

Supplementary Material

StagesCode.zip contains the information necessary to obtain the results in “Hidden Stages of Cognition Revealed in Patterns of Brain Activation”. It also contains a readme file that describes how to calculate the figures in the paper using 20 principal component scores for the 20,619 scans that constitute the correct trials. These principal component scores are derived from a very much larger data set, which is the activity for all scans for the 8755 regions for all problems for all subjects, as described in the main paper. Below we describe the steps in processing to derive the data in **zpcas** in the Matlab file associated with the paper.

- 1. Estimating the Bold Response.** We calculate the BOLD response as the percent change from a linear baseline defined from first scan (beginning of fixation before problem onset) to last scan (beginning of fixation before next problem). In our data this results in a matrix of 143,506 rows (scans) by 8755 (regions).
- 2. Extracting the Signal.** We assume BOLD response is produced by the convolution of an underlying activity signal with a hemodynamic response. The hemodynamic function we assumed is the SPM difference of gammas (Friston et al., 1998 – $g = \text{gamma}(6,1) - \text{gamma}(16,1)/6$). A Wiener filter (Glover, 1999) with a noise parameter of .1 was used to deconvolve the BOLD response into an inferred activity signal (Matlab: `deconvwnr(bold,g,.1)`). To an approximation, this results in shifting the BOLD signal backward by 2 scans (4 seconds) to compensate for its lag relative to neural activity. We have used this simple scan shift in earlier studies (e.g., Anderson et al., 2010, 2012) where there was not a constant baseline activity interspersed at regular intervals. The result is another matrix of the same size as the BOLD response matrix. For further analysis we only keep the scans during which the problem is presented. This results in a 38,627x8755 matrix representing 7003 trials across the 80 subjects (37 of the 7040 potential trials were lost due to scanner error, 3 subjects did not finish last scanner block)
- 3. Principal Component Analysis.** The inferred signals from the first step are organized into a matrix whose rows are all the scans for all the subjects and whose columns are the 8755 regions. Each region in this matrix is z-scored to focus on its sequential structure and not its mean activation and to guarantee that each region contributes equal variance. The resulting z-scored matrix is subject to a spatial principal component analysis that produced a same size matrix of orthogonal principal components. The factors are ordered by the amount of the variance they account for in decreasing order.
- 4. 20 Normalized Components.** The first 20 principal component dimensions account for 53.5% of the variance. Our analysis focuses on these components and only the values for problem-solving scans involving correct problems (from stimulus presentation to response completion). To assure that all dimensions are equally weighted and all subjects equally represented in the HSMM-MVPA analysis we take the zscores of each dimension for the correct solving scans for each subject. This constitutes the **zpcas** data file (20,619 x 20 matrix representing 4146 trials) that is the principal input to the HSMM-MVPA analysis.

The principal component values are selected to be orthogonal dimensions that capture

dimensions of greatest variation. Thus, the pattern in a scan, or more generally, in a problem-solving stage, can be represented as a linear combination of these principal components. However, there is no reason to suppose that individual principal component dimensions would be cognitively meaningful. Nonetheless, to give a sense of what the principal component dimensions are and how they vary across scans we looked at the two dimensions that showed the greatest variance across the 4 stages we identified (account for 1/3 of the variance among the brain signatures). These are the third and twelfth dimensions. Parts (a) and (b) of the following Figure S1 show those two dimensions (the end of this document gives them for all 20 dimensions). Part (c) shows their values for the 7 scans in each of the problems of Figure 1. It can be seen that Factor 3 weights heavily the left motor area and it tends to rise at the end of trial when the response is being prepared and executed. It can be seen that Factor 12 tends to prefrontal (particularly left) regions and the right motor area that would be active if participants were finger counting to keep track of the number of items to add. It tends to rise during the middle scans when the problem is being solved.

