Physics instruction induces changes in neural knowledge representation during successive stages of learning

Robert A. Mason, Carnegie Mellon University
Marcel Adam Just, Carnegie Mellon University
Physics instruction induces changes in neural knowledge representation during successive stages of learning

Robert A. Mason*, Marcel Adam Just

Center for Cognitive Brain Imaging, Psychology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

A R T I C L E   I N F O

Article history:
Accepted 29 December 2014
Available online 7 February 2015

Keywords:
Learning
FMRI
Causality
Mechanical systems
Knowledge representation

A B S T R A C T

Incremental instruction on the workings of a set of mechanical systems induced a progression of changes in the neural representations of the systems. The neural representations of four mechanical systems were assessed before, during, and after three phases of incremental instruction (which first provided information about the system components, then provided partial causal information, and finally provided full functional information). In 14 participants, the neural representations of four systems (a bathroom scale, a fire extinguisher, an automobile braking system, and a trumpet) were assessed using three recently developed techniques: (1) machine learning and classification of multi-voxel patterns; (2) localization of consistently responding voxels; and (3) representational similarity analysis (RSA). The neural representations of the systems progressed through four stages, or states, involving spatially and temporally distinct multi-voxel patterns: (1) initially, the representation was primarily visual (occipital cortex); (2) it subsequently included a large parietal component; (3) it eventually became cortically diverse (frontal, parietal, temporal, and medial frontal regions); and (4) at the end, it demonstrated a strong frontal cortex weighting (frontal and motor regions). At each stage of knowledge, it was possible for a classifier to identify which one of four mechanical systems a participant was thinking about, based on their brain activation patterns. The progression of representational states was suggestive of progressive stages of learning: (1) encoding information from the display; (2) mental animation, possibly involving imagining the components moving; (3) generating causal hypotheses associated with mental animation; and finally (4) determining how a person (probably oneself) would interact with the system. This interpretation yields an initial, cortically-grounded, theory of learning of physical systems that potentially can be related to cognitive learning theories by suggesting links between cortical representations, stages of learning, and the understanding of simple systems.
A number of tasks have shown brain-based changes in activation patterns due to training or instruction-based learning. Typically, these tasks examine brain changes in activation, in which it may be difficult to separate new representations from learned processes. Examples of tasks that examine activation in training-based learning include artificial grammars (Petersson, Folia, & Hagoort, 2012), perceptual category learning (Poldrack et al., 2001), and motor learning (Toni, Krams, Turner, & Passingham, 1998). Instruction-based learning tasks such as algebra (Anderson et al., 2012) have also resulted in activation changes. One study demonstrated both activation changes and white matter changes as a result of both instruction and repeated training in word decoding in children with dyslexia (Meyler et al., 2008; Keller & Just, 2009). Unlike these previous studies, here we look not for changes in tissues or in regional activation, but in the neural representations of specific concepts using recent methods that can identify the nature of the information that is being coded by a given fMRI activation pattern.

The neural changes in our study were assessed in terms of the multi-voxel fMRI-measured activation pattern that occurs when participants think about how a particular mechanical system works. More precisely, the study assessed how their neural representation of a system changed as they learned more about its workings. The change in knowledge about specific mechanical systems should produce measurable changes in the neural representations of those systems. Furthermore, the changes may be directly related to the content of the instruction, such that instruction that describes shared properties between systems may increase their neural similarity.

Participants were taught with a series of successive increments of information about mechanical systems. The first level of explanation provided information about the components of the mechanical systems. The second increment included limited functional information. The third increment of explanation included the entire functional and causal sequence of the components of the mechanical systems. Each of these instructional steps should result in discernable neural changes. More specifically, the progressive deepening of the explanations of the systems might be expected to produce increasing involvement of cortical systems implementing higher-level psychological processes, and unchanged involvement of lower level perceptual systems that process the visual stimulus.

Despite the absence of prior neuroimaging investigations of mechanical systems, previous research in the brain bases of general cognitive processes does provide guidance as to which cortical systems might be involved. A set of eight potential cognitive processes, which have previously been associated with cortical systems, are postulated to correspond to regions or small sets of regions (networks) involved in understanding how mechanical systems work. These eight processes (and postulated cortical systems) consist of: (1) mental animation (bilateral parietal: Boronat et al., 2005), (2) causal reasoning (right tempo-parietal and medial prefrontal: Mason and Just, 2011), (3) embodied cognition (pre-and post-central: Rueschemeyer et al., 2010), (4) semantic knowledge (left temporal: Price, 2000), (5) language in context (bilateral inferior frontal: Mestres-Missé et al., 2008), (6) biological/goal-directed motion (right temporal: Pelphrey et al., 2003), (7) rule learning (middle and superior frontal: Bunge, 2004), and (8) visual processing (occipital cortex). The contributions of these various systems might be expected to shift as the instruction and learning progresses.

The goal of this study was to examine the changes in the neural representation over the course of learning and instruction, rather than establishing the correspondence between cognitive functions and brain regions. We developed several hypotheses concerning changes in representation. First, prior to instruction, during the first exposure to only the diagram and label, the hypothesis is that the participating regions would be primarily visual in nature, loading on the occipital cortex. Subsequent neural representations should involve relatively less occipital participation. Second, following the introduction of causality information, the representation could be expected to be distributed across a large set of systems including causal inference related regions (medial frontal and right tempo-parietal) for inferring causal relations among the components’ motions. Third, bilateral parietal, particularly the intraparietal sulcus, should increase in participation once components of the mechanical systems are introduced as a result of imagining components moving with respect to each other. Intuitively imagining the components moving may be a part of mental animation (Hegarty, 1992). These hypotheses provide a starting point for examining the changing involvement of cortical systems during learning.

