basis of the central executive of working memory. *Nature, 378,* 279–281.

Falkenstein, M., Hohnbein, J., & Hoorman, J. (1995). Event related potential correlates of errors in reaction tasks. In G. Karmos, M. Molnar, V. Csepe, I. Czigler, & J. E. Desmedt (Eds.), *Perspectives of event-locked potentials research* (pp. 287–296). Amsterdam: Elsevier.

Fletcher, P. C., & Henson, R. N. A. (2001). Frontal lobes and human memory: Insights from functional neuroimaging. *Brain, 124*, 849–881.

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between the frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, & Behavioral Neuroscience, 1*, 137–160.

Friston, K. J. (2003). Introduction: Experimental design and statistical parametric mapping. In R. S. J. Frackowiak, K. J. Friston, C. Frith, R. Dolan, K. J. Friston, C. J. Price, et al. (Eds.), *Human brain function* (2nd ed., pp. 599–633) San Diego, CA: Academic.

Friston, K. J., Ashburner, J. T., Kiebel, S. J., Nichols, T. E., & Penny, W. D. (Eds.). (2007). *Statistical parametric mapping: The analysis of functional brain images*. San Diego, CA: Academic.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M., & Turner, R. (1998). Event-locked fMRI: Characterizing differential responses. *Neuroimage, 7*, 30–40.

Fujii, N., & Graybiel, A. M. (2005). Time-varying covariance of neural activities recorded in striatum and frontal cortex as monkeys perform sequential-saccade tasks. *Proceedings of the National Academy of Sciences USA, 102*, 9032–9037.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science, 4*, 385–390.

Gerardin, E., Lehericy, S., Pochon, J. B., Tezenas du Montcel, S., Mangin, J. F., Poupon, F., et al. (2003). Foot, hand, face and eye representation in the human striatum. *Cerebral Cortex, 13*, 162–169.

Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *Neuroimage, 9*, 416–429.

Handwerker, D. A., Ollinger, J. M., & D'Esposito, M. (2004). Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage, 21*, 1639–1651.

Heil, M. (2002). Early career award: The functional significance of ERP effects during mental rotation. *Psychophysiology, 39*, 535–545.

Just, M. A., & Varma, S. (2007). The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition. *Cognitive, Affective, & Behavioral Neuroscience, 7*, 153–191.

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience, 10*, 1625–1633.

Kao, Y., Douglass, S., Fincham, J., & Anderson, J. R. (in press). Traveling the second bridge: Using fMRI to assess an ACT–R model of geometry proof. *Cognitive Science*. Manuscript in preparation.

Kotz, S., & Adams, J. W. (1964). Distribution of sum of identically distributed exponentially correlated gamma-variables. *The Annals of Mathematical Statistics, 35*, 277–283.

Lange, J., & Lappe, M. (2006). A model of biological motion perception from configural form cues. *Journal of Neuroscience, 26*, 2894–2906.

MacDonald A. W., Cohen J. D., Stenger V. A., & Carter C. S. (2000). Dissociating the role of dorsolateral prefrontal cortex and anterior cingulate cortex in cognitive control. *Science, 288*, 1835–1837.

Martin, R. C., & Cheng, Y. (2006). Selection demands versus association strength in the verb generation task. *Psychonomic Bulletin & Review, 13*, 396–401.

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance. Part 1. Basic mechanisms. *Psychological Review, 104*, 2–65.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. Davidson, G. E. Schwartz, and D. Shapiro (Eds.), *Consciousness and self regulation: Advances in research and theory* (Vol. 4, pp. 1–18). New York: Plenum.

Poldrack, R. A., Prabakharan, V., Seger, C., & Gabrieli, J. D. E. (1999). Striatal activation during cognitive skill learning. *Neuropsychology, 13*, 564–574.

Posner, M. I., & Dehaene, S. (1994). Attentional networks. *Trends in Neurosciences, 17*, 75–79.

Posner, M. I., & DiGirolamo, G. J. (1998). Executive attention: Conflict, target detection and cognitive control. In R. Parasuraman (Ed.), *The attentive brain* (pp. 401–423). Cambridge, MA: MIT Press.

Purdon, P., & Weisskoff, R. (1998). Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates. *Human Brain Mapping, 6*, 239– 249.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology 1995* (pp. 111–196). Cambridge, MA: Blackwell.

Raichle, M. E., & Snyder, A. Z. (2007). A default mode of brain function: A brief history of an evolving idea. *Neuroimage, 37*, 1083–1090.

Ravizza, S. M., Anderson, J. R., & Carter, C. S. (in press). *Errors of mathematical processing: The relationship of accuracy to neural regions associated with retrieval or representation of the problem state*. Manuscript submitted for publication.

Reichle, E. D., Carpenter, P. A., & Just, M. A. (2000). The neural basis of strategy and skill in sentence–picture verification. *Cognitive Psychology, 40*, 261–295.