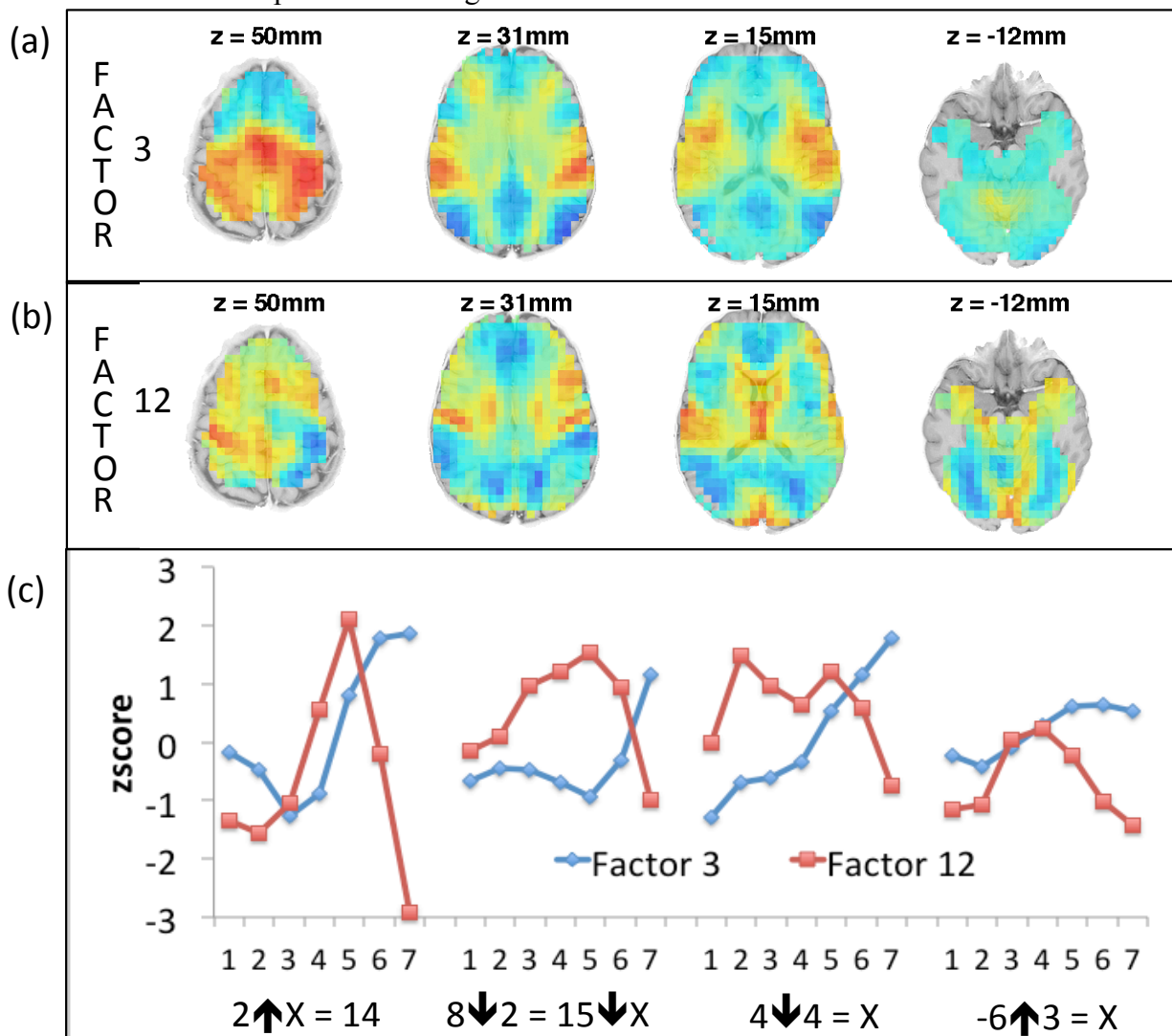


Figure S1. (a) & (b) Weighting of brain regions (dark blue = -.04; dark red = +.4) in defining two PCA factors (note radiological convention: left is on right). (c) The values of these two factors for the 7 scans of the four problems.

The remainder of this document provides some further detail about the results reported in the paper.

Split-Half Reproducibility.

To determine the reliability of the parameter estimates that underlie our conclusions, we performed a split-half analysis where we separately fit the data of the odd and even trials. The matrix below displays the correlations between the 20 principal component means that define each stage for odd and even trials:

		Even Trials			
		Encode	Plan	Solve	Respond
Odd Trials	Encode	0.993	0.154	-0.406	-0.382
	Plan	0.080	0.952	-0.269	-0.479
	Solve	-0.421	-0.247	0.973	-0.454
	Respond	-0.384	-0.516	-0.416	0.998

There are high values down the main diagonal supporting the consistency of results. It is worth noting that the off diagonal cells are near zero or negative, indicating that each state has a very different PCA representation.

The stage durations are characterized by gamma distributions. The matrix below gives the parameters of the gamma distribution when estimated for all the data and when estimated for each half:

	All Trails			Odd Trails			Even Trails		
	Shape	Scale	Mean	Shape	Scale	Mean	Shape	Scale	Mean
Encode	5.2	0.39	2.04	4.0	0.54	2.19	5.2	0.4	2.07
Plan	1.4	1.73	2.40	1.1	2.05	2.34	1.4	1.7	2.40
Solve	0.5	5.51	2.75	0.5	5.35	2.67	0.5	5.7	2.85
Respond	4.7	0.55	2.58	5.0	0.53	2.64	4.6	0.6	2.55

Shape and Scale are the parameters that define the gamma. Mean reflects the mean time of the gamma distributions, computed as the product of scale and shape. The correlations between odd and even trials are 0.960 for the shapes and 0.996 for the scales.

Four and Five Stage Models.

The following table displays the relationship between the 4-stage model that was the focus of the paper and the 5-stage model that appeared to improve on the HSMM-MVPA performance of the 4- stage model.

		5 Stage Model					Variance of Factor Scores
		1	2	3	4	5	
		Correlations Among Factor Scores					
4 Stage Model	1	1.000	0.162	-0.355	-0.564	-0.230	0.067
	2	0.141	0.999	-0.235	-0.430	-0.540	0.061
	3	-0.412	-0.275	0.996	-0.303	-0.491	0.060
	4	-0.378	-0.536	-0.517	0.941	0.968	0.104
Variance of Factor Scores		0.068	0.067	0.065	0.053	0.218	

The first three stages of the 4- and 5-stage models have factor score weights that are highly correlated and show similar variation. The mean durations of these stages are also very similar (see table below). Stages 4 and 5 of the 5-stage model both show strong correlations with stage 4 of the 4-stage model. They also have summed duration that is close to the duration of the last stage in the 4-stage model. Stages 4 and 5 in the 5-stage model are also fairly strongly correlated ($r = .828$), but the 5th stage shows much greater variability among the scores (see variability of factor scores in the table above). The reason there is an advantage to estimating a 5th stage is that its more extreme values capture the activity at the very end of the trial.