Several recently developed methods for assessing neural knowledge representations were used in the study. One of these was the machine learning and classification of the multi-voxel activation patterns associated with each of the mechanical systems (Just et al., 2010; Mitchell et al., 2008). A related method analyzed the locations of the types of voxels whose activation levels were modulated in a consistent way by the different mechanical systems (Just et al., 2010). A third method used representational similarity analysis to assess how similarly-described systems became neurally more similar (Connolly et al., 2012). These methods can be used to converge on an account of how the neural representations change as instruction and knowledge cumulative.

Materials and methods

Participants

Fourteen college students (6 females, all right handed and native speakers of English) between the ages of 18 and 26 years (M = 21.57; SD = 2.79) participated and were included in all of the analyses (no subjects were excluded). Each participant gave signed informed consent approved by the Carnegie Mellon University Institutional Review Board. Each participant received 5 minutes of practice with the experimental paradigm on a single training item (that was not included in the experimental stimuli) before performing the task in the scanner. In a debriefing session, all participants responded positively when asked if they felt they had gained an understanding of how the mechanical systems worked. Additionally, when they were asked if they had “prior knowledge of how any of the systems worked” only one participant said he had a very basic understanding of the systems. This participant’s data did not differ from the others’ so it was retained in the analysis.

Experimental Design

In the scanner, participants were taught how four familiar devices work (a bathroom scale, a fire extinguisher, car brakes, and a trumpet). The systems were selected to vary across some potentially interesting dimensions (e.g., manipulation by hand versus foot, being composed of different types of mechanical components) as well as meeting two criteria: (1) amenability of the explanation to segmentation into successive stages; (2) informal assessment that students who were not in science or engineering would be unlikely to know how the system worked.

The experimental design consisted of the four items presented in two types of blocks: thinking (or “test”) blocks and explanation (or “training”) blocks. The experimental design and timing of all events (presentations, blocks and scans) are shown in Fig. 1. In the thinking blocks, which provide the main data for this study, participants were presented with each of the four items and asked to “Think about how this mechanical system might function.” The explanation blocks cumulatively described how the components of each system work together to cause the system to function. During the explanation blocks, subjects were asked to read each sentence and “Think about the functioning of each stage of the mechanical system.”

In the thinking blocks, each stimulus item consisted of a realistic picture of the system above a schematic diagram and a verbal label for the system, as shown in Fig. 1A. In a thinking block, the mechanical systems were presented in six presentations (i.e., repetitions) of the four systems. The presentation order of the mechanical systems in each block was randomized using a Latin square design with an additional
constraint that a system did not appear first in a four-item presentation after appearing last in the previous four-item presentation. The six thinking blocks were separated from each other by one of the explanation blocks. This resulted in 144 presentations of systems in total across all the thinking blocks (6 blocks × 6 presentations × 4 items). Each item appeared for 4 sec during which participants were instructed to “Think about how this mechanical system might function.” In between each item an “X” appeared on the screen for 5 sec. Participants had been instructed to try to clear their mind during this rest interval.

There were three different types of explanation blocks: component description, partial causal information, and full functional information. (The second and third types were presented twice). The first explanation block (component description) consisted of a picture of the system above a schematic diagram of that system, a verbal label for the system, and a sentence describing the components of the system, a format shown in Fig. 2B. An example of a sentence that provided a description of components is A bathroom scale consists of a lever, a spring, a ratchet and a dial. Each explanation item was presented for 10 sec followed by a 7-sec rest period with an “X” on the screen. In the second explanation block (partial causal information), an additional two sentences were presented at the same rate (10 sec + 7 sec rest per sentence, in addition to the original component description sentence). During these blocks, the referenced component was highlighted in the schematic diagram by shading the rest of the diagram. This second type of explanation block introduced the first half of the causal sequence in the functioning of the system. The third type of explanation block (full functional information) included the same three sentences (component sentence and two causal sequence sentences) and appended the final two explanation sentences constituting a complete description of the mechanical system’s functioning. To assure that the language properties of the instructional sentences did not differ across stages of instruction, these texts were compared, with no major trends revealed with respect to the Flesch-Kincaid Grade Level scores (explanation block 1 = 6.7, explanation block 2 = 6.3, explanation block 3 = 5.8), and Kucera-Francis word frequency norms (excluding high-frequency function words (69.4, 81.3, and 80.6). The trend for lower frequency content words in the first explanation block reflects the introduction of all the components for each mechanical system.

fMRI scanning and data analysis parameters

Images were acquired on a Siemens Verio 3 T scanner at the Scientific Imaging & Brain Research Center (SIBR) of Carnegie Mellon University, using an EPI sequence with the following parameters for the 17 oblique axial slices: TR = 1000 ms, TE = 30 ms, 64 × 64-acquisition matrix, 5-mm thickness, 1-mm gap, flip angle = 60°, 32 channel head coil. Images were corrected for slice acquisition timing, motion-corrected, and were normalized to the Montreal Neurological Institute (MNI) space without changing voxel size (3.125 × 3.125 × 6 mm).

Although the main analyses used machine learning techniques (described in Sections 2.4 and 2.5), activation differences during the three types of explanation blocks (component, partial causal, full functional description) were assessed with conventional general linear modeling (GLM), to provide a converging measure of areas involved in processing the different incremental levels of instruction. The data here were resampled to 2 × 2 × 2-mm voxels, and the images were smoothed with an 8-mm Gaussian kernel. A GLM was constructed with regressors (a boxcar for each event duration convolved with a
hemodynamic response function) for each of the items (four) and each block type (thinking vs. explanation) and repeated presentations of thinking blocks (six) and fixation periods. Group t-test analyses were performed using a random-effects model (Friston et al., 1995) on the mechanical systems versus fixation contrast images (one per participant, per contrast). The t-maps from each contrast were calculated across the entire cortical volume, thresholded at $p < .05$ with a family-wise error correction and an extent threshold of 6 voxels. This technique is typically used to contrast the neural response during different cognitive processes, in this case, during reading and understanding how a mechanical system functions vs. fixation. The resulting activation maps would be expected to reflect the ongoing psychological processes.

