The Emergence of Explicit Memory during Learning

Michael Rose1, Hilde Haider2 and Christian Büchel1

1Department for Systems Neuroscience, NeuroImage Nord, University Medical Center Hamburg, 20246 Hamburg, Germany and
2Department of Psychology, EWF, University of Cologne, D-50931 Koeln, Germany

Address correspondence to Dr Michael Rose, NeuroImage Nord, Department for Systems Neuroscience, University Medical Center Hamburg Eppendorf, Martinistrasse 52, D-20246 Hamburg, Germany. Email: rose@uke.uni-hamburg.de.

In incidental learning situations, contingencies are extracted from the environment without the intention to learn and can change behavior without awareness for the extracted regularity. The development of explicit access to the learned regularity is an important learning mechanism that is rarely examined. With a series of behavioral, electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) studies, we were able to show that the emergence of awareness for a hidden regularity is accompanied by an increase in neural activity and in high-frequency coupling between distant brain areas as observed with a time-frequency resolved EEG analysis. More importantly, the increase in neural coupling was observed before awareness for the learned material was established behaviorally. In addition, coupling increases were paralleled by an fMRI-signal increase in the ventral striatum and the right ventrolateral prefrontal cortex directly preceding the emergence of awareness. The involvement of this system, which has already been linked to the processing of predictions and prediction errors, indicates the relevance of a reinforcement signal to generate awareness for the learned contingencies. Thus, our data provide direct evidence for the necessity of large-scale coupling and the evaluation of a predictive stimulus value as the basis for a transition from implicit to explicit memory.

Keywords: EEG, fMRI, learning, memory

Introduction

The transition from implicit to explicit memory during incidental learning offers an experimentally tractable paradigm to study neuronal processes underlying the development of a consciously available memory trace. This transition is an important process in order to adapt to changing environments (Koch et al. 2006) and can be observed in operant conditioning (O’Doherty et al. 2004). Initially, learning relies on implicitly formed associations, which is then followed by conscious awareness about the basis of what one’s decisions rely on. This is a very effective way to experience the complex contingencies in the environment and is assumed to be an important learning mechanism during development, when concepts of the world are often acquired incidentally. However, the pertinent question is how these explicit memories are acquired.

The observation of a sudden insight into underlying rules is a widely known finding in research on implicit learning (Seger 1994). During implicit learning, participants gradually adapt to the contingencies inherent in the task, mostly without being able to report anything about the underlying regularities. However, in nearly all implicit learning tasks, a variable number of participants also gains explicit knowledge about the underlying regularities. Furthermore, several findings suggest that this generation of explicit memories leads to a change in behavior and thus might be of great relevance for a rapid adaptation to a changing environment (e.g., Koch 2007; Tubau et al. 2007; Haider and Frensch 2009).

Despite the significance of these findings, the emergence of explicit memory in incidental learning situations and the neural basis for this transition process, in particular, is rarely examined. Theoretical models suggest a distinct cognitive process that initiates the transition from incidental learning to explicit memory formation (Runger and Frensch 2008), and the evaluation of the neural activity in relation to this process may provide evidence for the existence and the possible nature of this additional process.

Several studies have shown that the anterior lateral prefrontal cortex plays a role in situations in which the learning process is already explicit (Strange et al. 2001; Turner et al. 2004; Carter et al. 2006). In a previous study, we observed an increase in neural activity in the right ventral prefrontal cortex for the implicit detection of prediction errors during learning (Rose et al. 2005). This process is discussed as an important precondition for the development of explicit memory (Runger and Frensch 2008). An additional contribution of the striatum for the generation of predictions during learning is discussed for feedback-based learning (Shohamy et al. 2004; Bischoff-Grethe et al. 2009). Furthermore, there is evidence for an involvement of the ventral striatum (VS) in motivational learning (Seger 2008). In a previous experiment, it was suggested that the basal ganglia and the medial temporal lobe (MTL) interact during learning, depending on the implicit or explicit mode of processing (Poldrack et al. 2001). In this study, the MTL activity decreased, and the activity in the basal ganglia increased with implicit learning. However, in our own study, we found evidence for an involvement of MTL structures even in implicit learning of relational knowledge comparable with the present study (Rose et al. 2002). The present study further examines this area used to represent the regularity in order to directly test for possible changes due to the development of explicit memory.

Furthermore, studies on conscious awareness during learning (Mclntosh et al. 1999) and problem solving (Jung-Beeman et al. 2004) revealed the importance of interactions of the right prefrontal cortex with other areas to establish the neural correlate of the explicit aspect of the memory. In the actual study, we examined the neural activity and interactions in direct temporal relation to the emergence of explicit memory to test the hypothesis that frontal activity and interconnectivity increase in the time period directly preceding the emergence of explicit knowledge and therefore can be regarded as the neural basis for an additional process that is the basis for the generation of explicit memory.

For a detailed examination of the transition process from incidental learning to conscious knowledge, an objective measure of the development of awareness is necessary (Persaud et al. 2007). However, using confidence ratings or directly asking volunteers might bias participants to explicitly search for an underlying regularity (Dienes and Scott 2005).