An examination of the correlation of trial-by-trial estimates of stage durations suggests a similar conclusion with the duration of the last stage in the 5-stage model correlating strongly with the last stage in the 4-stage model:

		5 Stage Model					Mean Stage Duration (sec.)
		1	2	3	4	5	
		Correlations Among Stage Durations					
4 Stage Model	1	0.986	0.297	0.125	0.138	0.189	2.09
	2	0.324	0.992	0.205	0.188	0.241	2.26
	3	0.154	0.250	0.997	0.127	0.171	3.06
	4	0.205	0.254	0.185	0.426	0.895	2.53
Mean Stage Duration (sec.)		2.04	2.09	2.88	1.96	0.98	

It is also the case the duration of the response stage correlates quite strongly on a trial-by-trial basis with the actual time keying the answer. The last stage of the 4-stage model correlates .53 with the actual response duration (next largest correlation with a stage is .17 with Stage 2). The last stage of the 5-stage model correlates .50 with the actual response duration (next largest .20 for stage 4).

Comparison with Anderson & Fincham.

Figure S2 shows the correspondence between the activation patterns for the stages in this study and Anderson & Fincham (2014). Since each study had independent PCAs, we looked at the activation in the regions common between the two studies. The tables below provide the correlation between the stages and the mean deviation in activation values:

Correlations		Current Study				Mean Deviations		Current Study			
		Encode	Plan	Solve	Respond			Encode	Plan	Solve	Respond
AF (2014)	Encode	0.945	0.847	0.752	0.768	(2014)	Encode	0.07%	0.20%	0.24%	0.19%
	Plan	0.887	0.942	0.922	0.778		Plan	0.14%	0.14%	0.20%	0.19%
	Solve	0.760	0.813	0.855	0.810		Solve	0.21%	0.19%	0.15%	0.21%
	Respond	0.681	0.725	0.716	0.958		Respond	0.21%	0.22%	0.22%	0.12%

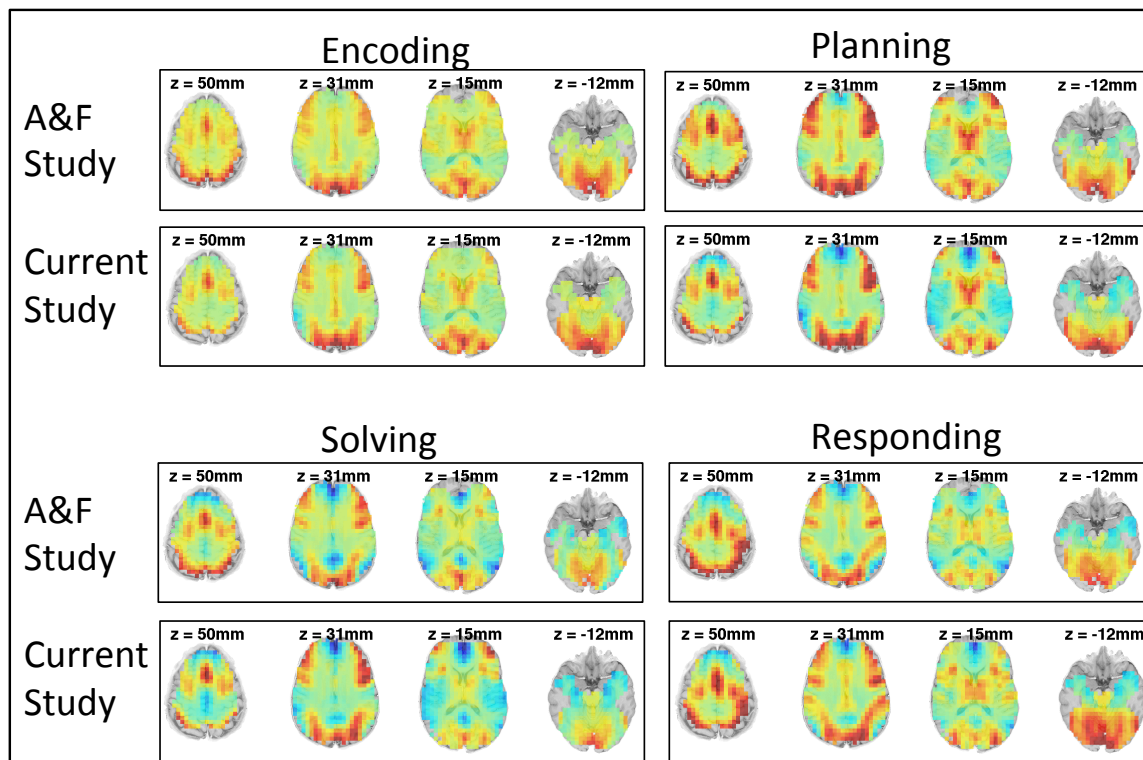


Figure S2. The brain signatures of the 4 stages. Dark red reflects activity 1% above baseline while dark blue reflects activity 1% below baseline. The z coordinate displayed for a brain slice is at $x=y=0$ in Talairach space. Note brain images are displayed in radiological convention with left on right. Compared are the results from Anderson and Fincham (A&F, 2014) and the current study.

