Main Manuscript for

Decoding task representations that support generalization in hierarchical task.

Woo-Tek Lee[1,2,3], Eliot Hazeltine[1,2,4], and Jiefeng Jiang[1,2,4]

[1]Department of Psychological and Brain Sciences, University of Iowa, Iowa City, IA 52242

[2]Cognitive Control Collaborative, University of Iowa, Iowa City, IA 52242

[3]Behavioral-biomedical Interface Training Program, University of Iowa, Iowa City, IA 52242

[4]Iowa Neuroscience Institute, University of Iowa, Iowa City, IA 52242

*Corresponding author: Jiefeng Jiang

Email: jiefeng-jiang@uiowa.edu

Author Contributions: **Woo-Tek Lee**: Conceptualization, Methodology, Investigation, Visualization, Writing—original draft, Writing—review & editing. **Eliot Hazeltine**: Conceptualization, Supervision, Writing—review & editing. **Jiefeng Jiang**: Conceptualization, Methodology, Investigation, Supervision, Writing—review & editing, Funding acquisition.

Competing Interest Statement: The authors declare there is no known conflict of interest to disclose.

Classification: Biological Sciences, Psychological and Cognitive Sciences.

Keywords: Hierarchical representation, task learning, generalization, integrative encoding, decoding

This PDF file includes:

Main Text
Figures 1 to 4
Tables 1

Abstract

Task knowledge is encoded hierarchically such that complex tasks are composed of simpler tasks. This compositional organization also supports generalization to facilitate learning of related

but novel complex tasks. To study how the brain implements composition and generalization in hierarchical task learning, we trained human participants on two complex tasks that shared a simple task and tested them on a novel complex task whose composition could be inferred via the shared simple task. Behaviorally, we observed faster learning of the novel complex task than control tasks (i.e., behavioral generalization effect). Using electroencephalogram (EEG) data, we could decode the constituent simple tasks when a complex task was performed (i.e., EEG composition effect). Crucially, the shared simple task, although not part of the novel complex task, could be reliably decoded from the novel complex task. The decoding strength was also correlated with EEG composition effect and behavioral generalization effect. The findings demonstrate how generalization in task learning is implemented via task reinstatement.

Significance Statement

Humans can generalize knowledge of existing tasks to accelerate the learning of new tasks. We hypothesize that generalization is achieved by decomposing a complex task into simple (sub)tasks and reusing the simple tasks to infer the structure of a new complex task and build it. Using electroencephalogram data, we showed that constituent simple tasks can be decoded from of humans learning new complex tasks. Crucially, when the structure of a new complex task can be inferred from simple tasks, the simple tasks can be decoded from the new complex task, even when they are not part of the new complex task. These findings demonstrate the importance of the reinstatement of simple tasks in task learning through generalization.

Main Text

Introduction

Humans have the remarkable ability to efficiently learn new complex and cognitively demanding tasks. To do this, we organize task representations hierarchically (1-4), such that more complex tasks are represented as compositions of simpler ones. That is, a complex task can be represented as a composition of two or more simple task representations, and new complex tasks can be learned efficiently by associating existing simple task representations (5, 6). For example, a barista can learn how to make ‘café latte’ by associating existing skills of making espresso and steaming milk. In previous studies, researchers found that combining multiple memory items enables us to adjust our behavior adaptively by predicting future events (7), or inferring novel information (8). This research has been extended to compositional representations including cognitive control parameters and task demands (5, 6, 9-13) that support adaptive, goal-directed behavior. Although the lateral prefrontal cortex (lPFC) has been shown to be involved in representing compositional task rules (14), neural evidence of how simple tasks are combined to form complex task representations is scarce.

This componential organization of task representations allows complex tasks to share components, which enables humans to generalize from previous experience and quickly adapt to new, related tasks (15). For example, humans can extract structure from experiences and infer novel knowledge from the structure, a phenomenon sometimes also termed zero-shot learning (16-18). Generalization is based on composition, but it requires additional inferential processes as it allows for learning without direct experience. This generalization in task learning may be attributed to associative inference, which refers to the ability to infer novel, indirect associations from partially overlapping associations (9, 11, 19-26). For example, the barista who is skilled at the ‘café latte’ task (espresso + milk) and the ‘café mocha’ task (espresso + chocolate) can infer how to make a hot chocolate via the ‘making espresso’ task shared by the ‘café latte’ and ‘café mocha’ tasks. In the laboratory, we also report that generalization via associative inference accelerates task learning

and improves task performance (6). Although generalization by associative inference plays an important role in task learning (3, 6, 9), how it is implemented in the brain is understudied.

Two prominent theories of association and generalization in episodic memory offer insights into the potential mechanisms of composition and generalization in task learning. The integrative encoding theory (21) posits that repeated associations between items (e.g., A and C) will create an associated memory representation (e.g., AC). Integration can occur between associated item pairs that share an item (e.g., AB and BC), which leads to an integrated memory representation (e.g., ABC) that encompasses all three memory items (19, 21, 24). The integrated memory representation will be retrieved if a subset of the memory items is present. For example, presentation of AC is expected to reinstate item B as well as items A and C. On the other hand, Kumaran and McClelland (20) suggest that an association between two memory items forms a conjunctive representation, and these conjunctive representations are connected to each of the constituent memory items. If a representation is activated (e.g., by sensory input), its connected representations will also be activated (though to a lesser extent). For example, if two conjunctions AB and BC share a memory item B, activating the conjunction AB will also activate memory items A and B. B will further activate conjunction BC, which in turn activates C. In the end, the coactivation of A and C will facilitate the formation of an indirect association of AC.

Thus, while these two major theories differ in terms of their proposed mechanisms of composition and generalization, they share predictions regarding the reinstatement of memory items. First, both theories hold that when an association is activated, its constituent memory items will also be reinstated (e.g., A and C in the AC pair). Applying this idea to the context of task learning, we predict that executing a complex task is expected to reactive the representations of the constituent simple tasks. Additionally, as discussed above, both theories predict that in associative inference when an association is presented (e.g., AC), the latent (i.e., not presented) item (e.g., B) will also be reinstated. Similarly, applied to task learning, the reinstatement of the latent task is expected to be key for generalization.

To test these predictions, we leveraged a behavioral paradigm of generalization in task learning (6). Subjects performed this paradigm while we obtained high temporal resolution and high-density electroencephalogram (EEG) data, to which we applied decoding and representational similarity analysis (RSA, 10, 27). To preview the results, we first replicated the behavioral generalization effect reported in (6). We then observed robust decoding of constituent simple task representations from the complex task EEG data, supporting the composition hypothesis. Critically, we observed in the complex task EEG data above-chance decoding of the latent task (e.g., B) that is reinstated by the indirect association (e.g., AC). We further found that the strength of the latent task representation in complex task EEG data is linked to both the strength of the EEG composition effect and generalization performance. Overall, the results suggest that complex task learning results in a network representing compositional information that enables zero-shot task learning (e.g., generalization) via the reinstatement of the latent task. The strength of generalization is linked to the strength of the composition.

Results

Experimental design.