Fig. 2. (A) An example of the four mechanical systems as presented in the thinking blocks. (B) Stimuli for each type of explanation block for the bathroom scale. The portion of the each schematic diagram that is highlighted corresponds to the explanation sentence below the diagram providing partial causal information.
and not necessarily the neural representations of the mechanical systems.

**Voxel selection**

To focus on the neural representations of the specific mechanical knowledge that was being learned, the voxel selection procedure applied was designed to optimize classification accuracy in machine learning of neurosemantic representations (Damarla and Just, 2013; Just et al., 2010; Mitchell et al., 2008). This method selects voxels whose activation profile (their set of activation levels over the four mechanical systems) remains consistent across the multiple presentations of the set of items (this description will be expanded below). These voxels, referred to as stable voxels, are those that display a consistent tuning curve. This selection method has led to successful classification of the neural representation of a variety of concepts, such as concrete objects (Just et al., 2010), numerical quantities (Damarla and Just, 2013), and emotions (Kassam et al., 2013).

The analyses below focused on the stable voxels. The assumption here is that only the relatively stable voxels provide information about the neural representations of individual mechanical systems, because it is these voxels that discriminate among the systems. The measure of a voxel’s stability was the consistency of its set of activation values (its tuning curve) over the four mechanical systems each time the set was presented. A voxel’s stability was computed as the average pairwise correlation between its 4-system activation profiles across the multiple presentations that served as input for a given classification analysis (the number of presentations over which stability was computed was four or six, depending on the analysis). Here the 4-system activation profile of a voxel for a particular presentation refers to the vector of 4 responses of that voxel to the 4 systems during that presentation.

**Voxel selection by thinking block**

To test the assumption that the underlying neural knowledge representation undergoes change as a participant reads a series of progressive explanations of how the systems operate, stability maps (locations of the voxels that responded with a consistent multi-item pattern of activation across the set of four systems) were computed separately for each of the six thinking blocks. The data from the two thinking blocks that followed two repeated identical explanation blocks were averaged (i.e., thinking blocks 3 and 4 were averaged as were thinking blocks 5 and 6), reducing the number of stability maps to 4. The original design made provision for observing a change in the learning state from block 3 to block 4. When no such change was observed, those data from the two blocks were averaged. The four resulting stability maps were thresholded to the 5% top most stable voxels (approximately 400–500 voxels per participant). This threshold was estimated by determining the number of significantly different voxels in within-subject pairwise t-tests between blocks on activation maps at an uncorrected p < 0.001 level (using a standard SPM contrast procedure in which all items were specified with one regressor). This analysis was done only to estimate the number of voxels which might be expected to be different across blocks and not used in any other analysis.

To assess differences among the four resulting stability maps (for each thinking block), the mean stable voxel counts across participants in each of the nine postulated cortical networks (see Section 3.2) were submitted to an omnibus analysis of variance (2 × 2: networks(9) by blocks (4)). The results of this ANOVA will be presented where appropriate in the results in Sections 3.2 (3.2.1–3.2.4).

**Classification analysis**

The training and testing used cross-validation procedures that iterate through all possible partitions of the data into training and testing datasets. Voxel stabilities were computed in the training set of data only. The analyses used the 120 most stable voxels of the 15,000–20,000 voxels in each image (each 3.125 × 3.125 × 6 mm).

The choice of a subset size of 120 was guided by several previous studies in which 80 to 120 of the most stable voxels constituted the smallest sets that could still attain the highest classification accuracies (Just et al., 2010; Mitchell et al., 2008). The activation values for the stable voxels were normalized (mean = 0, SD = 1) across the items, separately for the training and test set, to increase comparability across the six blocks of presentations. The training set and test set were always rigorously kept separate from each other. A subset of the data was used to train a classifier to associate fMRI data patterns with a set of labels (the mechanical systems). These subsets were four of the six presentations for training with the two remaining sets used for testing. The classifier used here was a GNB-pooled variance classifier. GNB is a generative classifier that models the joint distribution of class Y and attributes and assumes the attributes $X_1,...,X_n$ are conditionally independent given Y. The classification rule is:

$$Y = \arg \max_{\theta} p(Y = y_i | \prod_{j} p(X_j | Y = y_i))$$

where $p(X|Y = y_i)$ is modeled as a Gaussian distribution whose mean and variance are estimated from the training data. In GNB-pooled variance, the variance of attribute $X_j$ is assumed to be the same for all classes. This single variance is estimated by the sample variance of the pooled data for $X_j$ taken from all classes (with the class mean subtracted from each value).

The classifier was tested on the mean of the two left-out presentations of each system. This procedure was reiterated for all 15 possible combinations (folds) of leaving out two presentations. The rank accuracy (hereafter, simply accuracy) of the classification performance was computed as the normalized rank of the correct label in the classifier’s posterior-probability-ordered list of classes. For example, if the classification were operating at chance level, one would expect a mean normalized rank accuracy of 0.50. A rank accuracy was obtained for each fold, and these rank accuracies were averaged across folds, producing a single value characterizing the prediction accuracy for each mechanical system. The mean accuracy across items was then computed. Random permutation tests (10,000 permutations) were performed for each type of classification; accuracies are reported as significant with $p < 0.05$ (equivalent to a rank classification accuracy above 58%).

**Representational similarity analysis**

Representational similarity analysis (RSA) provided a converging measure to assess the neural representation changes that occurred over the course of the knowledge acquisition (Kriegeskorte, 2009; Raizada and Connolly, 2012). The representational dissimilarity matrix (RDM) was calculated within each thinking block by computing the inter-item distance as (1 – the pairwise correlation) of the signal within stable voxels. The top 2.5% of voxels across the whole brain (~500 voxels per subject per block) was assessed by their stability ranking were used in the analysis. In the second stage of RSA, the RDMs were projected into two dimensions using multi-dimensional scaling (cmdscale in the stats toolbox for Matlab 6.5 (R13)). These dimensions can potentially be interpreted as features of the systems.

**Results**

**Overview**

The progressive stages of instruction describing how a mechanical system works resulted in corresponding changes in the states of neural knowledge representation of the systems. These successive brain states are identifiable (serve as signatures of the mechanical systems to which they correspond) and reflect the stages of cognitive processes that occur during this learning. Results will be interpreted as showing that there was a shift from visualizing how a system might move, to understanding
why it moved that way, to imagining how one might interact with the system to make it move. In anticipation of this interpretation, the four successive states are labelled as follows: pre-training state, visual layout understanding, causal chain understanding, and embodied virtual interaction. The interpretation was guided by two main sources of evidence, as described below. One source was the stability maps obtained during the time when participants were thinking about each of the mechanical systems (in thinking trials). (Stable voxels are those that exhibit a consistency of response in each of the four items across six blocks of trials.) A second source was the changes in the areas of activation across the stages of instruction that showed patterns of activation consistent with these labels.

Neural signatures of the four successive states of knowledge in the thinking trials

The analyses attempted to assess the neural representations of the mechanical systems during the thinking trials before and after each iteration of instruction in terms of the locations of stable voxels (assigned to anatomical areas using Anatomical Automatic Labeling (AAL) masks). To focus on brain regions that would a priori be expected to be involved in understanding mechanical systems, a subset of the AAL regions were then allocated to eight cortical systems (either as individual anatomical regions or sets of regions that are postulated to constitute a cortical network). As previously described, these eight cortical systems are believed to implement higher level psychological processes such as mental animation, causal reasoning, embodied cognition, etc. The set of eight were: (1) right temporal, (2) bilateral parietal, (3) left temporal, (4) medial prefrontal gyrus and right temporal-parietal junction (RTPJ), (5) bilateral middle and superior frontal, (6) bilateral pre- and post-central, (7) bilateral inferior frontal, (8) occipital cortex. (RTPJ was defined as an 8 mm sphere. Any voxels assigned to the non-AAL defined RTPJ were excluded from any other region.) The stable voxels that did not fit into one of these eight networks were allocated to a ninth set (and the size of the ninth set remained similar across stages). All of these cortical systems were tentatively assigned a cognitive role based on previous research as mentioned in the introduction and labeled in Fig. 3 (a rendering of the stable voxels across participants is shown in Fig. 4).

An ANOVA of the mean stable voxel counts in the nine postulated cortical networks in each of the four representational states revealed a significant interaction ($F_{(24,312)} = 1.84, MSe = 414.88, p = .01$), indicating that the locations of the stable voxels differed across the instructional stages. Several additional contrasts of specific ROIs and blocks were conducted and are reported where appropriate in Sections 3.2.1–3.2.4.

Pre-training state

Prior to any training regarding mechanical function, when only the picture, label, and diagram had been presented without any explanation of the system, the preponderance of the stable voxels (i.e., the postulated neural representation) was in visual areas. The stable voxels were located primarily in the occipital lobe as well as in the right temporal sulcus, lying along the sulcus between the anterior temporal region and the right temporo-parietal junction (RTPJ), and in motor regions (blue bars in Fig. 3). Of the non-occipital stable voxels, the proportion was greater in RTPJ and the motor cortex than the remaining regions.

Fig. 3. The preponderance of stable voxels per cortical system in each of the eight regions or networks gradually shifts from posterior to anterior areas over the course of instruction. (The four columns are derived from assessments made from thinking trials at four successive points during instruction; bar labels indicate anatomical region and hypothesized cognitive process). Dark shading indicates which state has the greatest participation of a given network.
(as indicated by the unfilled bars for the first column in Fig. 3); this difference was significant ($F_{(1,13)} = 9.09, MSe = 43.89, p < .01$).

**Early intermediate state**

After the components of the systems had been described in the first explanation block, the resulting stability maps, calculated from the second thinking block, again included a large number of voxels in visual areas, consistent with the type of instruction that had been presented. (For example, in the bathroom scale system, the component information was given as: A bathroom scale consists of a lever, a spring, a ratchet and a dial.) The majority of the remaining highly stable voxels were in parietal and superior frontal regions (indicated by red bars in Fig. 3). In the previous state, the proportion of non-occipital stable voxels in the highlighted regions was greater than the remaining regions (indicated by the unfilled bars; $F_{(1,13)} = 58.81, MSe = 78.74, p < .0001$). In fact, it was in thinking block 2 that the highest proportion of stable voxels was in the parietal regions (22.7%). This increase in the proportion of stable voxels from block 1 to block 2 was significant ($F_{(1,13)} = 3.37, MSe = 414.88, p < .10$).