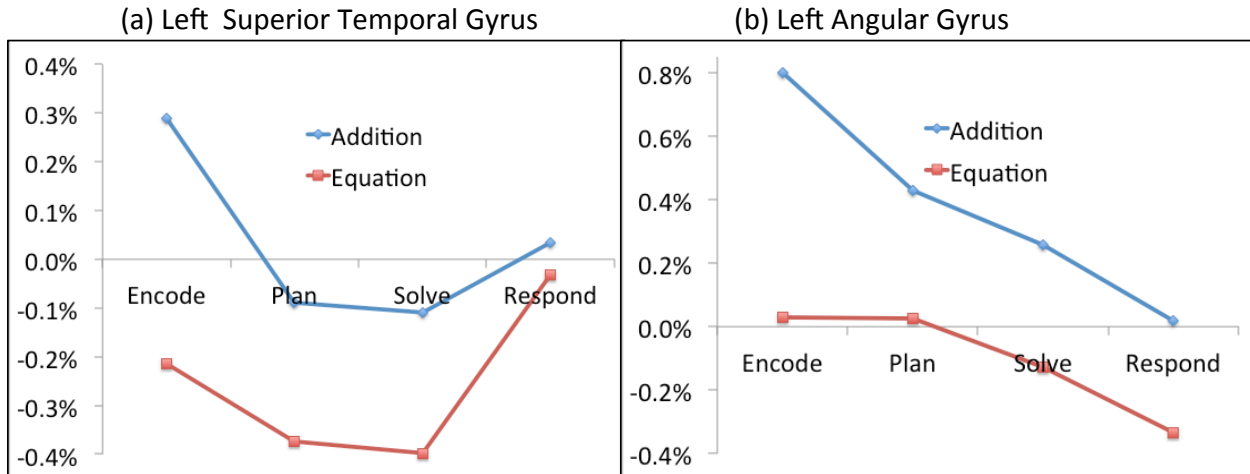
The correlations above are quite positive in all cases in contrast to the earlier correlations in PCA space because mean activation is subtracted before calculating the PCAs and because the PCAs are orthogonal and centered on 0. While there is similarity among all stages, in each case the corresponding stage signatures between experiments are more strongly related.

Analysis of Brain Imaging Data as a Function of Group and Stage

The brain signatures associated with the stages define complex multivariate patterns of activation over the brain. An exploratory whole-brain analysis (correct trials only) was conducted to determine which regions were active for each of the four stages. The imaging data were modeled using a general linear model (GLM). The first-level design matrix for each participant included 5 model regressors and a baseline model of an order-4 polynomial to account for general signal drift. Four of the model variables corresponded to the 4 stage occupancy probabilities over all trials and the fifth corresponded to the feedback period for each trial. The design matrix regressors were constructed by convolving the stage occupancy probabilities and the feedback boxcar function with the standard SPM hemodynamic response function (Friston et al., 2011). Each first-level GLM yielded 5 beta weights per voxel for each participant. Our focus was on the 4 beta weights corresponding to the 4 stages. A group-level mixed-design analysis of variance (ANOVA) was performed. We looked for regions showing either main effects of stage or Group (Addition vs. Equation). Regions of at least 28 spatially contiguous voxels with voxel-wise significance of 0.001 were identified. These constraints control familywise error at the 0.01 level, as estimated by simulation (Cox, 1996; Cox & Hyde, 1997).

The majority of the brain was selected as showing significant effects of stage, which is not surprising or interesting, given that the scans were assigned to stage on the basis of stage signatures. However, stage identification was not sensitive to training condition. The table below reports the four left hemisphere regions showing significant effects of training group. The table also gives the results of ANOVAs performed on average activity in these regions. Stage is always highly significant, reflecting the larger whole brain pattern. Unsurprisingly, group is also always significant, given that group difference were used to select the region. Of particular interest are the interaction terms which need not be significant, yet three of the regions do show significant interactions. The basic pattern shown in all four regions is the same: greater activity for the Addition group and this difference is largest for Encode Stage and least for the Respond Stage. The figure below on the next page shows two variations on this basic pattern for the two regions that show the strongest interactions – the Superior Temporal Gyrus and the Angular Gyrus.

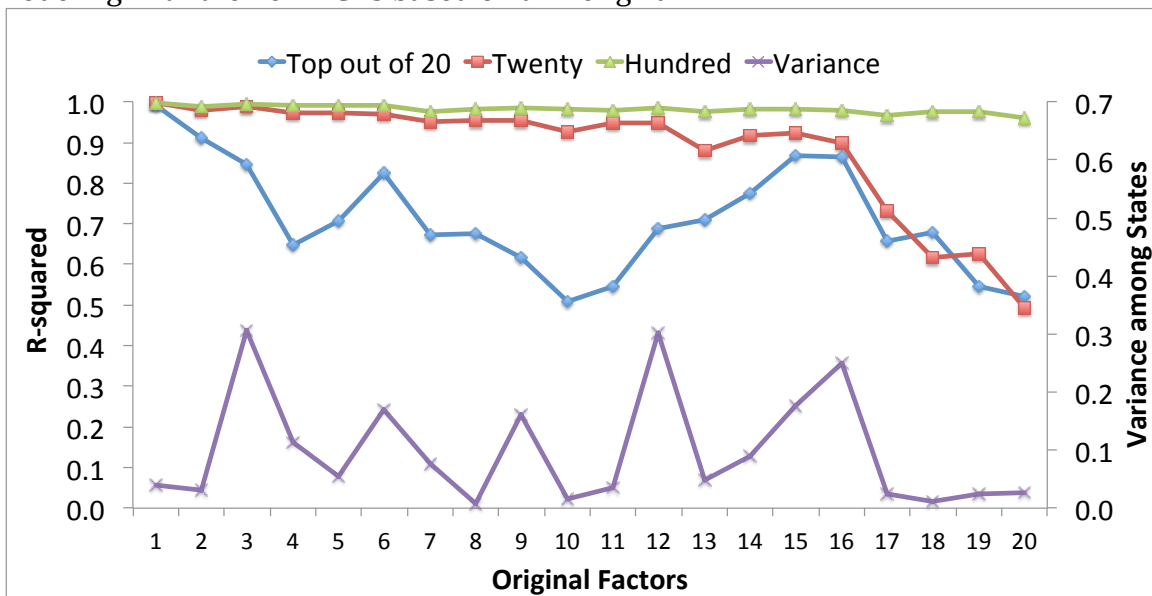
(Regions showing Significant effect of Condition (Addition vs. Equation))		Brodmann Area(s)	Coordinates (x,y,z)	Voxel Count	Anova Results			
					Stage F(3,234)	Condition F(1,78)	Interaction F(3,234)	Interaction p-value
1	L Superior Temporal Gyrus	39, 22	-48,-53,19	106	14.16	36.34	4.90	0.003
2	L Middle/Superior Temporal Gyrus	21	-58,-22,-4	53	14.57	27.36	1.34	0.261
3	L Angular Gyrus	39	-41,-71,31	31	23.68	18.72	3.88	0.010
4	L Superior/Middle Temporal Gyrus	22	-56,-40,6	29	20.79	20.44	2.70	0.046