Participants (n = 40) first learned six simple tasks (coded as A-F), each of which was a perceptual decision-making task defined on one out of the six stimulus features (Fig. 1A). On each trial, participants judged whether a cued feature appeared in a forthcoming target stimulus (Fig. 1B). Next, during the training stage, the participants learned four complex tasks, each consisting of two simple tasks. The complex tasks, denoted as the pair of constituent simple tasks (e.g., AB means a complex task consisting of simple tasks A and B), were presented on alternate blocks (odd blocks: AB, DE, even blocks: BC, EF; Fig. 1C). The complex tasks were divided into two sets, such that

each set contained two complex tasks sharing one simple task (e.g., AB and BC). After the training stage, participants performed four new complex tasks (AC, DF, AD and CF) in the test stage (Fig. 1C). Two of the new complex tasks could be generalized via associative inference because both simple tasks were associated with the same simple task during training (e.g., AC from AB and BC; Fig, 1D) and are hence termed generalizable tasks. In contrast, the other two complex tasks (e.g., AD and CF) could not be generalized using associative inference, as the constituent simple tasks did not overlap during the training stage. These were termed control tasks. This design allowed us to investigate the behavioral generalization effect using the difference in behavioral performance between generalizable and control tasks, and temporal dynamics of composition and generalization effects using the EEG data.

Behavioral data.

Group-average accuracy for both simple and complex tasks remained above 80% throughout the experiment (Fig. S1, S2, Note S1). To assess the generalization effect, we focused on the complex tasks in the test stage and compared the performance of the generalizable and control conditions (Table 1). Accuracy for the generalization and the control conditions did not differ significantly ($t(39) = 1.564$, $p = .126$, Cohen's $d = .25$). For RT analysis, we followed our previous procedure (6) and regressed out confounding factors such as performance difference among simple tasks, response switch cost and so on (see Behavioral Analysis in Methods for details). The resulting residual RTs were used to test the hypothesis. Replicating our previous finding (6), we found a significantly faster decrease in RT over time ($t(39) = -2.3446$, $p = .024$, Cohen's $d = .37$) in the generalizable complex tasks than in the control complex tasks (Fig. 1E), indicating faster learning for generalizable than control complex tasks.

Above-chance decoding in simple task EEG data.

For each participant and each time point (i.e., a 50 ms bin) of the simple task phase (Fig. 1C), we trained a multiclass decoder on the EEG data to classify the six simple tasks. The decoding analysis was performed after the onset of the target stimulus (Fig. 1B), so the decoder performance was not affected by perceptual processes related to the cue. Cross-validation decoding accuracy of simple tasks increased shortly after the target onset and was significantly above chance after 50 ms, peaking at around 300 ms (Fig. 2A, left panel). The longest epoch that showed significant decoding performance (indicated in red line in left panel of Fig. 2A) survived non-parametric permutation test for multiple-comparison correction ($p_{corrected} < .001$). The map of averaged absolute weights of the electrodes from the decoder revealed involvement of occipital and posterior electrodes around the peak decoding accuracy time point (Fig. 2A, right panel). This result suggests that the simple task representations can be robustly decoded in EEG data after target stimulus onset during their performance.

Overview of cross-phase EEG decoding analysis.

To assess the reinstatement of simple tasks during the performance of complex tasks, we trained decoders on simple task phase EEG data, as described above, and tested them on complex task phase data (see Method for details). To test the composition effect (e.g., decoding simple tasks A and C from complex task AC), the decoders were tested on complex-task EEG data from the training stage (blocks 1-8). To test the generalization effect (e.g., decoding simple task B from complex task AC), the decoders were tested on the test stage of the complex task phase (blocks 9-11, Fig. 1C). For each test time point, the decoder produced a probabilistic prediction of each simple task being represented. The logit-transformed decoding probability was regressed against a linear model with two predictors representing composition and generalization effects separately. The trial-wise regression coefficient of each predictor was averaged for each participant and tested against 0 using a one-sample t-test. The presence of a composition/generalization effect predicts that beta values will be greater than 0. Thus, we used the t-values to quantify the strength of the

composition and generalization effects. For the multiple comparison corrections, we applied a non-parametric permutation test (28).

EEG composition effect.

We predicted that the constituent simple task representations should be decodable during the complex task trials. To test this prediction, we first trained the decoder with simple tasks EEG data and tested it on the complex task EEG data. Figure 2B shows the time points that the constituent simple task representations can be decoded from the complex task EEG data (e.g., A & B from AB) in the training and test stages respectively. In the training stage, constituent simple task representations can be decoded above chance level from complex task EEG data between 150 ms and around 1,200 ms post target onset when the decoders were trained between around 150 ms and 1,000 ms post target onset (pcorrected < .001; Fig. 2B, left panel). In the test stage, this EEG composition effect remained significant (pcorrected = .041; Fig. 2B, right panel) yet was reduced and significant between around 150 ms and 700 ms post target onset in training data and between 150 and 550 ms post target onset in test data. The results indicate that the execution of complex tasks involves activation of constituent simple task representations, and this activation is more extensive during the training stage than the test phase.

EEG generalization effect.

While representing complex tasks as compositions of simple task representations can be beneficial for their initial learning, generalization can support the learning of a novel complex task if its composition can be inferred from other learned complex tasks. Both the integrative encoding theory (21) and the recurrent interaction theory (20) predict the reinstatement of the latent task representation during the test stage complex task trials (e.g., B from AC), despite the fact that the latent task is not present on the trial. To test this prediction, we applied the decoding analysis and RSA to assess the reinstatement of the latent simple task representations in the complex task EEG data (see Representational similarity analysis in Methods for more details). Consistent with the prediction, latent task representations were decoded above chance level between 850ms to 2,000 ms following target onset in the test stage using decoders trained between 200 ms to 400 ms following target onset (Fig. 2C, right panel, pcorrected = .047). As the integrative encoding theory also predicts that the integrative memory can be formed after both partially overlapping associations are experienced (11, 21), generalization may occur in the training stage after both complex tasks from the same set were experienced (e.g., AB and AC). We also tested the decoding of the latent task in the training stage (e.g., C from AB). Note that the latent tasks in the training phase did not contribute to generalization via associative inference. In contrast to the test stage data, no time points reached statistical significance after the cluster-based non-parametric permutation test. This result suggests that the generalization between neural representations of complex tasks emerged mainly during the test stage, when complex task execution may benefit from generalization.

Correlation between EEG composition and generalization effects.

The two major accounts of generalization in episodic memory (20, 21) hold that that the composition effect (i.e., that constituent tasks can be decoded during the performance of complex tasks) lays the foundation of generalization effect (i.e. that latent task can be decoded from complex tasks in the test stage). Therefore, we predicted that these two effects would be correlated. To test this correlation, we first calculated averaged EEG generalization effect for each participant within a time window defined in the following manner (Fig. 3A): Based on the EEG generalization effect observed in Fig. 2C, we trained decoders on simple task data between 200 ms and 300 ms post target onset. As the EEG generalization effect spans before and after responses, we divided the duration of the complex task showing the generalization effect (i.e., 850 ms to 2,000 ms post target onset, see Fig. 2C) into early and late parts using a threshold of the 75 percentiles of all complex task RTs across all participants (i.e., 1,100 ms). For consistency, the EEG composition effect was trained on simple task EEG data from the same window (i.e., 200-300 ms post target onset) and tested on time points showing statistically significant composition effect, which starts at 100 ms (Fig. 2B, left). The test

window was truncated at 850 ms to avoid temporal overlap between the periods used to measure the composition and generalization effects.

Consistent with the prediction, EEG composition effect in the training stage showed a significant positive correlation with the EEG generalization effect before the 75 percentile RT ($r(38) = 0.34$, $p = .032$) but not after ($r(38) = 0.06$, $p = .691$; fig 3B). To rule out the possibility that the finding depended on the truncation of the time window, we repeated the analysis using the full time window (from 100 ms to 1,350ms). The untruncated EEG composition effect exhibited a consistent pattern (before: $r(38) = 0.39$, $p = .013$; after: $r(38) = 0.09$, $p = .589$). To investigate the temporal dynamics of the correlation between composition and generalization effects in the EEG data, we computed the average EEG generalization effect in the cluster (Fig. 3A, center panel) for each time point of the complex task and correlated it with the average EEG composition effect. Note that this analysis did not rely on a threshold that segments trials into early vs. late conditions. We found a positive correlation between the EEG composition effect and EEG generalization effect between 850 and 1,100 ms post stimulus onset. The correlation gradually attenuated and reached chance level at 1,500 ms. The correlation coefficient showed a significant decrease over time (Fig. 3C, slope: $t(20) = -8.62$, $p < .001$). Overall, these results indicate that the composition of simple task representations during the training stage of the complex task phase was linked to the EEG generalization effect before and near the response in the test stage.

Correlation between behavioral and neural effects of generalization.
Lastly, we expected that the EEG generalization effect would benefit performance of generalizable complex tasks. Therefore, we conducted exploratory analysis to investigate the behavioral relevance of the EEG generalization effect. We first assessed individual behavioral generalization effects using the RT slope difference between generalizable and control conditions (generalizable – control) and correlated this behavioral generalization effect with the EEG generalization effect from the previous analysis (Fig. 1E). The slope differences showed a significant negative correlation with the early part of the EEG generalization effect ($r(38) = -0.38$, $p = .017$), indicating that a stronger EEG generalization effect is linked to larger difference in RT decrease in generalizable tasks compared to control tasks. The correlation with the later part of the EEG generalization effect was not significant ($r(38) = -0.23$, $p = .160$).

To examine the temporal dynamics of the brain-behavioral correlation over time, we conducted the time-resolved correlation analysis above using EEG and behavioral generalization effect. This analysis is similar to the time-resolved correlation analysis between composition and generalization effects in the previous section. The change of correlation over time showed a significant positive linear trend ($t(20) = 2.64$, $p = .016$), indicating that in the test stage, participants who exhibited stronger EEG generalization effect before and near response tended to show a stronger behavioral generalization effect.

Discussion

In the study, we investigated the role of task representation reinstatement in supporting composition and generalization in hierarchical task learning. To this end, we employed an established experimental paradigm that involves six simple tasks that can be combined to form complex tasks. Participants first trained with simple tasks, then learned complex tasks that consisted of two simple tasks. Specifically, participants learned complex tasks that shared a simple task (e.g., AB and BC) in the training stage. In the following test stage, participants performed generalizable complex tasks that consisted of indirectly associated simple tasks (e.g., AC) and control complex tasks that consisted of non-associated simple tasks (e.g., AD). Note that the generalizable complex tasks and control complex tasks are composed of constituent simple tasks that have received equivalent amounts of practice. Nonetheless, we hypothesized that participants would learn novel complex tasks efficiently by (1) compositing simple task representations and (2) generalizing complex tasks via associative inference. Behaviorally, we successfully replicated the behavioral finding of faster

decrease in RT for generalizable complex tasks than control complex tasks (6), thus supporting generalization and composition (as composition is the foundation of generalization).

From the EEG data, simple task representations were reliably decoded from the complex task EEG data, supporting the composition of simple tasks. More important, we also found a significant latent task representations from the test stage complex task EEG data despite the fact that the latent task was not part of the complex task, supporting the generalization of complex task representations. We also observed that a stronger EEG composition effect during the training stage was associated with a stronger EEG generalization effect (measured prior to response) during the test stage. We further found that a stronger EEG generalization effect during the test stage at early time points of the trial was associated with a larger behavioral generalization effect.

Previous research has shown that lower-level perceptual information can be decoded from the EEG data, including the color, orientation, spatial location, spatial frequency, and motion direction (29-34). Recent advances in EEG decoding further achieve decoding of different rules within a task (10, 27). In this study, we extended these findings by decoding different tasks. To tease apart task representations from stimulus representations, we used the same stimuli for all tasks. To further de-confound the perceptual information of task cues from the task itself, decoding analysis was conducted on time points in which task cues were absent. The successful decoding in the simple task phase (Fig. 2A) validates our decoding procedure to test the hypotheses.

Intriguingly, within complex task trials, the EEG composition and generalization effects showed contrasting temporal dynamics. While the constituent simple task representation was mainly observed shortly after the target onset in the complex task EEG data (i.e., the composition effect, Fig. 2B), the latent task representation mainly emerged 900 ms after the target onset (i.e., the generalization effect, Fig. 2C). Both associative episodic theories hypothesize that generalization relies on the reinstatement of the latent task, so the decoding of the latent task provides strong support for this hypothesis. Integrative encoding theory (19, 21, 24) suggests that the constituent simple tasks trigger pattern completion of the conjunction of all three tasks (e.g., ABC and DEF), leading to the delayed activation of latent task (35). On the other hand, Kumaran and McClelland (20) posit a recurrent network between individual item representations and conjunctive representations consisting of two associated item representations. Based on this account, activation of a conjunction will first activate the constituent simple tasks, which, after some delay, activate the latent task representation via recurrent connections. Thus, the observed temporal dynamics that the EEG composition effect precedes the EEG generalization effect is consistent with the latter account. Nevertheless, future research is needed to adjudicate between the two theories in hierarchical task learning.

Also consistent with the previous accounts of generalization, we observed a significant correlation between the EEG composition effect and the EEG generalization effect, suggesting that compositionality may support generalization. Specifically, early EEG generalization effect in the test stage significantly correlated with the EEG composition effect in the training stage. This pattern is predicted by the temporal order of the training and test stage, such that a stronger composition of simple tasks during the training stage would lead to stronger activation of the latent task representation during the test stage. Additionally, the relation between EEG composition and generalization effects decreased over the time-course of the trial (Fig. 3B, C). We speculate that this pattern may be attributable to the different durations of latent task reinstatement across subjects. In other words, some participants may attenuate the reinstatement of the latent task sooner than others, which would compromise the correlation at later time points. Underlying the connection between the composition and generalization effects are the associations linking the constituent simple tasks. In particular, when a simple task is shared by multiple complex tasks, it bridges multiple associations, which then form a network (20). A remaining question is how to efficiently organize this network. One possible approach is to embed the network into a cognitive map, implying that spatial and non-spatial information can be organized into a map like space for

goal directed behavior (36-40). This would allow for efficient inference (37, 41). Another non-exclusive approach is representation compression (42) that strengthens task associations to facilitate composition and generalization.

Our exploratory analysis further revealed that early EEG generalization effect in the test stage showed a significant negative correlation with the RT slope, suggesting that participants with stronger latent task reinstatement also exhibited better generalization in performance. This finding supports the behavioral relevance of the reactivation of the latent task representation. In other words, it is not an unwanted byproduct, but a mechanism that supports generalization via associative inference. The behavioral relevance is also consistent with the finding that the correlation between the EEG generalization effect and the behavioral generalization effect was more pronounced before the response was made (Fig. 3D, E).

In conclusion, we provide electrophysiological evidence that the reinstatement of simple tasks supports complex task learning in hierarchical task organization. Specifically, we report the decoding of simple task representations from the complex task EEG data, which suggests that complex task representations include constituent simple task representations. Furthermore, we found significantly above-chance decoding of latent task representation from the EEG data of complex tasks that consisted of indirectly associated simple tasks, which suggests that participants generalized complex task compositions via associative inference and the reinstatement of the latent simple task. Further supporting this conclusion, this latent task representation in the complex task EEG data was significantly related to the EEG measure of composition effect and behavioral generalization effect. Together, these findings extend our understanding of hierarchical task organization that supports efficient task learning.

Materials and Methods

Participants.

Forty-five healthy young adults participated in the experiment and received compensation at a rate of $25 per hour. Four participants were excluded from further analysis due to excessive non-stereotypical artifacts in the EEG data. One participant was excluded from the further analysis due to lower than 70% accuracy in a specific feature association during every block in training stages. The final sample consisted of forty participants (23 females and 17 males; age: M = 22.69, SD = 4.35). The experiment was approved by the University of Iowa Institutional Review Board.

Stimuli.

We used compound stimuli composed of six features. Each feature had two possible values (Fig. 1A). The features and their values are as follows: shape (oval or rectangle), color (green or red), outline (thick or thin), shadow (cast upward or downward), tilt (clockwise or counterclockwise), and pattern (parallel or diagonal). This design leads to 64 unique stimuli. Other stimuli included word and image cues indicating each of the feature values (Fig. 1A).

Procedure.

The experimental procedure was similar to Lee, Hazeltine and Jiang (6). The experiment consisted of two phases: a simple task phase (8 blocks of 48 trials each) followed by a complex task phase (11 blocks of 48 trials each). Participants had self-paced resting periods between blocks and had at least 5 minutes of rest between phases.

Each simple task trial started with the presentation of a cue of a stimulus feature value (e.g., filling pattern is diagonal) for 1,500 ms (Fig. 1B). The cue alternated between image and word format (Fig. 1A) after each trial so there were no cue repetitions. The cue was followed by the presentation of a fixation cross for 500 ms, after which the target stimulus was presented until

response or a response deadline. The initial response deadline was set to 1,500 ms and was adjusted based on a staircase procedure. Specifically, the response deadline decreased by 50 ms when the participants made one or no error and increased by 50 ms when participants made three or more errors in the last five trials. Otherwise, the response deadline remained unchanged. The participants were instructed to judge whether the cue matched the feature of the target stimulus by pressing the 'D' or 'K' button for yes or no, respectively. After each response, participants received feedback (correct or incorrect, presented in the center of the screen) for 500 ms. From the second block on, the feedback also included overall task accuracy. As a cue indicates which feature is task-relevant, the cues defined six different simple tasks (one for each cue). The simple tasks were coded using letters A-F (Fig. 1C), with the simple task-letter mapping randomized for each participant.

The complex task was similar to the simple task except for three modifications (Fig. 1B). First, the first screen showed two cues for 2,500 ms. Second, the rule of the complex task followed the 'exclusive or' rule. Specifically, participants were required to press the 'D' button if only one cue matched the target stimulus and the 'K' button otherwise. This rule was employed to ensure that the correct response can only be made after evaluating both cues. Third, the initial response time deadline was set to 2,500 ms for the staircase procedure. A complex task can be viewed as a combination of two simple tasks and was used to investigate composition and generalization in hierarchical task representation. Thus, we coded a complex task using constituent simple tasks (e.g., complex task AB is composed of simple tasks A and B).

The complex task phase consisted of two different stages (Fig. 1C): a training stage (block 1–8) and a test stage (block 9-11). To study generalization at the complex task level, the simple tasks were divided into two groups (A-C and D-F). In the training stage, participants performed complex tasks AB and DE in odd runs and complex tasks BC and EF in even runs. In the test stage, participants performed generalizable complex tasks AC and DF, as well as non-generalizable control complex tasks AD and CF. Complex tasks AC and DF are generalizable because the constituent simple tasks were practiced with a common simple task (e.g., B and E) during the training stage. In contrast, control complex tasks AD and CF cannot be inferred by complex tasks learned in the training stage based on integrative encoding theory (19, 21, 24) or the recurrent interaction of association theory (20).

Behavioral analysis.

The behavioral analysis focused on replicating the generalization effect in Lee, Hazeltine and Jiang (6). Specifically, we tested the generalization effect of faster learning of the generalizable than control complex tasks (Fig. 1D) using the analysis from Lee, Hazeltine and Jiang (6). We first removed error trials and trials with RTs that were above 5 standard deviations from the subject mean. Next, we built a nuisance effect design matrix consisting of trial-wise status of response, response repetition, post-error, cue modality (text or image), task repetition, and cue modality repetition and regressed out the effects of these confounding factors from RTs in the test stage. Next, we fit a design matrix, which consisted of binary regressors marking generalizable and control complex tasks (to estimate RTs at the beginning of test stage) and their temporal order (to estimate RTs change over trials) against the residual RTs. These steps were performed separately for each participant. We compared the individual regression coefficients of estimated RT improvement over trials (slope) between regressors representing generalizable and control conditions of feature associations at the group level using paired t-tests.

EEG recording & preprocessing.

EEG data were recorded using a 64-channel active Ag/AgCl electrodes connected to an actiCHamp plus amplifier (BrainVision) at a rate of 500 Hz (10 s time-constant high-pass and 1,000 Hz low-pass hardware filters). The reference and ground were set to Pz and Fpz electrodes respectively. Raw EEG data was preprocessed using customized MATLAB scripts and the

EEGLAB toolbox (43). The EEG data were first band-pass filtered from 0.3 Hz to 50 Hz using Hamming windowed sinc FIR filter (pop_eegfiltnew.m). From the filtered continuous EEG data, we removed epochs with non-stereotypical artifacts for further analysis using outlier statistics and visual inspection. The remaining EEG data were re-referenced to common average and decomposed using the independent component analysis (ICA) to reject components reflecting eye movements and electrode specific artifacts. After excluding the artifact components, the EEG data was reassembled and epoched relative to the target stimulus onset (simple task: -1,000 to 1,250 ms, complex task: -1,000 to 2,000 ms) for further analysis (Fig. 4A). The choice of using target stimulus locked data was to control for the confounds of perceptual information when training and testing the decoders (see below). The preprocessed EEG data was smoothed and down-sampled to 20 Hz to reduce computation time for the decoding analysis and noise in the data.

EEG Decoding analysis.
We trained decoders to discriminate among the six simple tasks at each time point using trial-level EEG data from all 64 electrodes (Fig. 4B). The decoders were L2 regularized multinomial linear regression with a tolerance (tol) of 1 × 10-4 and inverse of the regularization strength of 1 (44). Decoder training and test were implemented using the scikit-learn package (45) and customized Python scripts. This approach was applied to 5-fold cross-validation using simple task data for each participant. The performance of the decoders was assessed by comparing the logit transformed probabilities to logit transformed baseline decoding performance (i.e., 1/6) using a one-sample t-test.

Following cross-validation, a decoder was trained on each time point using all simple task data (Fig. 4C) and tested on each time point and each trial of the complex task data (30, 34, 46, 47) to assess the representation strength of simple tasks during the execution of complex tasks (Fig. 4D). As a result, for each test, the decoder generated a vector of six classification probabilities, each representing the decoder's belief of a simple task being represented in the complex task at the time point of test. These probabilities were transformed with multinomial logit transformation for further analysis. The decoding results were stored in a four-dimensional tensor for each participant in the format of [trials in the complex task] × [time points of simple task] × [time points of complex task] × [6 simple tasks] for each trial of the complex task phase.

Representational similarity analysis.
For the RSA, we first tested composition and generalization effects separately using the EEG data (Fig. 4E). The composition effect is based on the hypothesis that neural representations of the complex task should include the encompassing simple tasks (e.g. AC = A + C). The generalization effect is built under the prediction that neural representations of the generalizable complex task should reinstate the latent task (e.g. simple task B is expected to be activated during complex task AC, if AB and BC is generalized). Each effect is represented as a predictor (i.e., a column in Fig. 4E). Next, the two predictors were included in a linear model, which was regressed against the logit-transformed classification probabilities from each trial of the complex task EEG data. From the output of linear regression, we acquired trial-wise t-value for each predictor, which represents the composition/generalization effect on the current trial. For each regressor, the t-values were grouped as a vector with the length of trials in the complex task. In this vector, trials that showed extreme t-values (± 5 SD) were excluded following a previous protocol (10). Each vector from composition and generalization effects was averaged to examine the overall fitness of the regressor across trials. The regression was conducted on decoding probabilities obtained when the decoder was trained on each time point of the simple task EEG data and tested on each time point of the complex task EEG data. In the end, these analyses generated a t-value matrix with the size of the number of time points of simple task × the number of time points of complex task for each participant. The statistical significance was calculated by conducting one sample t-test across participants against 0 at each of combinations of simple task and complex task time points.

Cluster-based permutation test.
To control for multiple comparisons in the above the final t-value matrix, we conducted a cluster-based permutation test with customized python script based on Maris and Oostenveld (28). Specifically, with simple task EEG data, we repeated a 5-fold cross-validation decoding analysis using shuffled labels of the simple tasks 10,000 times. We pooled the clusters with the largest sum of absolute t-values (t-sum) from each repetition to create a null distribution of t-sums, which takes into consideration of both the cluster size and statistical significance. From the actual result, the cluster with the largest t-sum is compared to the null distribution to obtain corrected p-value. For the RSA, we repeated the analysis using shuffled labels of the simple tasks before training the decoder 10,000 times, while keeping the labels of the complex tasks the same between repetitions. The remaining procedure was identical to the non-parametric permutation test for simple task decoding analysis.

Data availability.
All data and analysis scripts related to this paper are available in Open Science Framework at (https://osf.io/mzf4a/).

Acknowledgments
This project was supported by the National Institute of Mental Health (R01MH131559 to J.J.). We thank the members of the Cognitive Control Collaborative and the members of the Human Perception and Performance Group at the University of Iowa for helpful discussions.

References

1. E. H. Schumacher, E. Hazeltine, Hierarchical task representation: Task files and response selection. *Current Directions in Psychological Science* **25**, 449-454 (2016).
2. S. Musslick, J. D. Cohen, Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences* **25**, 757-775 (2021).
3. S. Musslick *et al.* (2017) Multitasking capability versus learning efficiency in neural network architectures. (Cognitive Science Society).
4. M. W. Cole, P. Laurent, A. Stocco, Rapid instructed task learning: A new window into the human brain's unique capacity for flexible cognitive control. *Cognitive, Affective, & Behavioral Neuroscience* **13**, 1-22 (2013).
5. T. Ito *et al.*, Compositional generalization through abstract representations in human and artificial neural networks. *Advances in Neural Information Processing Systems* **35**, 32225-32239 (2022).
6. W.-T. Lee, E. Hazeltine, J. Jiang, Interference and integration in hierarchical task learning. *Journal of Experimental Psychology: General* (2022).
7. H. Eichenbaum, The hippocampus and mechanisms of declarative memory. *Behavioural brain research* **103**, 123-133 (1999).
8. D. Zeithamova, A. L. Dominick, A. R. Preston, Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron* **75**, 168-179 (2012).
9. C. Bejjani, Z. Zhang, T. Egner, Control by association: Transfer of implicitly primed attentional states across linked stimuli. *Psychonomic bulletin & review* **25**, 617-626 (2018).
10. A. Kikumoto, U. Mayr, Conjunctive representations that integrate stimuli, responses, and rules are critical for action selection. *Proceedings of the National Academy of Sciences* **117**, 10603-10608 (2020).
11. J. Jiang *et al.*, Temporal dynamics of memory-guided cognitive control and generalization of control via overlapping associative memories. *Journal of Neuroscience* **40**, 2343-2356 (2020).

12. A. F. Palenciano, C. González-García, J. E. Arco, L. Pessoa, M. Ruz, Representational organization of novel task sets during proactive encoding. *Journal of Neuroscience* **39**, 8386-8397 (2019).
13. R. D. Mill, M. W. Cole, Neural representation dynamics reveal computational principles of cognitive task learning. *bioRxiv*, 2023.2006. 2027.546751 (2023).
14. C. Reverberi, K. Görgen, J.-D. Haynes, Compositionality of rule representations in human prefrontal cortex. *Cerebral cortex* **22**, 1237-1246 (2012).
15. H. F. Harlow, The formation of learning sets. *Psychological review* **56**, 51 (1949).
16. B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum, Human-level concept learning through probabilistic program induction. *Science* **350**, 1332-1338 (2015).
17. B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, Building machines that learn and think like people. *Behavioral and brain sciences* **40**, e253 (2017).
18. A. R. Vaidya, H. M. Jones, J. Castillo, D. Badre, Neural representation of abstract task structure during generalization. *ELife* **10**, e63226 (2021).
19. D. Zeithamova, A. R. Preston, Flexible memories: differential roles for medial temporal lobe and prefrontal cortex in cross-episode binding. *Journal of Neuroscience* **30**, 14676-14684 (2010).
20. D. Kumaran, J. L. McClelland, Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. *Psychological review* **119**, 573 (2012).
21. D. Shohamy, A. D. Wagner, Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events. *Neuron* **60**, 378-389 (2008).
22. A. R. Vaidya, D. Badre, Abstract task representations for inference and control. *Trends in Cognitive Sciences* (2022).
23. H. Eichenbaum, Hippocampus: cognitive processes and neural representations that underlie declarative memory. *Neuron* **44**, 109-120 (2004).
24. Z. Zhou, D. Singh, M. C. Tandoc, A. C. Schapiro, Building integrated representations through interleaved learning. *Journal of Experimental Psychology: General* (2023).
25. H. C. Barron *et al.*, Neuronal computation underlying inferential reasoning in humans and mice. *Cell* **183**, 228-243. e221 (2020).
26. G. E. Wimmer, D. Shohamy, Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science* **338**, 270-273 (2012).
27. B. O. Rangel, E. Hazeltine, J. R. Wessel, Lingering neural representations of past task features adversely affect future behavior. *Journal of Neuroscience* **43**, 282-292 (2023).
28. E. Maris, R. Oostenveld, Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods* **164**, 177-190 (2007).
29. J. E. Hajonides, A. C. Nobre, F. van Ede, M. G. Stokes, Decoding visual colour from scalp electroencephalography measurements. *Neuroimage* **237**, 118030 (2021).
30. J.-R. King, S. Dehaene, Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences* **18**, 203-210 (2014).
31. G.-Y. Bae, S. J. Luck, Decoding motion direction using the topography of sustained ERPs and alpha oscillations. *NeuroImage* **184**, 242-255 (2019).
32. J. J. Foster, E. M. Bsales, R. J. Jaffe, E. Awh, Alpha-band activity reveals spontaneous representations of spatial position in visual working memory. *Current Biology* **27**, 3216-3223. e3216 (2017).
33. N. E. Myers *et al.*, Testing sensory evidence against mnemonic templates. *elife* **4**, e09000 (2015).
34. B.-I. Oh, Y.-J. Kim, M.-S. Kang, Ensemble representations reveal distinct neural coding of visual working memory. *Nature communications* **10**, 5665 (2019).
35. M. B. Jackson, Recall of spatial patterns stored in a hippocampal slice by long-term potentiation. *Journal of neurophysiology* **110**, 2511-2519 (2013).
36. E. C. Tolman, Cognitive maps in rats and men. *Psychological review* **55**, 189 (1948).
37. S. A. Park, D. S. Miller, H. Nili, C. Ranganath, E. D. Boorman, Map making: constructing, combining, and inferring on abstract cognitive maps. *Neuron* **107**, 1226-1238. e1228 (2020).

38. H. Eichenbaum, N. J. Cohen, Can we reconcile the declarative memory and spatial navigation views on hippocampal function? *Neuron* **83**, 764-770 (2014).
39. A. O. Constantinescu, J. X. O'Reilly, T. E. Behrens, Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464-1468 (2016).
40. G. Yang *et al.*, Conflicts are parametrically encoded: initial evidence for a cognitive space view to reconcile the debate of domain-general and domain-specific cognitive control. *eLife* **12** (2024).
41. S. A. Park, D. S. Miller, E. D. Boorman, Inferences on a multidimensional social hierarchy use a grid-like code. *Nature neuroscience* **24**, 1292-1301 (2021).
42. P. S. Muhle-Karbe *et al.*, Goal-seeking compresses neural codes for space in the human hippocampus and orbitofrontal cortex. *Neuron* **111**, 3885-3899. e3886 (2023).
43. A. Delorme, S. Makeig, EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods* **134**, 9-21 (2004).
44. J. Hubbard, A. Kikumoto, U. Mayr, EEG decoding reveals the strength and temporal dynamics of goal-relevant representations. *Scientific reports* **9**, 9051 (2019).
45. F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825-2830 (2011).
46. E. M. Meyers, D. J. Freedman, G. Kreiman, E. K. Miller, T. Poggio, Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of neurophysiology* **100**, 1407-1419 (2008).
47. M. G. Stokes *et al.*, Dynamic coding for cognitive control in prefrontal cortex. *Neuron* **78**, 364-375 (2013).

Figures and Tables

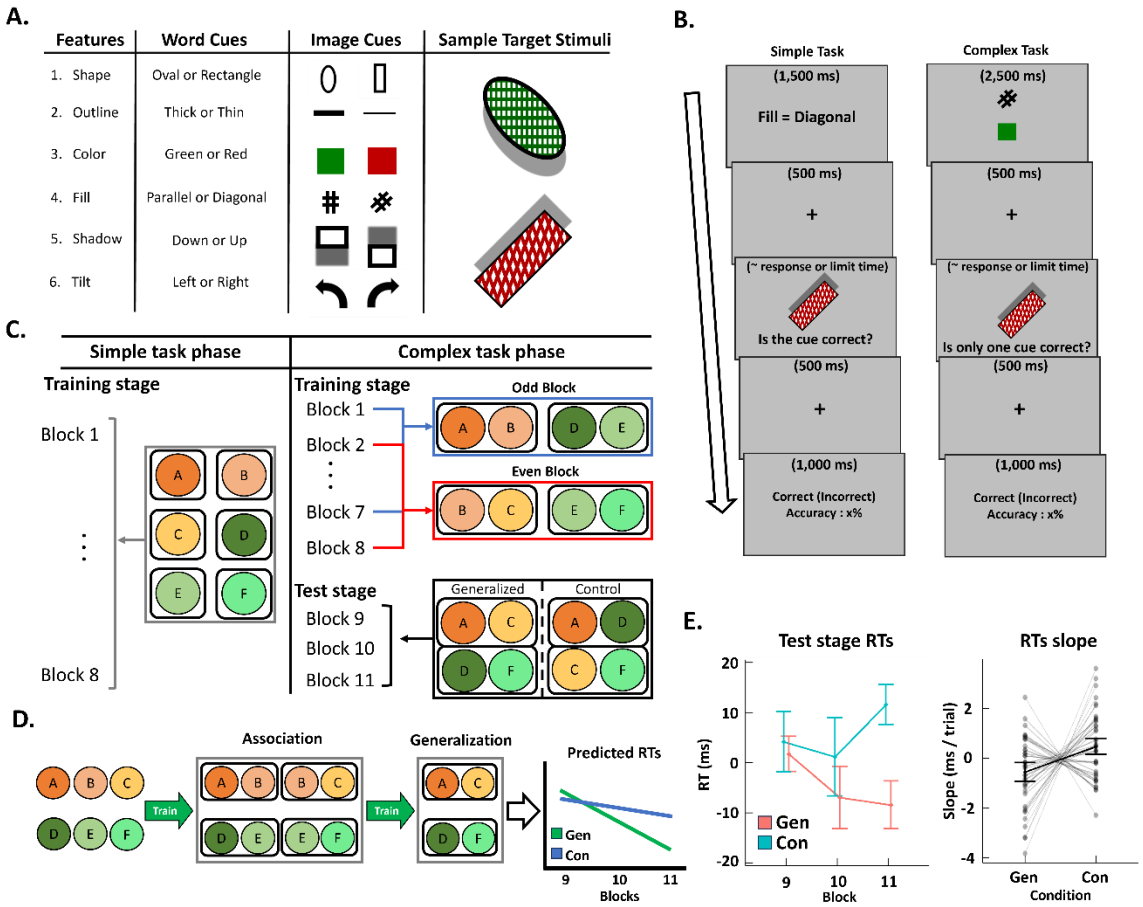

Figure 1. Experimental design and behavioral results. *(A)* Cue and target stimuli. Each cue indicated one stimulus feature and was presented as either text or image. Two example stimuli are illustrated in the rightmost column. *(B)* Trial events of the simple and complex tasks. In the simple task, participants first encountered a single cue that predicts a feature of an upcoming stimulus. After a delay of 500 ms, the target stimulus was presented until response or a deadline. Participants judged if the cue matched the target with button press. Visual feedback was provided after another 500 ms delay. The complex task is identical to the simple task except for two changes. First, participants encountered two cues instead of one. Second, participants decided whether only one cue matched the target (i.e., the 'exclusive or' rule). *(C)* Illustration of experimental design. *(D)* Predicted dynamics of task representations in the complex task phase. Participants were expected to associate simple task representations to build a complex task representation and to further generalize two complex tasks via associative inference. As a result, we predicted that the generalizable complex tasks present faster decrements in RT than the control complex tasks. *(E)* Block-wise RTs (left, group mean and standard error of means, SEM) and slope of RTs (right, group mean and SEM, dots represent individual slopes) of complex tasks for each condition during the test stage.

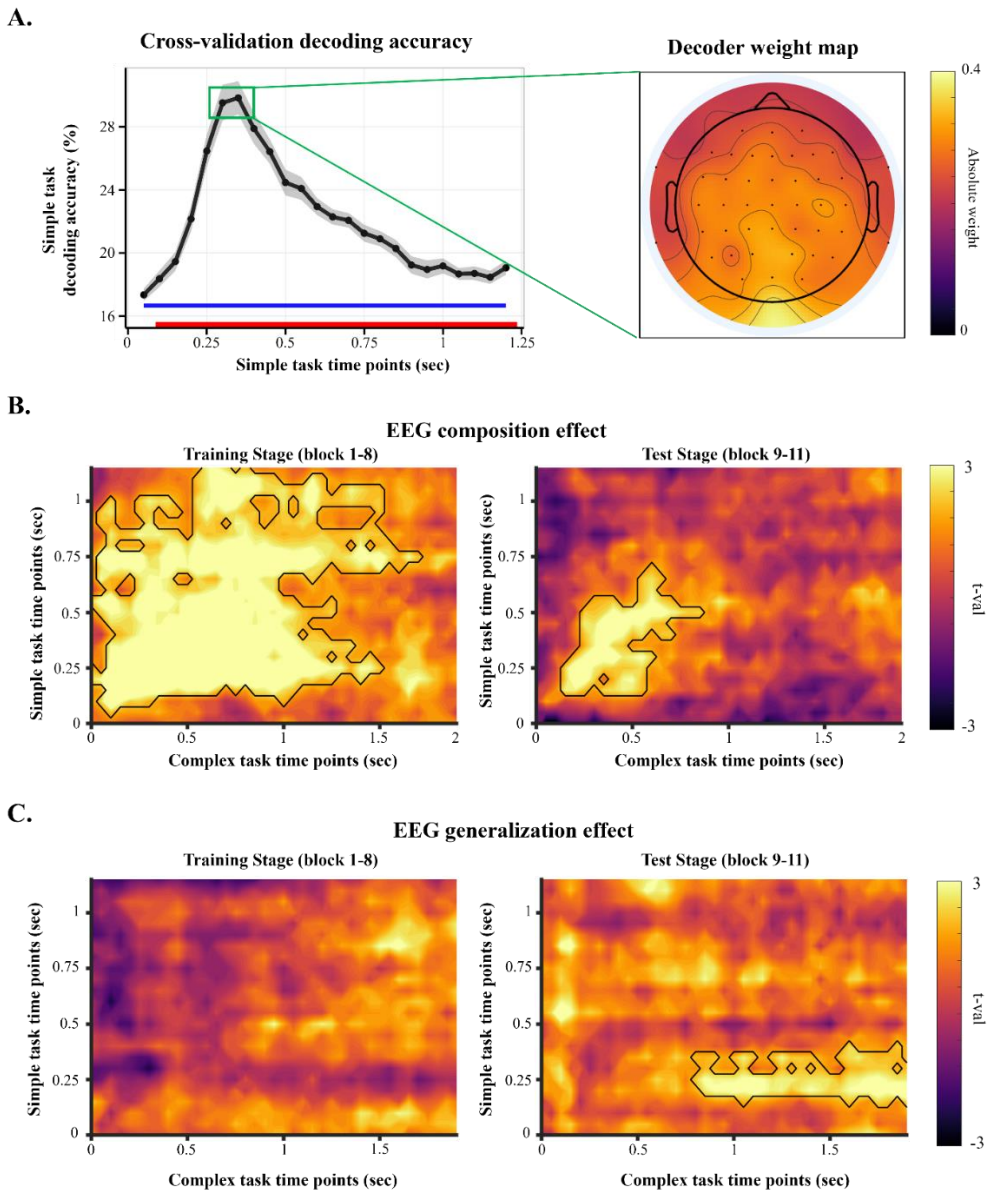

Figure 2. EEG decoding results. *(A)* Simple task EEG data decoding results. Left: Decoding accuracy over time within a trial. The grey shade indicates SEM of the decoding accuracy. The blue line indicates chance-level decoding accuracy. The red line indicates time points showing significantly above-chance decoding accuracy. Right: Visualization of mean electrode-wise contribution to the decoders. *(B)-(C)* EEG composition and generalization effects of complex task EEG data. In each panel, decoding performance (averaged t-values from trial-level decoding analysis, see Method) is plotted as a function of the time point of training (Y-axis) and the time point of test (X-axis). *(B)* Results of decoding constituent simple task representations from complex tasks in the training stage (left panel; e.g. A & B in AB) or in the test stage (right panel; e.g. A & C from AC). *(C)* Results of decoding the latent simple task in the training stage (left panel; e.g. C in AB), or in the test stage (right panel; e.g. B in AC). The time points are time-locked to the stimulus onset.