Richter, W., Ugurbil, K., Georgopoulos, A., & Kim, S.-G. (1997). Time-resolved fMRI of mental rotation. *Neuroreport, 8*, 3697–3702.

Rosenberg-Lee, M., Lovett, M., & Anderson, J. R. (in press). *Neural correlates of arithmetic calculation strategies*. Manuscript submitted for publication.

Shiffrin, R. M., Lee, M. D., Wagenmakers, E. J., & Kim, W. J. (2008). Model evaluation and selection: Established methods and recent developments. *Cognitive Science, 32*.

Sohn, M.-H., Goode, A., Koedinger, K. R., Stenger, V. A, Carter, C. S., & Anderson, J. R. (2004). Behavioral equivalence does not necessarily imply neural equivalence: Evidence in mathematical problem solving. *Nature Neuroscience, 7*, 1193–1194.

Sohn, M.-H., Goode, A., Stenger, V. A., Carter, C. S., & Anderson, J. R. (2003). Competition and representation during memory retrieval: Roles of prefrontal cortex and posterior parietal cortex. *Proceedings of the National Academy of Sciences USA, 100*, 7412–7417.

Sohn, M.-H., Goode, A., Stenger, V. A., Carter, C. S., & Anderson, J. R. (2005). An information-processing model of three cortical regions: Evidence in episodic memory retrieval. *Neuroimage, 25*, 21–33.

Stocco, A., & Anderson, J. R. (2008). Endogenous control and task representation: An fMRI study in algebraic problem solving. *Journal of Cognitive Neuroscience, 20,* 1300–1314.

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proceedings of the National Academy of Sciences USA, 94*, 14792–14797.

Thompson-Schill, S. L., D'Esposito, M., & Kan, I. P. (1999). Effects of repetition and competition on activity in left prefrontal cortex during word generation. *Neuron, 23*, 513–522.

Thompson-Schill, S. L., Swick, D., Farah, M. J., D'Esposito, M., Kan, I. P., & Knight, R. T. (1998). Verb generation in patients with focal frontal lesions: A neuropsychological test of neuroimaging findings. *Proceedings of the National Academy of Sciences USA, 95*, 15855–15860.

Van Veen, V., & Carter, C. S. (2005). Separating semantic conflict and response conflict in the Stroop task: A functional MRI study. *Neuroimage, 27*, 497–504.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review, 14*, 779–804.

Wagner, A. D., Maril, A., Bjork, R. A., & Schacter, D. L. (2001). Prefrontal contributions to executive control: fMRI evidence for functional distinctions within lateral prefrontal cortex. *Neuroimage, 14*, 1337–1347.

Wagner, A. D., Paré-Blagoev, E. J., Clark, J., & Poldrack, R. A. (2001). Recovering meaning: Left prefrontal cortex guides controlled semantic retrieval. *Neuron, 31*, 329–338.

Wise, S. P., Murray, E. A., & Gerfen, C. R. (1996). The frontal cortex-basal ganglia system in primates. *Critical Reviews in Neurobiology, 10*, 317–356.

Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., et al. (2002). A general statistical analysis for fMRI data. *Neuroimage, 15*, 1–15.

Zacks, J. M., Ollinger, J. M., Sheridan, M. A., & Tversky, B. (2002). A parametric study of mental spatial transformation of bodies. *Neuroimage, 16*, 857–872.

Appendix: Correlation in errors of measurement among scans

There are strong correlations between errors of measurement for adjacent scans because of various factors including scanner variability, movement, changes in blood flow, and general oxygen level of the blood. The most direct way to examine this issue is to look at the individual participant deviations from means for a scan:

$$Dev_{i,j} = \bar{B}_{i,j} - \bar{B}_j,$$

where the first term is the mean BOLD response of participant i for scan position j, and the second term is the mean for scan position j over all participants. Fig. 7 plots the average correlations r_n between these deviations for points that were n scans apart. As Friston, Ashburner, Kiebel, Nichols, and Penny (2007) noted, these correlations only differed a little by region. Fig. 7 shows how the average correlation over regions varies with n—the correlations are greater than 0.8 for Lag 1 and go down to about 0.3 for Lag 10. These lagged correlations do not differ substantially for the onset-locked averages versus the event-locked averages. Moreover, the fitted curve in the figure reveals that these correlations decay as a exponential function of lag where

$$r_n = r^n.$$

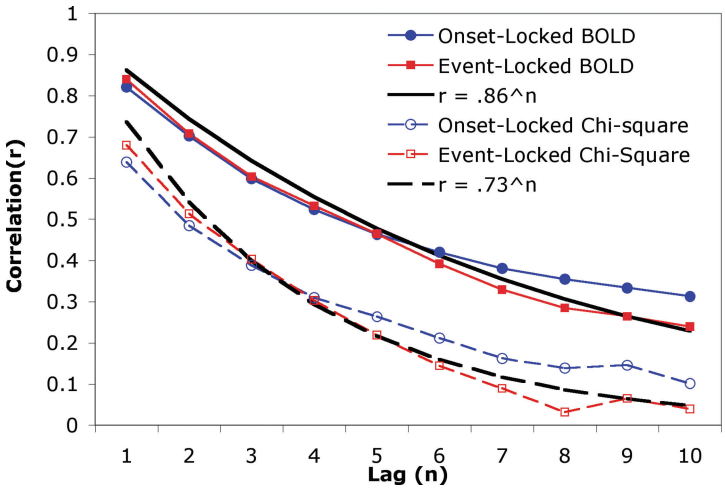


Fig. 7. Correlation in the deviations of participant means from overall means and the resulting correlation in the SSSEP (chi-square) statistics.

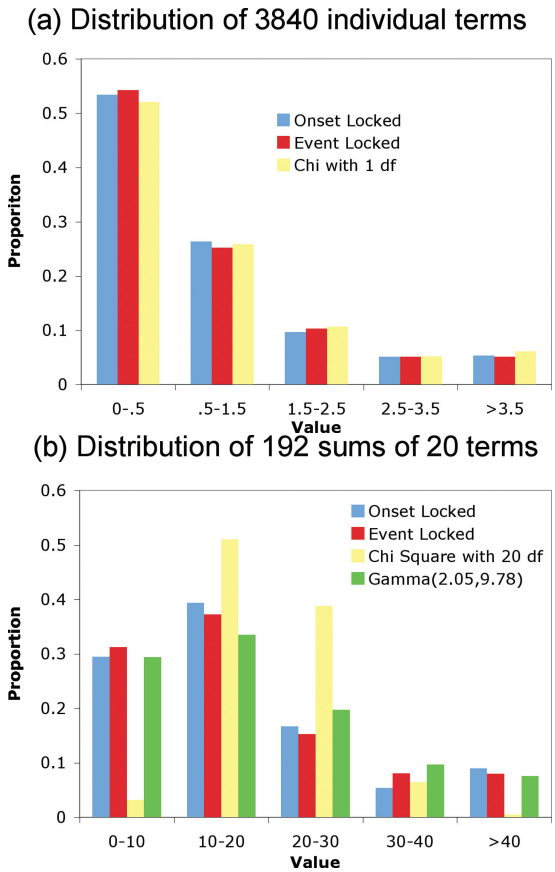


Fig. 8. Unit area histograms displaying the proportions of squared deviations with different values. Part (a) compares the empirical distribution of single squared deviations (3,840 observations = 16 subjects × 6 regions × 2 curves × 20 points) with the predictions from a chi-square distribution with 1 *df*. Part (b) compares the empirical distribution of sums of 20 squared deviations (192 observations = 16 subjects × 6 regions × 2 curves) with the predictions from a chi-square distribution with 20 *df* and a gamma distribution with $\alpha_n = 2.05$ and $\beta_n = 9.78$.

This is a sign of a first-order autocorrelative process that has been identified in fMRI data (e.g., Bullmore et al., 1996; Purdon & Weisskoff, 1998; Worsley et al., 2002).

The autocorrelation in the errors of measurement along a BOLD response curve leads to a correlation between the individual squared standard deviation terms that go into the SSSEP that we have been using to measure deviations in module predictions. The magnitude of these correlations can be assessed in a model-free way by using the average BOLD signal as the predictor for each participant i in position j. The squared standard deviation of position j for participant i can be calculated as:

$$\frac{Dev_{ij}^2}{s_{\bar{B}_{ij}}^2}.$$

Because the Dev_{ij} that have this autocorrelation appear in this quantity, these squared standard deviations would be correlated as well but with a correlation that is the square of the correlation for the deviations. Fig. 7 reveals that the correlations among these statistics also decrease as an exponential function of the lag but in this case dropping from an r of 0.7 for lag of 1 to a value near 0 for a lag of 10.

Although these squared standard deviations are correlated, individually they have a distribution of values very much like a chi-square distribution with 1 *df*. This is displayed in Fig. 8a that compares the proportion of values in various intervals and the expected proportion from a chi-square distribution with 1 *df*. Although the squared standard deviations individually behave as chi-squares, their sums do not because of the correlations. Fig. 8b shows the distribution of sums of 20 of these along a specific curve for a specific participant. If they were independent, these sums would behave like a chi-square distribution with 20 *df*. However, Fig. 8b has many more small and large values than predicted, reflecting the correlation among the values. Fig. 8b also compares the gamma distribution described in the main text of the article based on the Kotz and Adams (1964) formula for correlated chi-squares. The figure confirms that this gives a good approximation to the actual distribution of the sums. Again, these conclusions do not depend on onset-locked averages versus event-locked averages.