Effects of 2-by-2 Aggregation of Voxels

The results in the paper were based on an analysis stream that started with BOLD measures averaged over 2x2 squares of voxels in a brain slice. This is motivated because of uncertainty in the exact spatial correspondence of subjects. It also produces a factor of 4 speed-up in processing and reduction in storage space. To investigate whether anything would change if we had not aggregated into 2x2 regions, we applied the analysis stream to single voxel activity. Rather than 8755 2x2 regions we started the analysis with 34,945 voxels that composed the 8755 (some 2x2 regions only consisted of 3 voxels because they were at the edge of the brain). The end of this analysis stream produced a new 20,619x20 matrix based on the principal dimensions of variation in the finer grain-size data. This matrix and results with it are contained in the **newzpcas.mat** in the Matlab code folder.

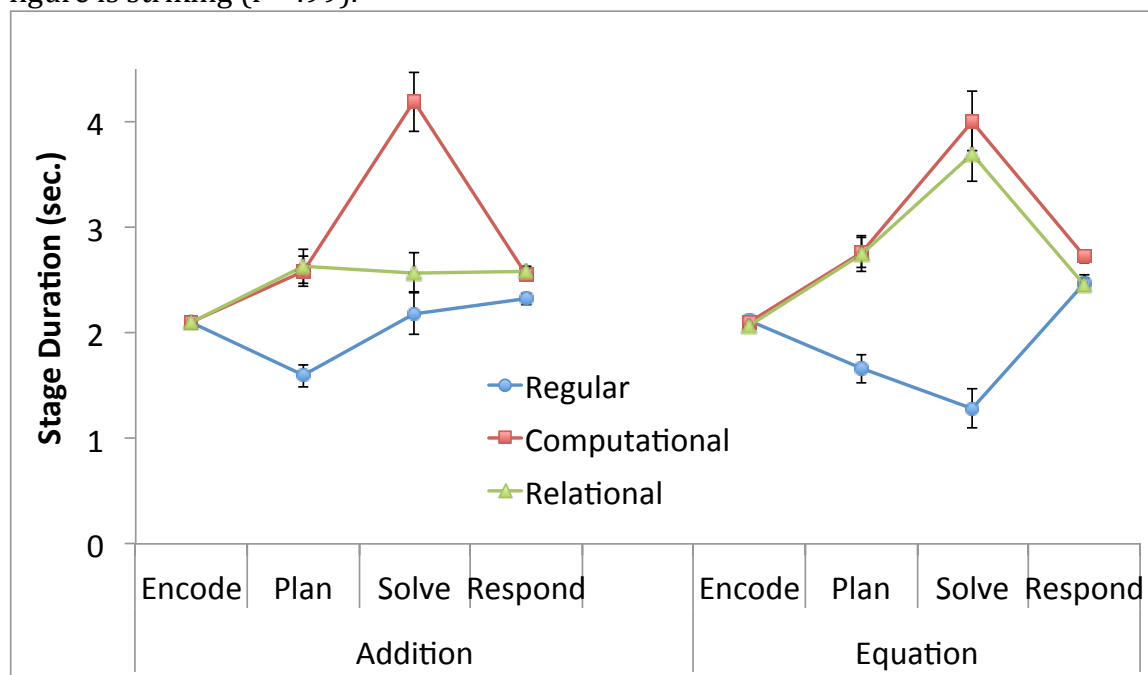
The two questions of interest are what are the differences between these dimensions and the old dimensions and what are the consequences of these differences for the conclusions we drew in the paper. The figure below contains a relevant analysis for the first question, showing how well one can account for our original 20 components used in the modeling with the new PCAs based on a finer grid.