**Late intermediate state**

The neural representations in the late intermediate state (stability maps calculated from the mean of thinking blocks 3 and 4) were obtained after participants had received a small amount of instruction about the causal relations in the systems but not yet a complete causal explanation. For example, in the case of the fire extinguisher, this limited functional information conveyed that Squeezing the handle of a fire extinguisher opens the gas canister, The pressurized gas is released into the water in the container. (Why the pressurized gas enables the device to function had not yet been specified at this point). The resulting neural representation included a diverse set of regions corresponding to a various types of information. In particular, stable voxels in the classical semantic regions of left temporal gyrus (primarily in posterior-superior temporal and inferior temporal) and the medial frontal (medFG) and RTPJ regions (the classic Theory of Mind/causal inference network) constituted a relatively large proportion during this stage (indicated by the green bars in Fig. 3). This change represented the largest shift in stable voxels to novel locations between blocks, with no set of regions being dominant. Thus, not surprisingly, a comparison of regions within the late intermediate state indicated there were no significant differences.

**Final state**

After the complete explanations of the mechanical systems had been provided, a larger proportion of stable voxels were located in the frontal lobe. The frontal cortex accounted for approximately 45% of the non-occipital stable voxels). This frontal lobe representation (indicated by yellow bars in Fig. 3) included the superior, middle, and inferior frontal gyri in both hemispheres, though the majority of the stable frontal voxels were close to the precentral gyrus. There was an increase in the number of stable voxels in the right pre- and post-central gyrus at this point (i.e., relative to the previous state), presumably indicating a motor-related component in the neural representation of the items. The comparison of the frontal regions to the other non-occipital regions indicated they comprised a significantly greater proportion of the stable voxels ($F_{(1,13)} = 5.87, MSe = 27.03, p = .03$).

**Classification of items using machine-learning techniques**

A machine learning classifier assessed the ability of the stable voxels to identify which of the four mechanical systems was being thought about in the thinking blocks. The mean classification rank accuracy across all 14 participants using the 120 most stable voxels was 62% (the threshold for rank accuracy reliably greater than chance at $p < .05$ was 57.9%). All participants had accuracies above a 50% level in the majority of blocks but there was variability in the number of blocks in which the accuracy was significantly above a chance level. Four participants had mean rank accuracies significantly above chance in all blocks. Eight participants had rank accuracies significantly above chance when averaged over blocks.

**Activation during explanation blocks**

The activation maps obtained during explanation blocks converged with the neural representations (stable voxel locations) during the thinking blocks. They indicated that there was a shift from more passive visual processing (watching how the system functions) to making inferences about causality and generation of hypotheses, and finally to a conceptual, perhaps motor representation (embodied cognition), a conclusion based on the following evidence. In the early stages before participants had been provided with any functional information, the preponderance of occipital activation suggests that participants simply scanned the pictures of the items. Similarly, the majority of the stable voxels before the first training stage were located in the occipital cortex. Occipital activation was prominent at all three stages, probably due to some combination of scanning visual depictions and reading the explanation sentences. The occipital activation is indicated by the green ellipses in Fig. 5. Fig. 5 shows the activation results in the three stages in the right hemisphere (all
activation results are significant at $T = 8.63$, 6 voxel extent, $p < .05$ family-wise errors corrected and appear in Table 1). After subjects had been presented with functional information about the systems, activation in explanation block 2 (given component + limited causal information) had four foci: right temporal sulcus including RTPJ, bilateral occipital gyrus extending up along IPS, bilateral inferior frontal gyrus (IFG), and bilateral pre-central gyrus. The frontal and pre-central activation was more extensive in the right hemisphere. The full set of activated clusters ($p < .001$ uncorrected, extent of 6 voxels) is shown in Table 1.

During the final two explanation blocks in which subjects were provided with the entire causal sequence, there was an indication of increased motor-related activity. There was increased pre-central as well as post-central/parietal activation. Also, frontal regions showed increased pre-central as well as post-central/parietal activation. Also, frontal regions showed

![Table 1](https://example.com/table1.png)

**Table 1**

Areas of activation for the time reading the sentences and examining the figures during the explanation sessions.

<table>
<thead>
<tr>
<th>Cortical region</th>
<th>Cluster size</th>
<th>Peak T-value</th>
<th>MNI coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$x$ $y$ $z$</td>
</tr>
<tr>
<td>A) Explanation block 1 (component information) as compared to fixation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilateral calcarine (occipital, fusiform)</td>
<td>1368</td>
<td>21.67</td>
<td>−10 −104 −4</td>
</tr>
<tr>
<td>Left fusiform (inferior occipital)</td>
<td>391</td>
<td>16.72</td>
<td>−38 −84 −8</td>
</tr>
<tr>
<td>Right hippocampus</td>
<td>25</td>
<td>6.1</td>
<td>30 −30 −2</td>
</tr>
<tr>
<td>B) Explanation block 2 (partial causal information) as compared to fixation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilateral calcarine (occipital, fusiform)</td>
<td>36346</td>
<td>30.26</td>
<td>−26 −92 20</td>
</tr>
<tr>
<td>Left inferior frontal</td>
<td>94</td>
<td>11.14</td>
<td>−60 18 10</td>
</tr>
<tr>
<td>Right medial frontal</td>
<td>49</td>
<td>10.43</td>
<td>12 32 46</td>
</tr>
<tr>
<td>Right superior temporal</td>
<td>40</td>
<td>10.27</td>
<td>58 −4 −8</td>
</tr>
<tr>
<td>Left orbital frontal</td>
<td>14</td>
<td>9.7</td>
<td>−46 44 −16</td>
</tr>
<tr>
<td>Left superior temporal</td>
<td>12</td>
<td>9.6</td>
<td>−66 −34 12</td>
</tr>
<tr>
<td>Left superior frontal</td>
<td>16</td>
<td>9.58</td>
<td>−14 50 48</td>
</tr>
<tr>
<td>Right middle frontal</td>
<td>6</td>
<td>8.83</td>
<td>40 50 32</td>
</tr>
<tr>
<td>C) Explanation block 4 (full causal information) as compared to fixation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left supramarginal (superior temporal, insula)</td>
<td>40542</td>
<td>26.35</td>
<td>−64 −34 36</td>
</tr>
<tr>
<td>Right supramarginal</td>
<td>47</td>
<td>11.45</td>
<td>66 −44 32</td>
</tr>
<tr>
<td>Right orbital frontal</td>
<td>33</td>
<td>11.08</td>
<td>40 48 −12</td>
</tr>
<tr>
<td>Right occipital</td>
<td>41</td>
<td>11.02</td>
<td>36 −96 8</td>
</tr>
<tr>
<td>Left inferior parietal</td>
<td>8</td>
<td>10.68</td>
<td>−56 −40 52</td>
</tr>
<tr>
<td>Left superior frontal</td>
<td>14</td>
<td>10.27</td>
<td>−18 44 46</td>
</tr>
<tr>
<td>Left occipital</td>
<td>9</td>
<td>10.18</td>
<td>−30 −98 16</td>
</tr>
<tr>
<td>Left middle temporal</td>
<td>7</td>
<td>10.03</td>
<td>−60 −64 6</td>
</tr>
<tr>
<td>Left middle frontal</td>
<td>12</td>
<td>9.98</td>
<td>−42 30 44</td>
</tr>
<tr>
<td>Left orbital frontal</td>
<td>8</td>
<td>9.84</td>
<td>−28 42 −10</td>
</tr>
<tr>
<td>Left occipital</td>
<td>19</td>
<td>9.55</td>
<td>−22 −80 8</td>
</tr>
<tr>
<td>Right occipital</td>
<td>22</td>
<td>9.20</td>
<td>46 −84 16</td>
</tr>
<tr>
<td>Right inferior occipital</td>
<td>7</td>
<td>8.81</td>
<td>32 −90 −10</td>
</tr>
</tbody>
</table>
more activation at this stage (middle and superior frontal gyri), possibly indicative of more goal directed or strategic processing. That this pattern of activation changes over stages further indicates a shift away from a passive viewing of the system to an interactive imagining of how the system would function under the individual’s control and conceptualizing the causal bases of the functioning.

Representational similarity within stable voxels

RSA revealed that the cortical similarity changes can be traced back to the information provided during the explanation blocks. When the instruction on two systems described some properties that were shared between them, the pairwise distance between those two systems (1 – r, where r is the correlation of the activation vector of stable voxels) decreased in the thinking session that followed. For example, the first instruction period for the disc brake system and the extinguisher system both referred to the concepts of pressure and force (brake: The increased pressure in the chamber forces a second piston against the brake pads; fire extinguisher: The difference in pressure forces the water out of the nozzle). The neural representation distance between these two systems in the temporal lobes decreased following the similar instruction. Similarly, the instructions describing the bathroom scale and disc brake system overlapped in the second instruction period in which bodily force was mentioned (scale: The person’s weight exerts a downward force on a lever; brake: Pressing the pedal pushes a piston in a fluid-filled chamber.) Thus, there is a decrease in neural representation distance between the two systems following commonality of words and the concepts they represent in the instructions, but no change in similarity in later thinking blocks. The temporal lobe stable voxels were explored because this region regularly activates during the comprehension of written language (Bookheimer, 2002; Ni et al., 2000).

In the three cases in which there was overlap of linguistic content in the instructions, the neural representation distance between the systems with shared descriptions decreased in each case. Fig. 6 indicates how two sets of systems decrease in distance (and increase in similarity). Although there were six possible pairs of systems, these two pairs of systems (pair 1: brake and fire extinguisher; pair 2: brake and scale) and instruction points were analyzed because they were the only cases in which there was a measureable overlap in instructions (in terms of shared physics concepts and vocabulary) across items (in total, only 12 words were repeated across mechanical systems). Overall, there was a decrease in distance across time for these systems in the temporal lobe stable voxels ($F_{(2,22)} = 6.55, \text{MSE} = 0.099, p < .01$). The hypothesis that changes in similarity in the three pairs with conceptual overlap in their instructions were different than the one pair with no conceptual overlap (brake and bathroom scale, time point 2 and time point 3) was tested using a planned contrast; this contrast was significant ($F_{(1,11)} = 4.79, \text{MSE} = 0.017, p = .05$).

Discussion

Overview

This research indicates the progression of neural changes the brain undergoes during the acquisition of knowledge of mechanical systems. The primary finding is that a sequence of knowledge states can be inferred by examining the snapshots of an evolving cortical representation as well as the activation during instruction. We propose that evidence from the domain of mechanical learning suggests a sequence of neurally-identifiable knowledge states, a sequence which may be general to other technical and scientific domains, such as learning in physics, biology, neuroscience, etc. Additionally, this research provides an early glimpse of how mechanical concepts are neurally represented, with the activation locations providing face-valid clues about the content of the representation. Furthermore, the similarity relations among the neural representations of different systems are driven in part by the similarities among the system descriptions.

In the case of the instructional materials used here, the neural knowledge representations evolved through distinct states, including an initial state before instruction and three additional states following three types of instruction. The brain locations that responded consistently to the items during test (thinking) trials (as well as the activation locations during explanation trials) provided evidence for the following states of the representation of knowledge:

0. Pre-training state (e.g. thinking about bathroom scale: a base visual component [largely occipital] and as well as a goal-directed motion component [right temporal sulcus];

1. Early-intermediate state (following a description of the components of the system): an imagery-based representation (parietal and superior frontal) of how the components of a system might move;

2. Late-intermediate state (following incomplete causal information): a multi-process state which involved inferring hypotheses about the causal relations of the components in the system (diverse activation across many non-occipital regions, in particular RTPJ and medFG);

Fig. 6. Mechanical systems become neurally more similar following the presentation of commonalities in the explanations (significant decreases are indicated by red arrows. Error bars indicate the mean square error of the interaction term).
3. **Final state** (following complete explanation of causal chain): a representation of how one might interact with the system to cause its motions (motor cortex and frontal cortex).

These four states of learning-based representations overlapped each other. Some regions, such as the motor cortex, played a role in all states (though the role was strongest in the final state), whereas other regions were closely tied to a single state. For example, the number of stable voxels in the parietal lobe was significantly greater after component information had been presented. These states were identifiable by changes in the spatial distributions of the stable voxels across the set of mechanical systems. Neural signatures comprised of these voxels can be used to distinguish items. The sequence of states may generalize to the learning sequence that leads to the understanding of causality in physical systems.

We begin with a discussion of the various states of representation based on the stability maps. The simple fact that the systematically responsive locations in the cortex change during learning is unique and novel. In addition, we use “reverse inference” to attribute cognitive functions to various brain regions. That is, based on prior neuroimaging findings in which the regions of the stable voxels have been shown to activate, we attribute particular cognitive functions to various brain regions, acknowledging the attendant uncertainty of such attributions. We have prefaced the reverse inference portions in the text with an “interpretation” label. The web-based Neurosynth database (Yarkoni et al., 2011) was used to relate locations of stable voxels to possible functions. This large scale database has the potential application to “draw more rigorous reverse inferences when interpreting results by referring to empirically established mappings between specific regions and cognitive functions; and extract the terms most frequently associated with an active region or distributed pattern of activity, thus contextualizing new research findings based on the literature” (Yarkoni et al., 2011). This exploratory analysis is a precursor to future planned studies that vary the learning task properties and permit less uncertain conclusions.

**Modifying the neural representation of mechanical systems via instruction**

**Pre-training state: visual and goal-directed regions**

Prior to training, the neural representation as measured by stable voxels in the first thinking block had a decidedly visual nature. The locations of the stable voxels in the untrained representation, occipital and right temporal sulcus, suggest that this representation contained a visual representation of the systems and how they would move during their operation. The right temporal sulcus, stretching from anterior temporal to RTPJ, has been associated with processing of goal-directed motion and biological motion (Carter et al., 2011; Peelen et al., 2006; Pelphrey et al., 2003, 2005). Interpretation: Thus, the first knowledge state contains information about spatial layout and possible motion.

**Early intermediate state: imagery-related regions**

The second thinking block indicated a visualization of the system components moving to perform the goal of the system. Three types of evidence support this conclusion: stability maps that were primarily composed of occipital cortex and IPS (visual imagery-related) voxels, activation maps during the first stage of instruction, and a high accuracy in classification of the systems from visual voxels. After the components of the systems were identified/described in the first explanation block (e.g., for the bathroom scale, they were informed of a ratchet and a dial that they could see were connected), resulting stability maps included a parietal (and IPS) component in addition to the occipital voxels. This region has repeatedly been shown to be active during mental imagery tasks (Deiber et al., 1998; Ishai et al., 2000; Just et al., 2004; Kosslyn et al., 1998).

A second notable pattern is the decrease in the percentage of stable voxels from the occipital cortex. The prominence of visuo-spatial processing gradually decreased during learning as the prominence of causal relations processing increased. A similar early learning decrease in visual attention (attributed to familiarity) has been previously shown in a task described as visuo–motor transformation, during which participants had to learn the mapping of a joystick movement to a visual cuing (Graydon et al., 2005).

**Interpretation.** The IPS stable voxels may have been coding imagery-based chunking and transformations of the components. This region has been seen to be active in imagined geometric transformations of objects such as a rotation of a 3D clock around an axis (Just et al., 2001). A key point is that although all explanation blocks directed the participants’ attention to the visual representation, the first block specifically targeted the components without describing any causal transformation. This may have resulted in participants ascribing causality by intuitively imagining the components moving. This proposed mental animation is consistent with behavioral evidence that early in the learning of mechanical systems eye movements are focused on the components of mechanical systems (e.g., pulleys) but the units are grouped into larger sub-assemblies later. Thus with familiarity comes an increasing focus on systems as a whole (Hegarty, 1992). This partitioning of the learning process illuminates the early learning state focused on visual transformations of component units involving IPS and the subsequent causality processing, leading to a sequenced understanding of the mechanical systems that occurred in later states. This hypothesis is consistent with evidence from neuropsychological literature that indicates ideational apraxia patients with right-sided or diffuse brain damage (primarily in the frontal lobe and potentially in the parietal lobe) have problems with multi-step actions involving technical devices such as coffee makers (Hartmann et al., 2005).

**Late intermediate state: causality regions**

The third thinking block indicated active generation of causal hypotheses of how something works as it is being explained. The introduction of causality information in explanation block 2 evoked a diverse network supporting both a continuation of the imagery-based mental animation as well as inference related regions (suggesting that learners engaged in a type of active causal hypothesis generation). As noted in the methods section, the second explanation block consisted of an incomplete description of how the systems moved to achieve the goal. Consider the given information for the bathroom scale example (the person’s weight exerts a downward force on a lever). The participant has to understand that force downward on a lever has a consequence on the end of the lever that is connected to a spring. The participant then can infer that the spring is stretched.

**Interpretation.** This type of postulated active causal hypothesis generation could account for the presence of mediFG-RTPJ stable voxels (see Fig. 3). These are regions that activate during the reading of narratives that require the reader to make inferences (i.e., mentally add information to the text that is not present to maintain causal coherence). This type of processing has also been associated with activation in language regions, such as bilateral IJG and posterior temporal gyr (Ferstl et al., 2008; Mason and Just, 2004, 2006; Virtue et al., 2008). Although some of the regions whose role in the neural representation shows increases in later stages are part of the default mode network (DMN) (Raichle et al., 2001), their increases are far more likely to be attributable to increased text comprehension and learning rather than to increased DMN activity. Some of the key components of the default-mode network (RTPJ and mediFG) are typically activated (compared to a rest baseline) during text comprehension tasks (see Ferstl et al., 2008; Mason and Just, 2009 for overviews). It is likely that the increasing
Interpretation.
The increase in stable voxels in the right pre- and post-encoding how to interact with the system. This supports an embodied systems suggests that understanding how the components of a me-

46
turbation of the mechanical systems (presented in 4.3.2) suggests what

activation in these regions during the later explanation blocks reflects the processing of increasing lengths of texts.

In addition to the regions associated with causal coherence process-

In the final state, the locations of the stable voxels included indication of knowledge of how to interact or use the mechanical system. The frontal lobe voxels appeared in superior, middle, and inferior gyri in both hemi-

Interpretation. The increase in stable voxels in the right pre- and post-

central gyri in the final thinking trials is consistent with activation in motor areas while imagining motor interaction with an object (Just,

Interpreting the stable voxel analysis, another subset of the stable voxels suggests that the representation at this stage was in a transitory period. This snapshot of the neural knowledge indicated it was composed of voxels found in the early inter-

Final state: motor/embodied regions

In the final state, the locations of the stable voxels indicated inclusion of knowledge of how to interact or use the mechanical system. The frontal lobe voxels appeared in superior, middle, and inferior gyri in both hemi-

In addition to the regions associated with causal coherence process-

Anatomical and functional overlaps among mechanical systems.

Interpretation. The increase in stable voxels in the right pre- and post-

central gyri in the final thinking trials is consistent with activation in motor areas while imagining motor interaction with an object (Just,

Interpreting the stable voxel analysis, another subset of the stable voxels suggests that the representation at this stage was in a transitory period. This snapshot of the neural knowledge indicated it was composed of voxels found in the early inter-

Generalizations and limitations

This cortically-grounded account integrates cognitive learning theo-

There are several provisos concerning these conclusions above. First, while the stable voxels maps indeed show evidence of a changing knowledge state, the attribution of each knowledge state (cortical map) to a set of cognitive processes involves reverse inference, as men-

Common instructions increase item similarity

RSA indicated that instructions can affect changes in the neural rep-

Activation in these regions during the later explanation blocks reflects the processing of increasing lengths of texts.

In addition to the regions associated with causal coherence process-

Anatomical and functional overlaps among mechanical systems.

Interpretation. The increase in stable voxels in the right pre- and post-

central gyri in the final thinking trials is consistent with activation in motor areas while imagining motor interaction with an object (Just,

Generalizations and limitations

This cortically-grounded account integrates cognitive learning theo-

There are several provisos concerning these conclusions above. First, while the stable voxels maps indeed show evidence of a changing knowledge state, the attribution of each knowledge state (cortical map) to a set of cognitive processes involves reverse inference, as men-

Common instructions increase item similarity

RSA indicated that instructions can affect changes in the neural rep-

Activation in these regions during the later explanation blocks reflects the processing of increasing lengths of texts.

In addition to the regions associated with causal coherence process-

Anatomical and functional overlaps among mechanical systems.

Interpretation. The increase in stable voxels in the right pre- and post-

central gyri in the final thinking trials is consistent with activation in motor areas while imagining motor interaction with an object (Just,

Generalizations and limitations

This cortically-grounded account integrates cognitive learning theo-

There are several provisos concerning these conclusions above. First, while the stable voxels maps indeed show evidence of a changing knowledge state, the attribution of each knowledge state (cortical map) to a set of cognitive processes involves reverse inference, as men-

Common instructions increase item similarity

RSA indicated that instructions can affect changes in the neural rep-

Activation in these regions during the later explanation blocks reflects the processing of increasing lengths of texts.

In addition to the regions associated with causal coherence process-

Anatomical and functional overlaps among mechanical systems.

Interpretation. The increase in stable voxels in the right pre- and post-

central gyri in the final thinking trials is consistent with activation in motor areas while imagining motor interaction with an object (Just,

Generalizations and limitations

This cortically-grounded account integrates cognitive learning theo-

There are several provisos concerning these conclusions above. First, while the stable voxels maps indeed show evidence of a changing knowledge state, the attribution of each knowledge state (cortical map) to a set of cognitive processes involves reverse inference, as men-

Common instructions increase item similarity

RSA indicated that instructions can affect changes in the neural rep-

Activation in these regions during the later explanation blocks reflects the processing of increasing lengths of texts.

In addition to the regions associated with causal coherence process-

Anatomical and functional overlaps among mechanical systems.

Interpretation. The increase in stable voxels in the right pre- and post-

central gyri in the final thinking trials is consistent with activation in motor areas while imagining motor interaction with an object (Just,

Generalizations and limitations

This cortically-grounded account integrates cognitive learning theo-

There are several provisos concerning these conclusions above. First, while the stable voxels maps indeed show evidence of a changing knowledge state, the attribution of each knowledge state (cortical map) to a set of cognitive processes involves reverse inference, as men-

Common instructions increase item similarity

RSA indicated that instructions can affect changes in the neural rep-

Activation in these regions during the later explanation blocks reflects the processing of increasing lengths of texts.
References


Acknowledgments

This research was supported by the Office of Naval Research (Grant N00014-13-131-0250). We thank Tim Keller, Vladimir Cherkassky, and Andrew Bauer for help in data analysis and comments on earlier versions of the manuscript. We thank Chelsea McGrath for assistance in testing participants and running experiments. We thank the CCBI reading group for comments on an earlier draft of the article.