Recent findings have revealed that participants, who develop explicit memories during incidental learning, show a dramatic rise in confidence ratings or directly asking volunteers might bias participants to explicitly search for an underlying regularity (Dienes and Scott 2005).
decrease in reaction time (RT) (Haider and Frensch 2005; Haider and Rose 2007), which indicates that the actual stimuli become less important for response preparation (Tubau et al. 2007). In these experiments, participants were instructed to solve tasks on the basis of simple explicit rules. Unbeknownst to the participants, the tasks also contained a hidden regularity. This hidden regularity, if explicitly noticed, allows participants to prepare their responses in advance to the occurrence of the stimuli resulting in a clear decrease of RT. The results demonstrated that explicit memory in these kinds of experiments correlated with an RT drop across several consecutive trials. In contrast, single fast responses in individual trials did not signal explicit memory.

The goal of the behavioral experiments therefore was to evaluate the reliability of this RT-drop as an indicator for the point in time when explicit memory emerges. If we can show that the RT drop reliably indicates the emergence of explicit memory, we should be able to investigate the neural events accompanying and, in particular, preceding this process of insight in subsequent functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) studies.

Materials and Methods

Eighteen healthy volunteers (age 19–36, 12 females) participated in the fMRI experiment; 19 volunteers participated in the EEG study (age 20–36, 9 females), and 77 healthy subjects were recruited for the behavioral studies (age 19–41, 49 females). All subjects were right handed and had normal or corrected-to-normal vision. The study was approved by the ethics committee, and all participants gave written informed consent prior to participating in the study.

Stimuli

The colored squares (0.6° × 0.6°; all 4 responses together 2.3°) were presented on a gray background controlled by a PC that ensured synchronization with the MR scanner using the software “Presentation” (http://nbs.neuro-bs.com). For the MR experiment, an LCD projector projected the stimuli on a screen positioned on top of the head coil that was viewed by the subjects through a mirror (10 × 15° field of view), and the stimuli for the EEG and behavioral studies were presented on a monitor. Participants entered the responses by pressing buttons on an MR-compatible device or on a keyboard with the right hand (EEG and behavioral studies).

Designs

In the employed experimental design, participants practiced a simple sequential task. This task was identical in all experiments. During the task, volunteers were shown pairs of colored squares (red, green, or blue) and had to simply compare the colors of 1 pair of squares in order to generate a response according to 2 simple rules (see Fig. 1). The first rule stated that the result of 2 identical colors is the same color (i.e., the first stimulus red and red = red). The second rule stated that the response to 2 different colors is the remaining of the 3 colors (e.g., in the second stimulus, green and red = blue). Each color response was assigned to a distinct button that should be pressed. Each stimulus pair was presented for a maximum of 3 s. It disappeared after the response was made, and the response appeared in the row below. Regardless of the RT to the previous stimulus pair, the next stimulus pair was always presented after an interval of 3 s. The response is presented in the row below the stimulus pair containing a total of 4 responses per task. Importantly, unbeknownst to the participants, a hidden regularity was embedded into this response row, such that the final response in each trial was identical to the first one (determined), whereas the other responses were unpredictable (undetermined). Thus, participants could incidentally learn this hidden regularity while performing the explicit color comparison task. The stimuli that the volunteer had to compare to make the final response varied across trials. Thus, although the response pattern always follows the hidden regularity, the presentation of the stimulus pairs was not regular (i.e., a response to red could result from a red-stimulus pair or from a green-blue-stimulus pair). As in our previous experiments, participants who gain an explicit insight in the hidden regularity should show an abruptly occurring decrease in RT for the final determined response in comparison with the undetermined responses.

The goal of the behavioral studies was to test whether this abrupt RT drop indicates the emergence of explicit memory. The aim of the fMRI and EEG studies was to evaluate the neural events that accompany and precede this process.

In all experiments reported, the interstimulus interval between consecutive stimulus pairs was 3 s, allowing for a separate estimation of the evoked fMRI signal for each input (i.e., convolution). Incorrect responses or no response within a 3-s window were considered as errors, and the complete trial was excluded from further analyses. Visual feedback was given to the participants in error trials only, and error trials were excluded from all analyses. The intertrial interval was randomized between 8 and 12 s for the fMRI study and between 1 and 3 s for the behavioral and EEG studies. Before the experiment started, participants were trained on the task using unstructured material (i.e., no hidden rule). One study block consisted of 30 trials. In the fMRI and EEG experiment, participants performed a total of 6 blocks.

Behavioral Studies

In all 3 behavioral studies, the reliability of the abrupt RT drop was examined by interrupting the experiment according to the amount of fast responses to the determined input and assessing the degree of explicit knowledge (Table 1). If none of the criteria was reached, the participants also performed...
a total of 6 blocks like in the fMRI and EEG studies. To detect a reliable fast response to the determined inputs, the RTs were measured online, and the determined inputs were compared statistically against the undetermined inputs. This was done by a method identical in all experiments (online for the behavioral studies and after the experiment for the fMRI and EEG experiment).