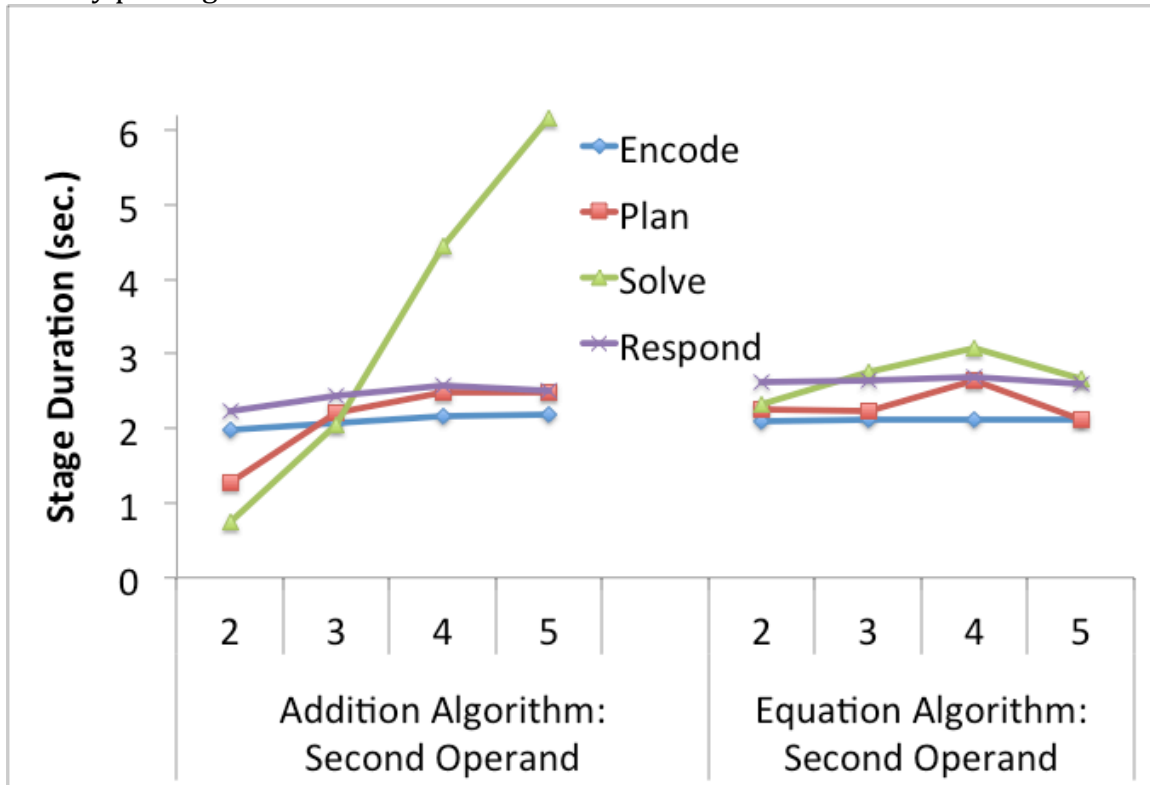
The figure displays 4 values for each of the 20 original factors (x-axis – these 20 are displayed at the end of the Supplementary Material):

- 1. Top out of 20:** This shows the strongest correlation between each original factor and any of first 20 new factors computed over the 20,619 scans. The first 6 old factors correlate most strongly with the corresponding new factors, but the correspondences change for later factors. For instance, the 7th old factor correlates most strongly with the 9th new factor. The 1st old factor (which largely reflects brain-wide fluctuation) seems totally reproduced by the 1st new factor, but the story for the remainder is more complicated:
- 2. Twenty:** This shows how well we can reconstruct the each of the original factors as a linear combination of the new 20. Reconstruction is high for the first 16 factors. Thus, the new space largely contains the space spanned by the first 16 original factors, but the axes through that space have rotated (and perhaps shrunk or expanded).
- 3. Hundred:** This shows how well we can reconstruct the original 20 factors from the top 100 new factors. This reveals that the some of the variation in the last 4 original factors has shifted down to later factors in the new analyses.
- 4. Variance:** Indicates how much that old factor varied among the 4 Stages -- i.e., its importance to the stage definitions. The last 4 factors were not important.

Since the new factors reflect most of the variance that drove the old analysis one would suspect the results would be the same if we used the new factors. We checked to see if the residual differences would affect our conclusions about the factors the timing of stages in conditions. The figure below should be compared with Figure 6 in the published paper. It shows how the time in the stages (derived in the new analysis stream) varied with problem type and algorithm. Its correspondence with the original figure is striking ($r > .99$).