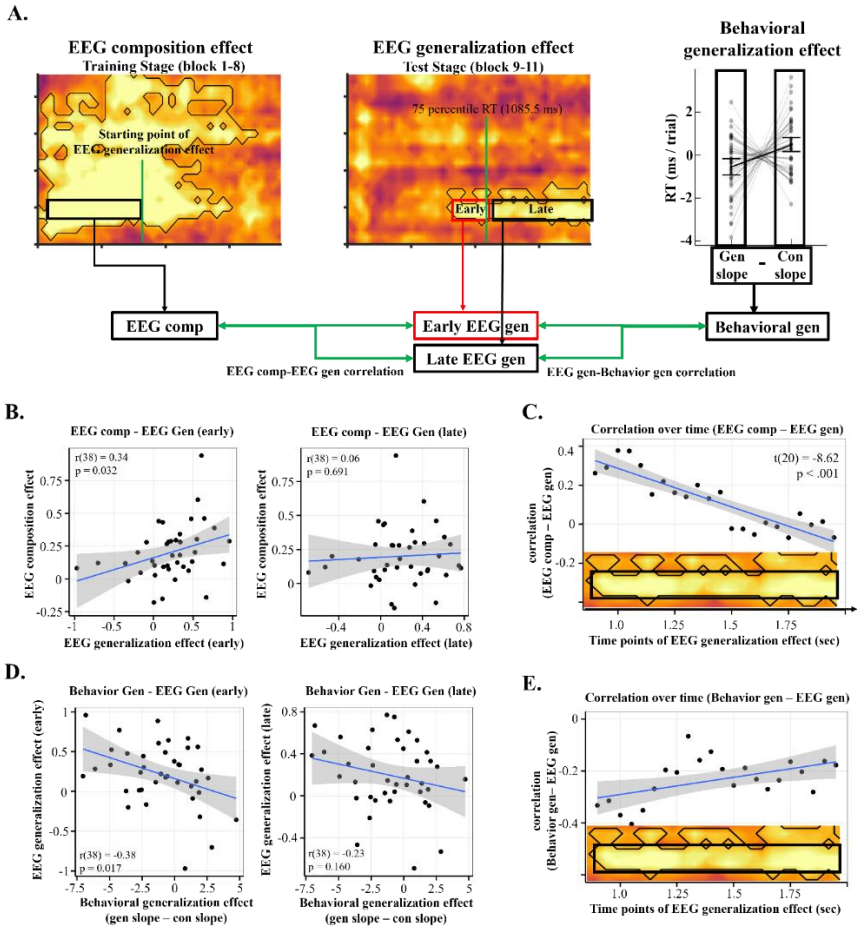

Figure 3. Correlation analysis results. *(A)* Data used for correlation analyses. For each individual, EEG composition effect (left panel), EEG generalization effect (center panel), behavioral generalization effect (right panel) were used. The EEG composition effect was calculated by averaging single-trial t-values from the decoders trained on simple task time points between 200 ms to 300 ms and tested on training phase complex task time points between 100ms and 850ms.. The EEG generalization effect was calculated by subject-mean single-trial t-values from the decoders to be trained on simple task time points between 200 ms to 300 ms and tested on complex task time points between 850 ms to 1,100 ms (early) or between 1,100 ms to 2,000 ms (late). The complex task time points were divided based on the 75% percentile RT (1085.5ms). The EEG composition effect was calculated by subject-mean single-trial t-values from the decoders to be trained on simple task time points between 200 ms to 300 ms and tested on complex task time points between 100 ms to 850 ms. The behavioral generalization effect was calculated as the difference in RT slope (Fig. 1E) between generalizable and control conditions. *(B)* Scatter plots showing individual EEG composition and early (left) or late (right) EEG generalization effects. *(C)* Correlation between overall EEG composition effect and EEG generalization effect at each time point. *(D)* Scatter plots showing individual EEG generalization effect and early (left) or late (right) behavioral generalization effect. *(E)* correlation between behavioral and EEG generalization effects at each time point. The time points are time-locked to the onset of the stimulus.

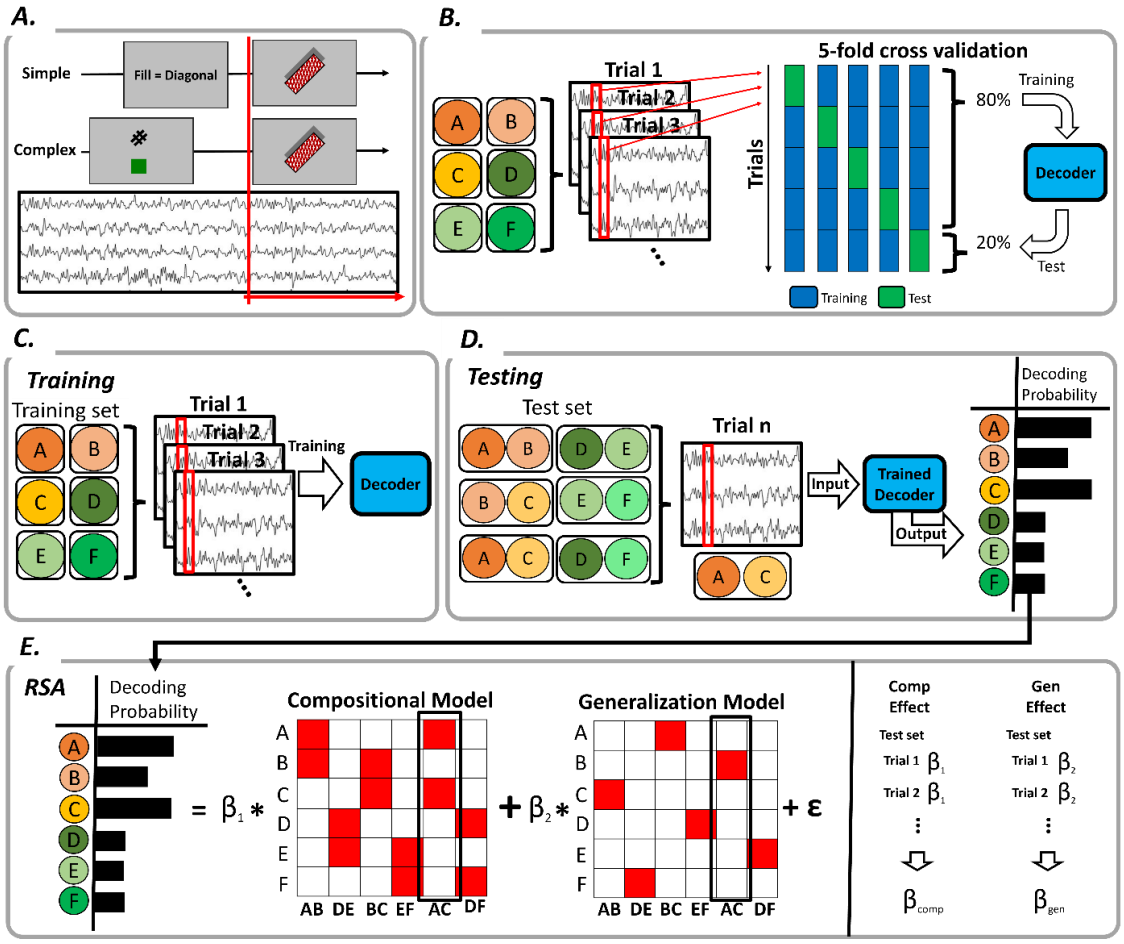

Figure 4. EEG decoding analysis procedure. (A) Illustration of the epochs for the EEG data analysis. The EEG epochs were time-locked to the onset of the target (red lines) to de-confound the effect of perceptual information on decoding performance. (B) Decoding analysis using simple task EEG data. At each time point, we trained and tested a decoder of six simple tasks using 5-fold cross-validation. (C) For cross-phase decoding analysis, the decoder was trained on simple task EEG data. (D) The trained decoder was tested using complex task EEG data and generated a decoding probability for each of the six simple tasks. (E) The RSA approach. The logit-transformed decoding probability was regressed against two model predictors representing composition and generalization effects. This procedure was repeated on each complex task trial. We averaged the trial-wise composition and generalization effects for each participant and then ran a one sample t-test against 0 for the statistical analysis.

Table 1. Behavioral results of the test stage in complex task phase. Group mean (SEM) of accuracy and residual RT in each condition and block of the test stage.

Condition	Block	Residual RTs (ms)	Accuracy (%)
Gen	9	1.80 (3.51)	78.54 (7.69)
	10	-7.31 (6.17)	79.17 (7.92)
	11	-8.31 (4.75)	81.77 (7.07)
	Slope	-0.54 (0.23) per trial	1.67 (1.11) per block
Con	9	4.34 (6.03)	74.69 (7.19)
	10	1.25 (7.77)	79.48 (5.80)
	11	11.67 (4.00)	79.17 (7.55)
	Slope	0.48 (0.22) per trial	1.88 (1.11) per block