For each participant, the time series of RTs across trials for each input (undetermined inputs 1–3 and determined last input) were filtered using a moving median filter (lag 1, width 5, and calculation of median). Furthermore, in each 5-trial-long window of the moving median filter function, confidence intervals (90%) for the undetermined inputs (inputs 2 and 3 only) were estimated. The mean confidence interval of both undetermined inputs was used for a statistical comparison to define the starting point of the RT drop as that point, where the median filtered RTs of the determined input position significantly deviate from the RTs of the undetermined input position. Thus, the confidence interval estimated from the moving median filter was adaptive to general changes in RTs across time, such as habituation or unspecific learning, and the lower confidence value reflected the fast RTs of the undetermined inputs within the specific time interval. If a single RT (median filtered) to a determined input fell below the confidence interval of the undetermined inputs, the RT was counted as significantly faster.

In Experiment 1, the task was interrupted when the participant’s responses to the determined input were significantly faster than the confidence interval in 3 successive trials (later referred to as triple RT decrease [TRT]). Thus, the starting point of the TRT was defined as that point when 3 consecutive responses to the determined input position fell below the confidence intervals of responses to input positions 2 and 3. In Experiment 2, the task was interrupted after each single fast response and then continued until the TRT criterion was observed.

It could be argued that explicit memory is already established after the first fast response, and therefore, Experiment 2 directly assessed the relevance of single fast RTs. To compare the reliability of the TRT criterion against the occurrence of single fast responses, the training in Experiment 2 was interrupted after a participant had exhibited a total of 10 fast responses to the determined input, irrespective of whether or not these fast responses occurred in successive trials. Thus, here explicit knowledge was assessed after the occurrence of 10 fast trials that did not result in a constant RT drop.

**Imaging**

In the fMRI experiment, the participants performed 6 blocks (with 30 trials each), and explicit knowledge was assessed after the whole experiment. For the participants who were able to verbalize the hidden rule, the RTs were analyzed offline with the method described above to detect the start of the constant RT drop (TRT criterion). In each participant, the first fast RT within this constant RT drop was taken as the reference point for the boundaries of the experimental parts where the blood oxygen level-dependent (BOLD) signal was contrasted statistically. Beginning with this first fast RT, neuronal activity across the next 10 trials was analyzed to detect changes that accompany the emergence of explicit knowledge. The 10 trials before the first fast RT were analyzed to examine neural activity related to the initiation of explicit memory, thus in a time range in which the behavioral experiments demonstrated nearly no occurrence of explicit knowledge. All trials preceding this time period were combined and labeled as the Start period and all trials following the 10 trials after the TRT were subsumed to the End period.

fMRI was performed on a 3-T system (Siemens Trio) with a gradient echo-planar imaging (EPI) 

T2-sensitive pulse sequence in 38 axial slices (2-mm thickness with a 1-mm gap, time repetition 2.2 s, time echo 25 ms, flip angle 90°, field of view 192 × 192 mm², and matrix 64 × 64). SPM2 (http://www.fil.ion.ucl.ac.uk/spm) was used for image processing and statistical analysis. All volumes were realigned to that first acquired, spatially normalized to the standard EPI template, and smoothed with a 12-mm full-width at half-maximum isotropic Gaussian kernel.

The fMRI data were analyzed by an estimation of the BOLD signal for each single trial and each input in each trial and modeled as a hemodynamic response function (HRF). Error trials were modeled as a separate regressor and were excluded from further contrasts. Using a general linear model, regression coefficients were obtained for each regressor (trial). In each subject, contrasts for the different time periods were calculated with respect to the individual RT decrease. Therefore, the regressor weights of the individual trials were combined by the contrasts against the individual time point of the emergence of explicit memory (TRT criterion). Thus, in each subject, contrasts for the four different time periods (referred to the individual TRT) were estimated, and all contrasts were entered into a group analysis. To detect changes with learning for the determined inputs within the time period of the emergence of explicit memory (10 trials after the RT decrease), this time period was compared with the Start period. To test the specificity of this effect, it was compared against changes evoked by the undetermined inputs. This was statistically implemented at the level of the group analysis as an interaction contrast testing for an increase of activity from the Start period to the period 10 trials after the TRT specific for the determined input (factor time period × determined/undetermined, corrected for multiple comparisons, False Discovery Rate (FDR) whole brain).

The effects within the time period of 10 trials before the RT decrease were tested using a comparable interaction effect comparing the activity during Start period with the period 10 trials before the TRT between the determined and undetermined inputs (corrected for multiple comparisons, region of interest analysis, 20-mm sphere centered around regions of interest: VLPFC according to Turner et al. 2004, striatum according to Bischoff-Grethe et al. 2009).

Two additional analyses were performed to address possible changes in the area involved in the representation of the learned material. Therefore, 1 contrast tested for an increase of the BOLD signal of the determined inputs from the Start period to both periods around the TRT (10 trials before and after the TRT) and another contrast for a decrease. Both contrasts again were modeled as an interaction effect between time periods and the determined and undetermined inputs (corrected for multiple comparisons, region of interest analysis, 20-mm sphere centered around regions of interest MTJ according to Rose et al. 2002).

**EEG**

In the EEG experiment, the participants also performed 6 blocks (with 30 trials each). The explicit knowledge was assessed after the whole experiment, and the TRT criterion was determined in each participant with explicit memory. Using the individual TRT occurrence, the EEG was assigned to comparable time periods as in the fMRI study (Start, 10 before the TRT, 10 trials after the TRT, and End).

EEG was recorded from 57 channels (BrainVision System), impedance <5 kΩ, low cutoff of 0.03 Hz, high cutoff of 250 Hz, 500-Hz sampling rate, referenced to the nose. Vertical and horizontal electrooculographs (EOGs) were recorded from above versus below the left eye and from the outer canthi of the eyes, respectively. Data segments (~800 until 1200 ms around the appearance of a stimulus pair) were extracted for correct trials only, using the individual TRT and the 4 different time periods, checked for artifacts, and decomposed using a time-frequency analysis based on multitapers (Fieldtrip, www.ru.nl/fcdonders/fieldtrip/; window length 256 ms, spectral concentration over 8 Hz). To correct for the bias due to different numbers of trials, a z-transform was applied (Jarvis and Mitra 2001). Each resulting spectrum was normalized by subtracting the baseline...
values (~300 to ~100 ms). To calculate a global index of synchronization, coherence spectra within the time-frequency domain were estimated for all possible channel combinations, and the mean overall spectra were calculated. Single-subject z-transformed mean coherence values were extracted in the frequency range between 35 and 45 Hz (450–550 ms) and 5–20 Hz (300–400 ms). Statistical comparisons were based on interaction effects between undetermined and determined inputs and between different time periods (time ranges with respect to the TRT) comparable with the fMRI contrasts. Of most interest here was the evaluation of an increase of synchronization from the Start period to the period of 10 trials before the TRT limited to the determined inputs.

Results

Behavioral Experiments

In all experiments, the development of explicit insight into the regularity of the determined last response resulted in a clear and abrupt decrease of reaction times (RTs) for the final stimulus pair due to anticipating the correct response even before the stimulus had appeared (Fig. 2). In 3 behavioral experiments (Experiment 1: \( n = 22 \), Experiment 2: \( n = 24 \), and Experiment 3: \( n = 31 \)), we evaluated the predictive value of this decrease in RT for the development of explicit memory in further detail.

Although participants practiced the task, their RTs were monitored online, and a confidence interval of the undetermined inputs was estimated (see the Materials and Methods section). The confidence interval was used to detect those RTs to the determined last inputs that were significantly faster than the RTs to the undetermined input position during the ongoing experiment. Each reliably fast response to the determined input position was counted, and training was interrupted when the number of fast responses reached a predetermined criterion that differed in both experiments, and the status of explicit knowledge was assessed. In Experiment 1, the task was interrupted if participants produced 3 fast responses to the determined input position in immediate succession. Reanalyses of former experiments had revealed that participants showing an abrupt RT drop during training never produced more than 1 or 2 fast responses in immediate succession before the drop occurred (Haider and Rose 2007). Therefore, we assumed that this criterion might reflect the starting point of the

Figure 2. Behavioral results: (a) RT data from the fMRI experiment for a representative subject. RTs for the determined last input are shown in red, RTs for undetermined inputs in gray. The RT decrease (TRT) for the determined last input (red) as compared with the undetermined responses (gray) can clearly be seen (around trial 48). The time point of a significant difference between the RTs was statistically verified using the confidence intervals of the undetermined RTs (dotted lines) and was determined for each participant individually. (b) Group-averaged RTs for the determined (red) and undetermined (gray) inputs before (left) and after (right) the occurrence of the RT decrease for the determined inputs (error bars represent the standard error). The reliable difference between the input categories showed that the RTs to the determined input are faster for the whole time series following the TRT.
abrupt and robust decrease in RT. If so, participants showing 3 fast responses in immediate succession should be able to deliberately describe the underlying hidden rule. The status of explicit knowledge was assessed directly after the training had been interrupted. The participants were first asked, whether they noticed something special in the task and if yes, they were asked to explain it. If they did not start to describe the regularity on their own, they were told that a regularity was built into the tasks and were asked to describe it and to write down examples. The memory was categorized as explicit if a participant was able to deliberately verbalize the hidden regularity. In all other cases, participants’ knowledge was classified as implicit. The results indicated that 3 fast responses (compared with the undetermined inputs) in successive trials (TRT) were a reliable indicator for a robust and lasting RT decrease and the development of explicit knowledge. At this point in time, 88% of the participants who reached the criterion of Experiment 1 \((n = 17)\) were able to deliberately verbalize the hidden rule, whereas 70% of the participants who did not reach the criterion \((n = 7)\) could not report anything \((\chi^2(1) = 71.69, P < 0.001)\). Thus, nearly all participants immediately interviewed after the occurrence of 3 fast successive responses were able to explicitly name the hidden rule. By contrast, nearly all participants who never reached this criterion during training were unaware of the presence of a structural regularity. Only 2 of the 7 participants who did not reach the criterion of Experiment 1 mentioned that there was a regularity in the material but were not able to further describe the rule. For those participants, it remained unclear whether they really detected a regularity or whether the interview induced the feeling of a regularity. Consequently, participants with such partial explicit knowledge were excluded from further EEG and fMRI analyses.

In summary, the results of Experiment 1 demonstrated that 3 fast responses in immediate succession (TRT) are a sufficient behavioral predictor for the development of explicit memory. However, an outstanding question was whether explicit knowledge develops in the course of training before the 3 fast responses occurred or whether the 3 fast responses indicate the point in time when explicit memory emerges. Therefore, the goal of Experiment 2 was to test these 2 alternatives with the procedure described above.

In Experiment 2, we additionally controlled for the role of single fast responses. To do so, we interrupted participants’ training after they had produced a total of 10 fast responses irrespective of the distribution of these 10 fast responses during training. If these single fast responses already result from explicit memory about the hidden rule, then these participants should also be able to deliberately name the regularity.

The results from Experiment 2 revealed that the predictive value of single fast responses was low. All 22 participants were interrupted after they produced 10 fast responses in nonsuccessive trials distributed over the entire course of training, but only 12 participants could report the hidden rule (55%). A comparison of the distribution of the fast responses in Experiment 2 between volunteers with and without explicit memory showed that participants with explicit memory produced significantly more fast RTs within the last 10 trials before training was interrupted \((4.2 \text{ vs. } 1.3 \text{ responses}; t(20) = 2.74, P < 0.05)\).

Experiment 3 was designed to directly evaluate the existence of explicit processes in the time period preceding the RT criterion (TRT). It can be assumed that explicit memory develops gradually over a broader time range before behavioral changes. To test this possibility, the training was interrupted whenever the determined response in a single trial was faster than the undetermined response, using the online estimated confidence interval. Then, the participant was asked for the reason of the fast response. The participant had to select 1 possible response out of several given answers (‘‘not noticed,’’ ‘‘a possible mistake,’’ ‘‘practice effect,’’ or ‘‘knowing the correct answer in advance’’) and then the experiment continued. As before, the experiment ended when the RT criterion (TRT) was reached or after 6 blocks and the amount of explicit memory was assessed \((n = 23)\) reached the criterion, 87% of these participants were able to deliberately verbalize the entire hidden regularity after the experiment; \(n = 8\) did not reach the criterion and were also not able to verbalize the hidden regularity. Before the occurrence of the TRT, participants were interrupted in 28% of all trials and responded in 7% of all trials that they knew the correct answer in advance. Neither the rate of fast responses nor the rate of answers—possibly indicating explicit knowledge—differed between groups \((F(29) = 1.35 \text{ and } T(29) = 1.27, \text{ respectively; } P > 0.15)\). Importantly, although the questions may have induced a search process different from all other experiments, the results clearly demonstrated that before the occurrence of the TRT only few of the responses suggested that participants had some explicit knowledge about the hidden rule. By contrast and most interestingly, participants reaching the criterion responded in 92% of the TRT trials that they knew the answer in advance. For these participants, we additionally analyzed the 10 trials before the TRT. Here, only 11% of the responses indicated probable explicit knowledge. Thus, the results of Experiment 3 indicate a strong correlation of the TRT and the sudden emergence of explicit memory.

Overall, the behavioral results of Experiments 1–3 indicate that 1) the development of explicit knowledge is reliably signaled by an abrupt decrease in RT (TRT), and 2) awareness for the regularity starts to emerge at this point in time. The results from Experiment 3, in particular, demonstrated that the TRT signals the emergence of the explicit knowledge with a high temporal precision. As a consequence, the mechanisms that initiate the emergence of explicit knowledge should precede this abrupt decrease in RT, and the results from Experiment 3 suggested that an important time period starts around 10 trials before the RT decrease. Based on these behavioral results, we were in a position to investigate the neuronal basis of this transition process from incidental learning to explicit memory using fMRI and EEG. We used the design of the behavioral experiments (Fig. 1). However, instead of using the prespecified criteria for interrupting the training within a session, the experiment was finished if a participant deliberately named the regularity during the break between 2 successive blocks, and the RT time series was analyzed offline, or the experiment ended after 6 blocks.

**fMRI Experiment**

Overall, 10 participants (55%) stated during the break between successive blocks that they have noticed a hidden rule in the preceding block and were able to deliberately describe the rule. The error rate was very low (6%), and error regressors were not included in the analyses. No reliable change of error rate was detected across the sessions. Analyses of their performance revealed that they also reached the RT criterion evaluated in the behavioral experiments (TRT). After having reached this criterion, RTs to the determined input position remained consistently faster than RTs to the undetermined inputs in both the fMRI and EEG experiments (repeated measures analysis of variance [ANOVA], interaction: input by time [before/after RT decrease]; MR study: \(F(3,27) = 100, t^{2} = 3.4\)).
To dissociate the RT decrease from implicit learning effects, the RTs after the TRT were compared with the RTs of the last session of the implicit group. A reliable interaction of group by input (fMRI study: \( F(3,24) = 86.9, P < 0.001; \) EEG study: \( F(3,24) = 40.2, P < 0.001 \)) demonstrated that with explicit knowledge, the determined position is processed faster than with implicit memory. This further demonstrated that the TRT signaled an additional process that is specific for the development of explicit memory.

Thus, the personal report together with the behavioral data allowed a precise temporal prediction of the emergence of explicit knowledge. This temporal information was then used in a single-trial analysis of fMRI data to evaluate the neural responses in different brain areas in relation to the transition process. For each participant, the whole time series of learning was divided into 4 segments of interest around the individual time point of the TRT (see Fig. 3). From our behavioral studies, it was concluded that the first 10 trials of the RT drop are the trials when explicit memory has emerged. Within that time segment, a clear increase of neural activity in several brain areas, including bilateral VS, frontal, parietal, and anterior cingulate cortex was observed (Table 2). The increase in the BOLD signal was limited to the processing of the determined input position and was time locked to the individual occurrence of the TRT.

Most interestingly, individually determining for each participant the point in time when explicit memory starts to emerge, additionally enabled us to investigate changes in neuronal activity that preceded the transition to explicit knowledge; that is, in a time period when the following transition process is likely to be initiated with no behavioral or attentional confounds (Koch and Tsuchiya 2007). Therefore, the time period of 10 trials before the first fast RT within the TRT was evaluated; thus, we focused the analysis on a time period with no reliable difference between the RTs of the determined and undetermined inputs. Within the 10 trials before the TRT, activity increased within the ventrolateral striatum and the right ventrolateral prefrontal cortex (VLPFC), indicating the relevance of these structures for the development of explicit insight into the hidden rule. Again this effect was limited to the determined inputs.

To evaluate the specificity of the BOLD-signal increase for the development of explicit memory, we compared participants with and without explicit memory. First, the determined input was compared within the first block between both groups to test for possible differences starting from the beginning of the experiment. No region showed a reliably higher BOLD signal for the explicit group (threshold \( P < 0.05, \) FDR whole brain). This result indicated that the process relevant for the development of explicit memory develops over time (i.e., with learning). To test

\[ \text{Figure 3.} \quad \text{Signal changes preceding the emergence of explicit memory.} \]

In the VS (a, left, maximum x, y, z = 12, 9, −3 mm) and the right VLPFC (b, right, maximum x, y, z 51, 45, −6 mm) the fMRI signal increased even before the TRT. In the middle row (c–e) data from a representative single subject are shown. The bottom row (f,g) shows the result for the whole group. For the individual data, the RTs (e) and the BOLD signal changes across single trials (c–d) for the undetermined second input (gray) and to the determined input (red) are plotted (filtered using a median function (lag 1, width 5)). The line at trial 85 represents the time point of the significant RT decrease (TRT). Importantly, the BOLD signal shows a significant increase 10 trials antecedent. The group data (f,g) show the mean BOLD signal averaged within the different time periods referenced to the TRT. It can be seen that the signal increase is limited to the determined inputs (red bars) as compared with the undetermined inputs (gray bars) and already started in the trials preceding the TRT (error bars represent the standard error).
whether the increase of the BOLD signal to the determined inputs was specific for the explicit group, the increase in 10 trials before the TRT was compared against determined inputs of the implicit group within the last block. One could assume that the BOLD signal to the determined inputs also increased with implicit learning. In this case, the fMRI signal to the determined inputs should also increase in the implicit group and should be largest in the last block. The increase was calculated as the difference to the undetermined inputs 10 trials before the TRT in the explicit group and within all trials of the last block for the implicit group.

The mean BOLD signal difference 10 trials before the TRT was significantly higher in the group that developed explicit memory than the BOLD signal difference in the last block of the implicit group (striatum: $t(16) = 4.5$, $P < 0.01$; right VLPFC: $t(16) = 5.7$, $P < 0.05$).

The additional analysis that used both time periods around the TRT revealed an increase in the BOLD signal for the determined inputs within the right MTL ($t(30), p = -21, z = -15$, $t(65) = 3.86$, $P < 0.05$, corrected). In contrast, the second additional contrast did not reveal any reliable decrease of the BOLD signal in these time periods in the whole brain.

**EEG Results**

Again, only participants with a reliable RT criterion (TRT) were analyzed ($n = 10; 53\%$), and the EEG of these participants was again segmented into the 4 time periods of interest with respect to the starting point of the TRT. The error rate was low (4%), and errors were not included in the analyses. No reliable change of error rate was detected across the sessions.

In addition to an increase in neural activity as observed in the fMRI data, theoretical models (Dehaene et al. 2003; Baars 2005) suggest an increase in the long-range information exchange of distributed brain areas, resulting in the general availability of the information acting as a basis for the explicit nature of information.

To obtain a direct measure for large-scale functional coupling of neural activity between several areas, we examined the coherence of EEG signals between different brain regions in an additional experiment. Because we were first of all interested in the overall long-range coupling of neural activity, we estimated a global index of synchronization (Uhlhaas et al. 2006). Therefore, all segments of the different time periods were transformed in the time-frequency domain. Next, coherence spectra for all possible channel combinations were estimated, and the mean overall coherence spectrum was used as the global index of synchronization. This global measure shows an increase in overall synchronization in the time period within 10 trials directly preceding the TRT. The mean coherence evoked by the processing of the determined inputs increased within this time period but not for the undetermined inputs. Between 300 and 800 ms after stimulus onset, coherence in the gamma band (35–45 Hz) and in lower frequencies (LFs, 5–20 Hz) increased during the time period of 10 trials before the TRT (repeated measures ANOVA, interaction of input position (determined vs. undetermined) by time period for the gamma-band coherence: $R(1,9) = 6.6$, $P < 0.05$, for the LF: $R(1,9) = 7$, $P < 0.05$). The coherence increase in lower frequencies was accompanied by a comparable increase in spectral power ($F(1,9) = 6.2$, $P < 0.05$), whereas in the gamma band, no reliable power changes were observed ($F(1,9) = 0.4$, n.s.). An increase in gamma-band coherence in the absence of changes in power underscores the fact that the gamma-band coherence effect relies on a true synchronization of neural activity. Recently, it was suggested that induced gamma-band responses reflects properties of miniature saccades (Yuval-Greenberg et al. 2008). To control for this possible confound, gamma-band power of a horizontal EOG channel of the identical time range was compared for the determined input within the first time periods. No difference was found for the gamma-band power of the EOG channel ($F(10) = 1.8$, $P = 0.1$) indicating that the effect is not related to possible miniature saccades.

To evaluate the topographic distribution of the gamma-band coherence increase, the mean coherence from each electrode to all other electrodes was calculated, and the difference between the time periods was estimated. In accord with the fMRI results, the topographic distribution of gamma-band coherence increase was maximal over right frontal electrodes. We further estimated coherence from that location (maximum at electrode AF8) to all other electrodes and again observed a maximal increase of gamma-band coherence over neighboring right frontal electrodes (Fig. 4). Thus, although there was a widespread increase in gamma-band coherence, the synchronization changes over the right prefrontal cortex were maximal.

---

**Table 2**

Maxima of fMRI signal changes during the 10 trials after the RT abruptly started to decrease and within the 10 trials preceding the TRT

<table>
<thead>
<tr>
<th>Coordinates (mm) and magnitudes of main activations</th>
<th>10 Trials before (interaction inputs × time period, $P &lt; 0.05$, corrected)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T(XY Z T)</td>
</tr>
<tr>
<td>Left ventral striatum</td>
<td>5.02</td>
</tr>
<tr>
<td>Right ventral striatum</td>
<td>4.04</td>
</tr>
<tr>
<td>Right VLPFC</td>
<td>5.06</td>
</tr>
<tr>
<td>Left VLPFC</td>
<td>4.65</td>
</tr>
<tr>
<td>Medial orbitofrontal cortex</td>
<td>4.98</td>
</tr>
<tr>
<td>Cingulate cortex</td>
<td>3.63</td>
</tr>
<tr>
<td>Left lateral parietal cortex</td>
<td>3.70</td>
</tr>
<tr>
<td>Parietal cortex</td>
<td>3.81</td>
</tr>
<tr>
<td>Left temporal</td>
<td>3.84</td>
</tr>
<tr>
<td>Right temporal</td>
<td>3.84</td>
</tr>
</tbody>
</table>

Both effects were tested as an interaction contrast between the time periods and the input category (determined against undetermined), that is, a steeper increase of the BOLD signal intensity across the training trials for the determined inputs.
These results show that in addition to the activity increase within the rPFC as shown in the fMRI experiment, this area is part of a dynamically assembled network, which subserves the development of explicit knowledge by participating in the synchronized coupling across many areas.

Discussion
The present data elucidate the neural correlates of different stages of the transition from incidental learning to explicit memory formation. In accord with our predictions, the neural activity within the right VLPFC and the VS increased in the time period directly preceding the emergence of explicit memory. This demonstrates that key areas observed during explicit and motivational learning may play an important role for the initiation of the transition from implicit to explicit memory formation. In conclusion, processes such as the evaluation of predictions during learning may form the cognitive and motivational basis for this change in the mode of learning.

The objective behavioral measure used here allows the examination of neural activity that precedes the emergence of explicit memory. The behavioral studies clearly confirmed that
the RT criterion (TRT) can be used to determine the individual time point when explicit knowledge or conscious awareness about the hidden rule begins to develop. The experimental design implemented the hidden rule as a relational link between the first and the last response. Thus, the hidden rule could not be learned as a simple motor sequence and could not be predicted from the stimulus pair. Therefore, the learned association has an abstract, rule-based character. The ability to deliberately verbalize the hidden rule occurs immediately after a participant has shown fast responses in direct succession (3 trials) and is strengthened across the next trials. The results from Experiment 3 further demonstrate that during the period before the TRT, nearly no explicit knowledge can be assessed. This statement is supported by the behavioral data from all 3 studies reported here that show that after single fast responses explicit knowledge is rare, but after 3 successive fast responses, most of the participants (ca. 85%) were able to verbalize the rule. More importantly, the RTs to the determined inputs remained faster than to the undetermined inputs for all participants after the TRT criterion. Therefore, the TRT can be characterized as indicating the starting point of a constant RT drop that signals explicit knowledge with a very high reliability. The TRT can thus be used as a reliable measure of the emergence of explicit knowledge. To ensure the absence of explicit knowledge in the period directly preceding the TRT, only trials before the first fast RT of the TRT, thus in a time period where no differences between the RTs could be observed, were used in the fMRI and EEG analyses.

The results from the fMRI study show a clear change in the neural activity in the trials directly preceding the individual TRT and before the consecutive insight into the hidden rule occurs. This increase in neural activity is in close temporal relation to the emerging awareness and indicates a qualitative change in information processing as the basis for the development of explicit memory. Thus, our results suggest that the insight is not simply due to a strengthening of memory representation alone as proposed in a theoretical model (Cleeremans and Jimenez 2002) but rather requires additional processes, such as restructuring (or integrating) task representations on the basis of the ongoing evaluation of predictions. Theoretical models suggest such an additional process that initiates or triggers insight and should precede the development of awareness (Runger and Frensch 2008). A frequently discussed mechanism here is the generation and evaluation of event predictions. During incidental learning, predictions about upcoming events such as a required response, response speed, or feedback are generated. Unexpected deviations of those predictions are detected, and attention is allocated to the unexpected events that resulted in the generation of awareness for the event (Haider and Frensch 2009). Previous studies established a role of a cortico-subcortical network including the striatum and the frontal cortex for implicit learning (Grafton et al. 1995; Berns et al. 1997), for the generation of predictions during learning (Schultz and Dickinson 2000; Waeli et al. 2001), for feedback-based learning (Shohamy et al. 2004; Bischoff-Grethe et al. 2009) and for the coding of abstract rules (Wallis et al. 2001). Within this network, midbrain dopaminergic projections to the VS are assumed to be involved in the representation of predictions (Schultz 2000; Elliott et al. 2003; O’Doherty et al. 2004; Yacubian et al. 2006) and enhance synaptic plasticity within several regions (Seger 2008). Previous results demonstrated that the striatum is involved in the evaluation of feedback valence independent of reward (Bischoff-Grethe et al. 2009). Recently, it has been demonstrated that the striatum generates sustained activity that can be used to stabilize the learned information (Histed et al. 2009). Thus, there is strong evidence for an involvement of the VS in motivational learning, and our results may indicate that an evaluative process of predictions about the determined response within the frontostriatal network can be regarded as the process that triggers the consecutive insight in the hidden rule. It could be speculated that the effort or speed of the processing of the determined input differed due to implicit learning, and this resulted in a discrepancy between the predicted and actual costs for the production of the response. The prediction and the unexpected violation of the prediction may be coded within the frontostriatal network that generates a signal that enhances learning. This signal presumably supports learning and resulted in the acquisition of additional explicit representations of the task structure. This assumption is in accord with a previous study that reported an involvement of the right VLPC in the implicit detection of a deviation from learned associations (Rose et al. 2005). The acquisition of additional representations could also explain why no decreases in the BOLD signal were observed. Due to the development of explicit knowledge about the determined last response, there was no need for the participant to further apply the given processing rules, which could have resulted in a decrease of activity in regions related to this mode of processing. However, the signal from the frontostriatal network enhances learning in many regions related to learning that may overlap with regions activated by applying the explicitly given rules. Alternatively, the participants may not fully rely on the hidden rule and may have crosschecked both results.

In previous experiments, it has been suggested that the basal ganglia and the MTL interact during learning, depending on the implicit or explicit mode of processing (Poldrack et al. 2001). In accord with this observation, we found an increase of MTL activity around the TRT with the generation of explicit memory. However, no decrease within the basal ganglia was observed. Furthermore, the MTL effect was located close to the area that showed an effect of implicit learning of a comparable hidden rule in a previous study (Rose et al. 2002). Taken together, the results indicate no change in the area involved for the representation of the information but that the signal generated by the frontostriatal network may enhance learning and consequently strengthens the representation of the task structure within the MTL.

In addition, the fact that our data show an increase in coupling between cortical areas in agreement with a top-down binding and amplification signal, which is presumably modulated by the increasing strength of incidentally acquired predictions. Such a top-down binding process is postulated by many frameworks of consciousness in relation to the emergence of large-scale coupling between the neural populations that represents different aspects of the information (Engel et al. 2001; Crick and Koch 2003; Dehaene et al. 2003). Furthermore, some authors assume that the widespread connectivity between those neural populations can make represented knowledge available to a variety of processes, which is subjectively experienced as a conscious state. This global availability of information is also required for the explicit character of incidentally learned knowledge that consists of memory traces that are processed in different neural circuits. However, one implication of these models is that the coherent neural network has to be established before the information can become consciously available. Therefore, the observed increase in gamma-band coherence that directly preceded the starting point of the following decrease in RT may reflect the formation of such a global network and therefore represents the driving neural process for the emergence of conscious awareness.
The temporal and spatial localization of the fMRI and EEG effects and the accordance with the results from experiments on perceptual awareness strongly suggest an interpretation of the results in terms of a reinforcement process and a top-down binding signal, which is important for the development of explicit memory.

Overall, our approach provides direct experimental evidence for a qualitative change in information processing as a prerequisite for the emergence of awareness during incidental learning in accord with several theoretical models (Dehaene et al. 2003; Baars 2005; Seth et al. 2006). An important cognitive process for this transition process is the generation of awareness, but possible other processes may also play a role. The characterization of the neuronal events that initiate the emergence of explicit memory is the basis for further studies to evaluate the contribution of distinct processes such as working memory or attention. Taken together, the fMRI and the EEG results strongly support the role of a subcortico-cortical network including the lateral frontal cortex and the VS in establishing explicit memory during incidental learning.

The capacity to generate explicit representations of complex rules is also the basis for higher cognitive functions such as reasoning and language. It has therefore been speculated that an evolutionary advantage in the architecture of the human frontal cortex may explain the constraints of nonhuman primates for such processes (Fitch and Hauser 2004; Friederici et al. 2006).

Funding
Deutsche Forschungsgemeinschaft and Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie.

Notes
Conflict of Interest: None declared.

References
Haider H, Frensch PA. 2009. Conflicts between expected and actually performed behavior lead to verbal report of incidentally acquired sequential knowledge. Psychol Res. 73:817–834.
activity during Gestalt perception in schizophrenia. J Neurosci. 26:8168-8175.