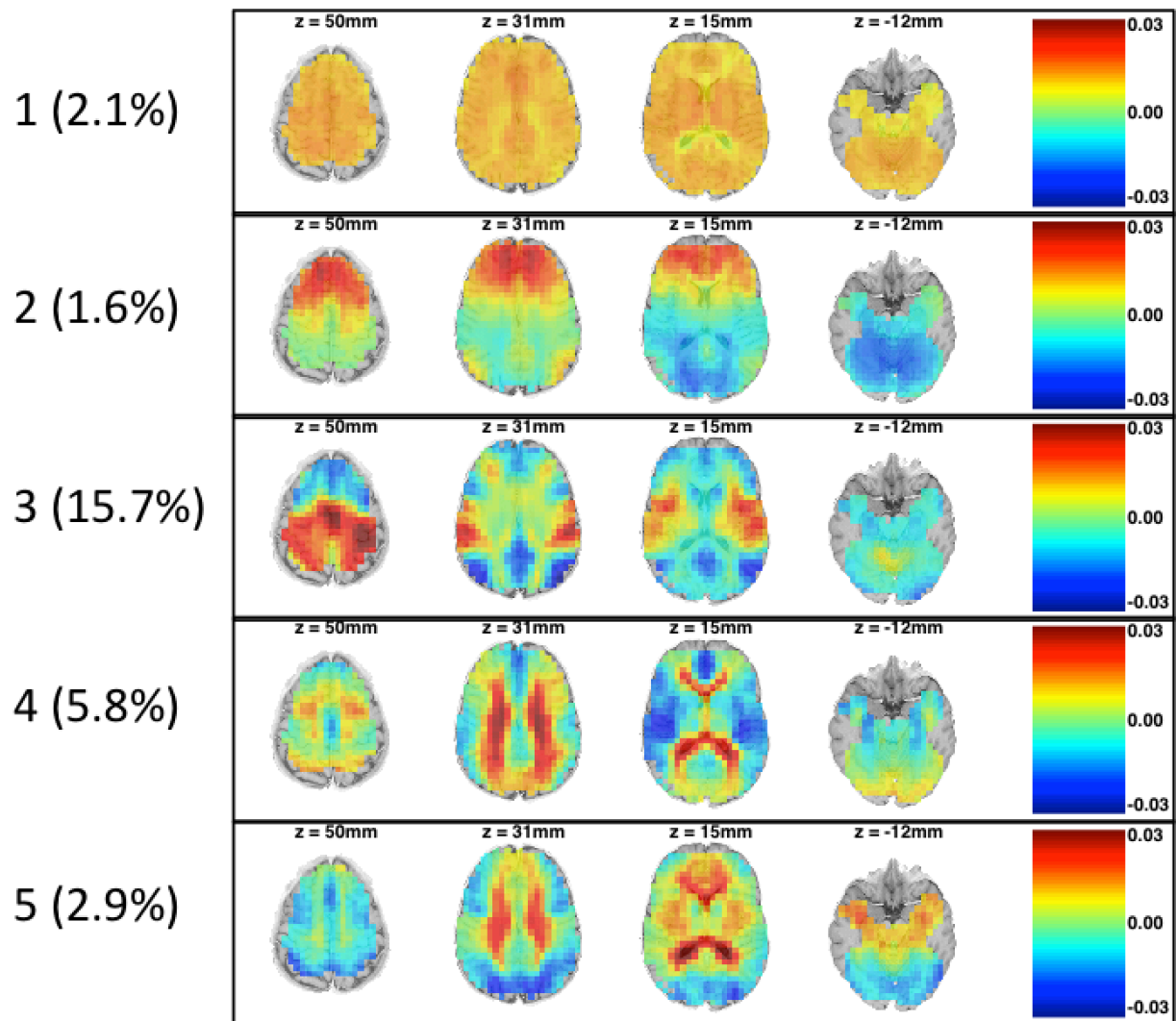
Similarly, we can calculate the results displayed in Figure 7 given the new parsing of the trials. These data are displayed below and again show a strong correlation ($r > .99$) with the latency parsing based on factors extracted from the 2x2 voxels.



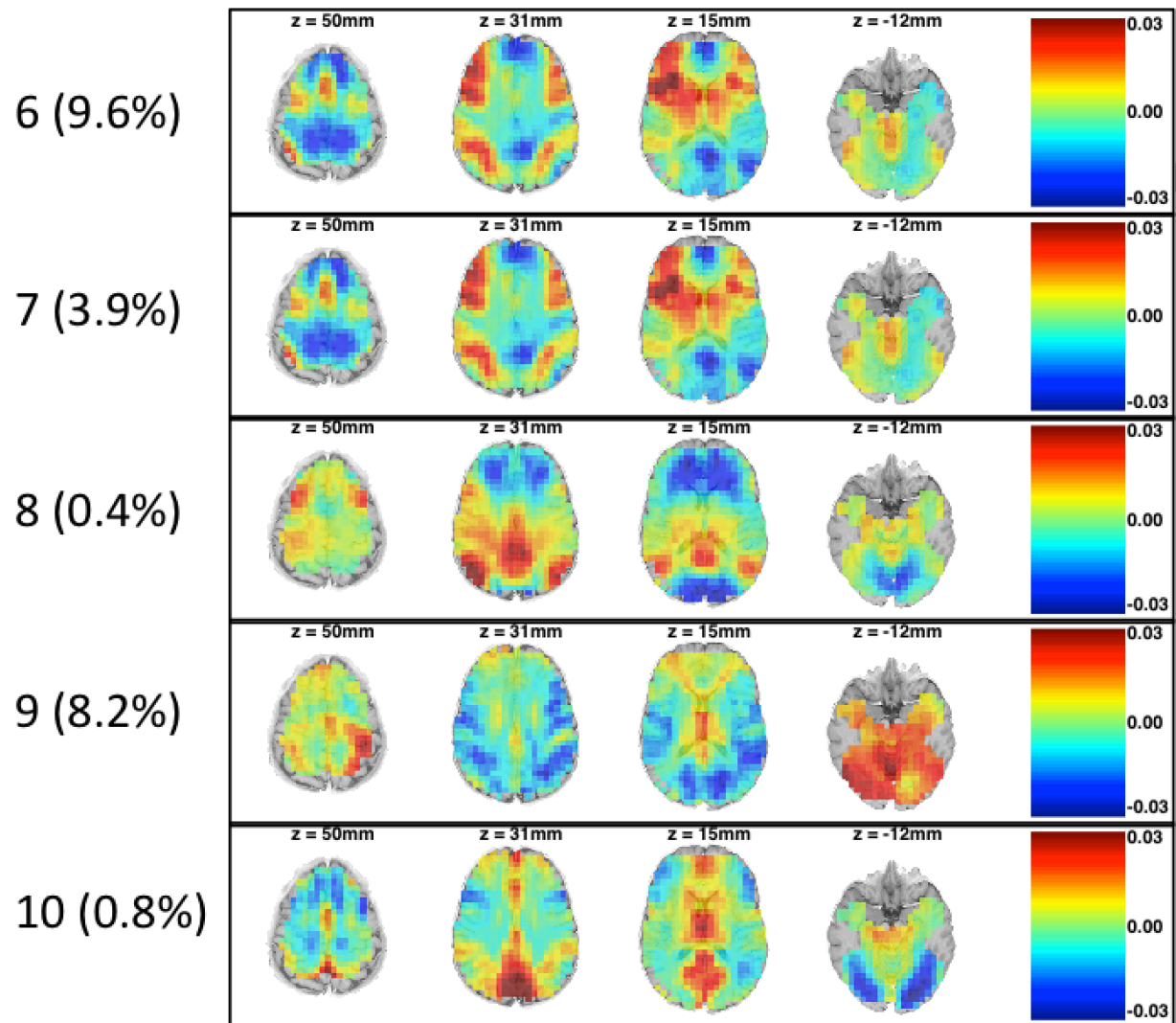
References

- Anderson, J. R., Betts, S., Ferris, J. L., Fincham, J. M. (2010). Neural Imaging to Track Mental States while Using an Intelligent Tutoring System. *Proceedings of the National Academy of Science, USA, 107*, 7018-7023.
- Anderson, J. R., Betts, S. A., Ferris, J. L., & Fincham, J. M. (2012). Tracking children's mental states while solving algebra equations. *Human Brain mapping, 33*, 2650-2665.
- Anderson, J. R., Fincham, J. M., Yang, J. & Schneider, D. W. (2012). Using Brain Imaging to Track Problem Solving in a Complex State Space. *NeuroImage, 60*, 633-643.
- Friston, K. J., Ashburner, J. T., Kiebel, S. J., Nichols, T. E., & Penny, W. D. (Eds.). (2011). *Statistical Parametric Mapping: The Analysis of Functional Brain Images: The Analysis of Functional Brain Images*. Academic Press.
- Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage, 9*, 416-429.

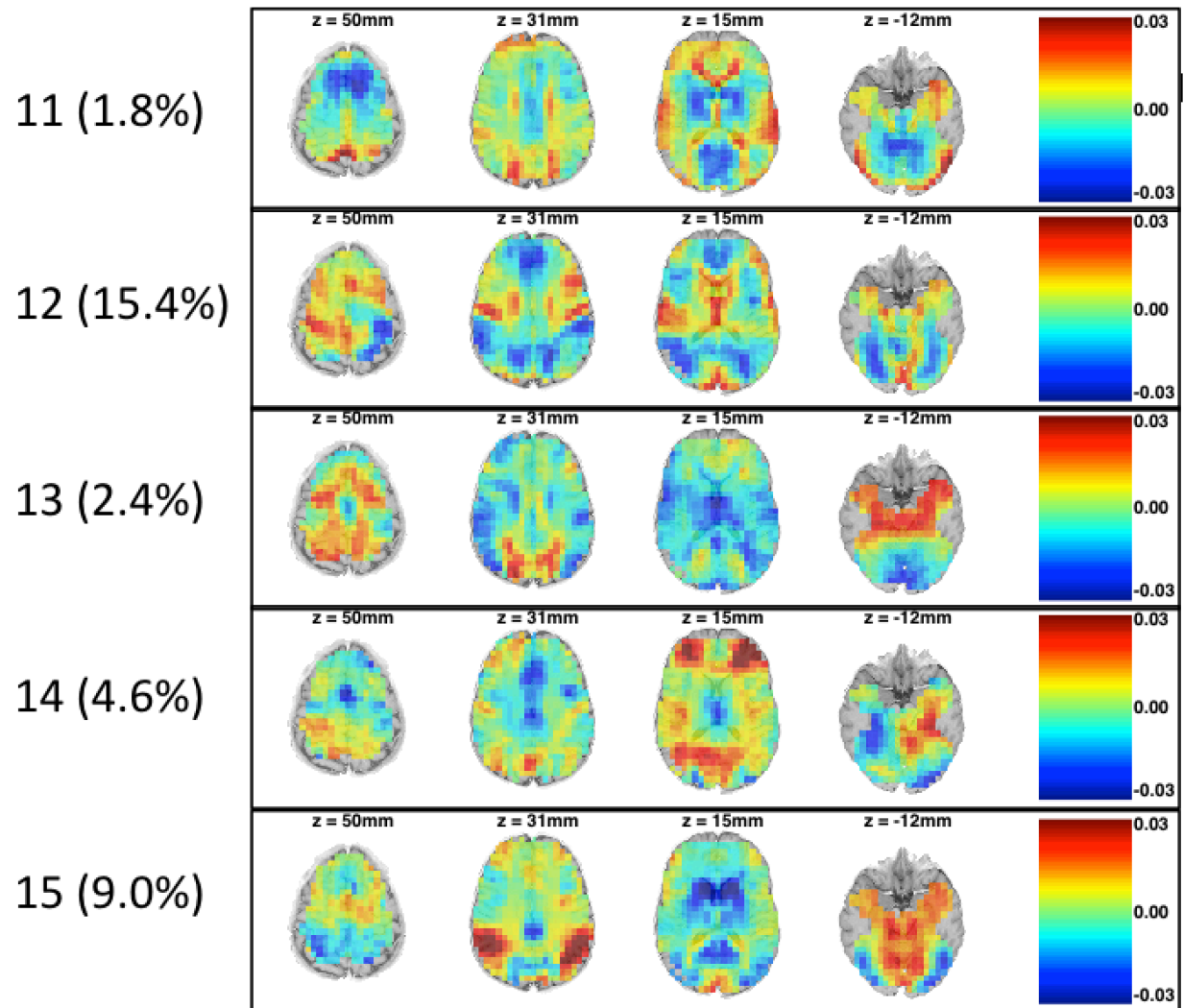
Factors 1-5 and proportion of variance among stages explained:



Factors 6-10 and proportion of variance among stages explained:



Factors 11-15 and proportion of variance among stages explained:



Factors 16-20 and proportion of variance among stages explained:

