Word count: 10'383

# What changes with representational change?

Mirko Thalmann*, Theo Schäfer**, Stephanie Theves**, Christian Doeller**, and Eric Schulz*

*Max Planck Institute for Biological Cybernetics

Tübingen - Germany

**Max Planck Institute for Human Cognitive and Brain Sciences

Leipzig - Germany

## Author Note

Correspondence concerning this article should be addressed to Mirko Thalmann, Research Group Computational Principles of Intelligence, Max Planck Institute for Biological Cybernetics, Max-Planck-Ring 8, 72076 Tübingen, Baden-Württemberg, Germany. E-mail: mirkothalmann@hotmail.com

Experimental scripts, modeling scripts, and analysis scripts are available on the following osf webpage: https://osf.io/pgyr4/files/

Pre-registrations of Experiments 2-4 are available on the following osf webpage: https://osf.io/pgyr4/registrations/

**What changes with representational change?**

Research from several areas suggests that mental representations adapt to the specific tasks we carry out in our environment. In this study, we bring the idea of adaptive representational change to a systematic test. Thereby, we propose a computational model, which portrays representational change as an adaptation to specific task goals. We test the predictions of the model in four behavioral experiments using healthy young adults as participants. In each experiment, we assess participants' baseline representations in the beginning, then expose participants to one of two tasks intended to shape representations differently according to our model, and finally assess any potential change in representations. Across the four experiments, we vary the measurement level of mental representations. We use a continuous reproduction task to measure visual-perceptual representations, pairwise comparison judgments to measure perceived similarities between stimuli, and pairwise category comparison judgments to measure differences between task-specific representations. The results suggest that representational change is task-specific. That is, it is not the low-level representation of an object, which changes when we practice a task, but rather a task-specific interpretation of that object. The results align with recent findings that performance improvements due to task practice do not generalize broadly over and above the specific task practiced.

## 1 Introduction

How does our cognitive system react when we learn to perform a task, for example to categorize a flavor as belonging to a fruit. A simple solution is to rely on a strategy that we already know. For example, an allergic reaction against the specific fruit may tell us that it is a Kiwi. A harder solution is to learn a strategy, which maps fruit taste profiles to fruit identities, from the ground. This could be achieved, for example, by using the rule that a highly acidic fruit is a grapefruit (Ashby & Gott, 1989), by using individually stored episodes of fruits previously eaten to generalize to the current fruit (Nosofsky, 1986), by using representations of prototypical fruits and go with the closest prototype (Homa et al.,

1982; Minda & Smith, 2001), or by using a combination of the latter two (Vanpaemel & Navarro, 2007). The speed of the involved basic cognitive processes may even increase when we use the same strategy over extended periods (Case et al., 1982). A different, intriguing idea is that the representations of the fruits themselves change with learning.

The idea of representation learning has gained a lot of traction in the field of machine learning in the last decades. The reason for that is that the performance of machine learning algorithms in terms of predictive accuracy heavily depends on the representation of the data (i.e., the features, Bengio et al., 2013). For example, the success of speech recognition models depended to a large degree on learning good representations of speech signals. By analogy, it could be assumed that achieving high performance on a given task by humans also depends on learning good mental representations. A main question to be tackled then is how representations become more favorable to carry out a task. Or in other words, how do representations change? Research targeted to understanding mental representations and their change broadly comes from three fields: neuroscience, cognitive science and cognitive psychology. Despite continued research in these areas and findings interpreted in favor of the idea that representations change (Dubova & Goldstone, 2021; Goldstone et al., 2001; Karagoz et al., 2022) it is currently not clear what - if anything - changes with representational change.

We start this article with a short literature review summarizing different lines of previous research targeting the topic of representational change from different viewpoints. Even though all of these lines touch the topic of representational change, only few have examined it explicitly. In our current research we take inspiration from all of these lines and bring the idea of representational change to test.

We develop a test bed on how representational change can be examined behaviorally. We further introduce a computational model, which portrays representational change as a response to task-specific practice. The model makes qualitative predictions on how representations change, which we test in a series of four behavioral experiments, three

of them pre-registered. The results are mostly in disagreement with the model's predictions and show that representations of objects do not change according to practice in a category learning task. Thus, participants' representations of objects seem to be relatively stable over time. Over the progression of the four experiments we varied the task intended to measure representations. When the task was designed to measure predominantly perceptual details about objects, participants' responses did not change as a function of task practice. However, when the task highlighted higher-level information about the stimulus as category membership, participants' responses changed. We interpret the pattern of results across the experiments in such a way that representational change is narrow and domain-specific. That is, whereas higher-level representations of objects such as their interpretation within a given task context change, their perceptual representations remain relatively stable over time.

## 2 Background

### 2.1 Neuroscience

Neuroscientific studies examined how brain responses to the same stimuli change depending on the context in which they appear (e.g., Schlichting et al., 2015, Wammes et al., 2022). More specifically, several studies explored changes in neural activation patterns in the hippocampus when concepts are learned using a method called repetition suppression. Repetition suppression is assumed to measure habituation of single cells or neural populations to repeated stimuli or repeated feature information (Barron et al., 2016). Those studies show that the brain reacts differently to the same stimulus depending on what stimulus information has been learned to be important to infer a concept. Based on that, they suggest that representations of stimuli adapt to how the stimuli relate to the task goals. Mack and colleagues show that hippocampal object representations during categorisation are modulated by attention to the currently decision-relevant feature (Mack et al., 2016). Theves and colleagues show that the hippocampus specifically integrates categorization relevant feature dimensions in a common representational space, thereby

mapping distances between exemplars and category boundaries (Theves et al., 2019; Theves et al., 2020). To summarize, the neuroscientific literature suggests that the representation of a stimulus is shaped by the contexts, in which the stimulus appears, and by the concepts a person has learned to associate the stimulus with.

### 2.2 Categorical Perception

The idea that concept learning affects how we perceive the world has also been studied in the literature of categorical perception. The phenomenon of categorical perception describes the observation that two stimuli separated by a fixed distance in feature space are harder to discriminate when they come from the same category than when they come from different categories. This phenomenon was first observed in the domain of speech perception. For example, Liberman et al., 1959 found that discrimination between auditorily presented phonemes is more difficult when they come from the same phoneme category (e.g., two b's) than when they come from different phoneme categories (e.g., b vs. d). While the emergence of categorical perception for natural speech might be targeted from a developmental perspective, several studies examined it experimentally in the visual domain.

Goldstone, 1994 showed that participants' perceptual sensitivity to discriminate between visually presented squares varying in size and brightness increases after category learning. Importantly, the sensitivity increased particularly for the category-relevant dimension but not for the irrelevant one. Moreover, Goldstone et al., 2001 tested the idea that besides increased sensitivity, representations of stimuli become biased due to category learning. That is, representations of stimuli belonging to the same category may be rated as more similar to each other after category learning than before. The results, however, provided mixed evidence for the idea. Whereas within-category items were not rated as more similar afterwards than before, surprisingly between-category items were. Consistent with their theory was the observation, though, that the difference of similarity ratings from two stimuli to a neutral stimulus became smaller, when the two stimuli came from the

same category, but not, when they came from different categories. To summarize, research on categorical perception pursued the idea that mental representations of objects in our environment may be affected by how these objects were conceptually used in a different context.

### 2.3 Effects on Short-Term Memory

A third line of research examined influences from categorical knowledge on responses in short-term memory tasks. Several studies suggest that categorical knowledge assists in reproducing information from short-term memory. For example, Huttenlocher et al., 1991 presented participants with a location on a circle to remember for immediate recall. An important finding was that participants' responses were biased towards certain stereotypical angles (45 degrees, 135 degrees and so on) and towards locations halfway between the center of the circle and its circumference. In a similar vein, Hasantash and Afraz, 2020 presented participants with one target color patch in one of two experimental conditions. In the simultaneous matching condition, participants adjusted a variable color patch to match the simultaneously presented target color patch. In the sequential matching condition, they adjusted the variable color patch after the target patch had disappeared for a ten second retention interval. At the end of the experiment participants were given a color naming task. The authors found that the precision of participants' responses in the sequential matching task but not in the simultaneous matching task correlated with the size of the color vocabulary. Moreover, precision was higher in regions of the color space where the average density of the color vocabulary was higher. These two studies render the possibility likely that the way how we represent stimuli with continuous feature dimensions in short-term memory may be affected by previous experience with these stimuli in different task contexts. For example, painters may have developed a fine-grained color vocabulary to effectively communicate with colleagues. If we think of short-term memory as providing access to representations for goal-directed processing (Oberauer, 2009), the way how we learn concepts essentially may affect how we interact with the world.

The lines of research introduced in the previous three sections suggest the possibility that mental representations, and at the same time the way we represent the world, are shaped by task-specific practice. In the following, we introduce a controlled setup, which allows us to observe such a representational change in an experimental behavioral setting. We then introduce a computational model, which makes specific predictions for that setup, and delineate the road map, how we test the model predictions in a series of four experiments.

## 3 A Model of Representational Change

In the following, we introduce a computational model that portrays representational change as an adaptive response to task-specific practice. In order to explain the model and simulation results of the model we first give a brief sketch of the experimental setup. Knowledge about the latter should help to understand the decisions we took in our modeling approach.

### 3.1 Experimental Setup

An important point to show in our approach is that representations change differently according to different tasks. We chose two tasks that differed in their respective goals. One task was a category learning task, in which biased representations are particularly helpful. The other task was a sequential comparison task. Biased representations are not helpful in the latter, because participants have to learn the identity function to perform well. In both tasks, participants were confronted with the same 100 two-dimensional stimuli designed and tested by Schäfer et al., 2022. The two dimensions of the stimuli were represented as the spikiness of the head and the fill of the belly of "monsters" (see Figure 1). We defined the 100 stimuli according to equally-spaced locations in the two-dimensional feature space (see Figure 2). In the category learning task, participants were required to learn to assign these stimuli into two categories (Experiment 1, category structure similar as in Milton and Pothos, 2011; Nosofsky et al., 2005; Schäfer et al., 2022) or into four categories (Experiments 2-4, similar as in Nosofsky et al., 2005).

In the sequential comparison task, participants were required to judge how similar two consecutively presented stimuli were to one another.

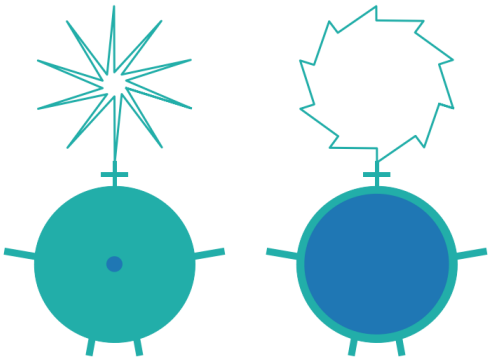

**Figure 1**

*Left: Stimulus with minimum values on both feature dimensions. Right: Stimulus with maximum values on both feature dimensions.*

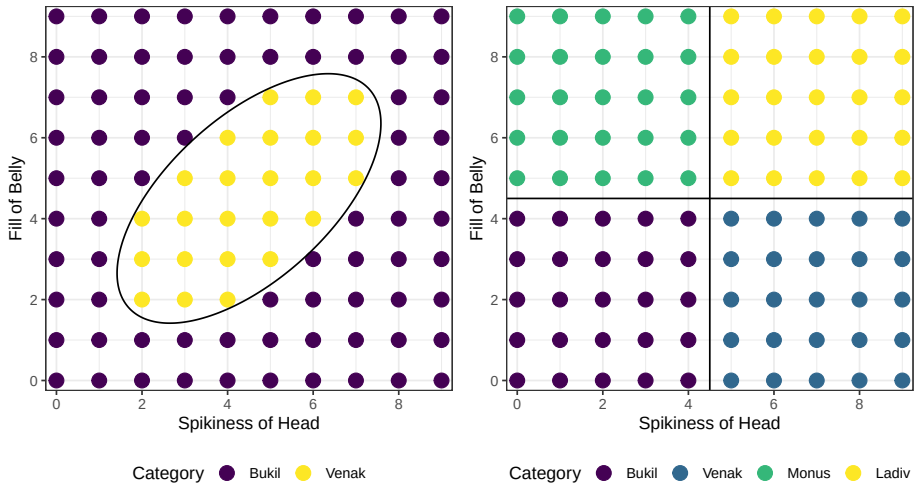


**Figure 2**

*Two-dimensional category structures in Experiment 1 (left) and Experiments 2-4 (right).*

All experiments followed the same sequential logic. Before and after practice in one

of those tasks we measured representations. Measuring representations before and after task practice with the same stimuli but different goals allowed us to get an estimate of representational change by the logic of subtraction. Any task-specific differences can be attributed to representational change due to the different goals. The way how we measured representations varied across experiments, though. We progressively increased the cognitive level of representations across experiments. In Experiments 1 and 2, we measured visual-perceptual representations using a continuous reproduction task (Pertzov et al., n.d.; Souza et al., 2014). In Experiment 3, we used a simultaneous comparison task, in which participants were required to rate how similar two stimuli were to each other. In Experiment 4, we used a simultaneous comparison task, in which participants were required to rate how likely two stimuli were to come from the same category.

### 3.2 Model Specification

The computational model consists of three sequential stages. In the first stage, one stimulus out of the set of 100 stimuli is presented. The perceptual representation is modeled via a sample from a prior distribution, implemented as a multivariate normal with uncorrelated variances:

$$\mathrm{x}_i \sim mvnormal(\mu_i, \Sigma)$$
$$\mu_i = [\mathrm{mod(i/10)} + 1, \lceil i/10 \rceil + 1] \quad \text{for} \quad \mathrm{i} \quad \text{in} \quad 0{:}99$$
$$\Sigma = \mathrm{I} * sd_{prior}$$

The idea that perceptual representations are stochastic follows population coding accounts (e.g., Bays, 2014; Pouget et al., 2000). In the second stage, perceptual representations are mapped to responses in the category learning task and in the sequential comparison task. For the category learning task, we used three different models of category learning: an exemplar model, a prototype model, and a rule-based model. We used the generalized context model as a representative exemplar model (Nosofsky, 1986), a naïve Bayes classifier (e.g., John and Langley, 1995) as a prototype model, and a self-constructed

rule-based categorization model. The latter was self-constructed because the idea that representations of stimuli themselves change contrasts with previous implementations of rule-based categorization models, such as decision-bound models (e.g., Maddox and Ashby, 1993). Each model takes the perceptual representation as an input and returns category probabilities $p(cat_k|\ x_i)$ for every category k out of m categories. The exemplar model, the prototype model, and the rule-based model arrive at those probabilities by calculating

$$\frac{\sum_{i=1}^{n_k} \eta_{ij}}{\sum_{j=1}^{m} \sum_{i=1}^{n_m} \eta_{ij}},$$

$$\frac{p(x_i|\ cat_k) * p(cat_k)}{\sum_{j=1}^{m} p(x_i|\ cat_j) * p(cat_j)},$$

$$\sum_{i=1}^{N_{thx}} \int_{lo_i}^{hi_i} mvnormal(x_i, \Sigma)\, dx_i$$

$$\Sigma = I * sd_{prior}$$

, respectively. For simplicity, and because both dimensions were equally important for category learning, we omitted response bias parameters in the generalized context model. After feedback has been provided, the model decides in the third stage whether the perceptual representation is advantageous for performance. An important feature of the model is that representations are more likely to be stored in memory when they are helpful in the given task.

We varied the definition of helpfulness across two sampling algorithms. The first algorithm accepted every perceptual representation if it improved the true category probability:

$$if \quad p(cat_k|\ x_i)_t > p(cat_k|\ x_i)_{t-1}$$

, with k being the index of the true category and t indexing the time point within the experiment. The second algorithm used a Metropolis-Hastings acceptance mechanism. A uniformly sampled value s between 0 and 1 was compared to the ratio of the category probability after feedback had been provided to the category probability before feedback has been provided. Whenever the sampled value was below that ratio, the perceptual representation was accepted:

$$s \sim uniform(0,1)\ if \quad \text{s} < \frac{p(cat_k|\ x_i)_t}{p(cat_k|\ x_i)_{t-1}}$$

Accepting or rejecting samples in the sequential similarity task works similarly. The model accepts every perceptual representation if it is most likely under the true stimulus prior (improvement sampling). For the Metropolis-Hastings scheme, the model first calculates a normalized cumulative distribution of the likelihoods of the perceived stimulus given all true stimulus priors. It then again uniformly samples a value between 0 and 1, which lands on one stimulus portion of the cumulative distribution. The model then accepts the current representation under that stimulus prior.

We conducted simulation studies with different model implementations in order to get qualitative predictions for representational change in the three different tasks. Two of the tasks were the category learning tasks with different category structure (i.e., ellipse or squared). The third task was the sequential comparison task. For the two category learning tasks, we varied the category learning model, the sampling scheme, and whether perceptual representations with values outside of the borders of the stimulus space exist (i.e., $< 0$ and $> 9$ in Figure 2). The latter was intended to explore potential effects that arise solely due to the edges of the feature space. For example, perceptual representations of stimuli close to the edges are more likely to be rejected solely due to the fact that a larger portion of their probability density is located outside of the feature space. For the sequential comparison task, we varied the sampling scheme and again whether samples

outside of the feature space could be accepted or not. The standard deviation of the prior distribution was fixed to .75 for all simulations. We ran every simulation for 5000 trials.

The main prediction of the model is that representations change in the category learning task, but not in the sequential comparison task. As can be seen in Figure 3, mostly representations of stimuli close to the decision boundaries are predicted to change due to categorization practice. That is, these representations are pushed away from the decision boundaries, which leads to representations from closed categories being assembled more closely to each other. The change is qualitatively similar across the two category learning structures. When translated to the average distance of representations to the associated category center, category learning decreases that distance for closed categories (i.e., Venak category in ellipse structure and any category in squared structure) and increases that distance for the residual category in the ellipse setup (i.e., Bukil category). These average predictions are displayed in the right panels of Figure 3.

A second point to be mentioned is that the qualitative predictions of the model are relatively stable across different model implementations. The predictions do not depend on the used category learning model. That is, whether people use rules, prototypes, or exemplars to categorize stimuli affected the predictions of the model only marginally. Neither did the simulation results vary qualitatively as a function of the sampling algorithm. In general, though, the predicted effects were quantitatively larger for improvement sampling than for Metropolis-Hastings sampling. We plot exemplary simulation results for a prototype model with the improvement sampling algorithm without constraint of the feature space in the right panels of Figure 3.

## 4 Experiments

All four experiments were designed to test the qualitative prediction of the computational model that representations change differently according to task practice in the category learning task and in the sequential comparison task. The main difference between experiments was that we varied the measurement level of representations, which

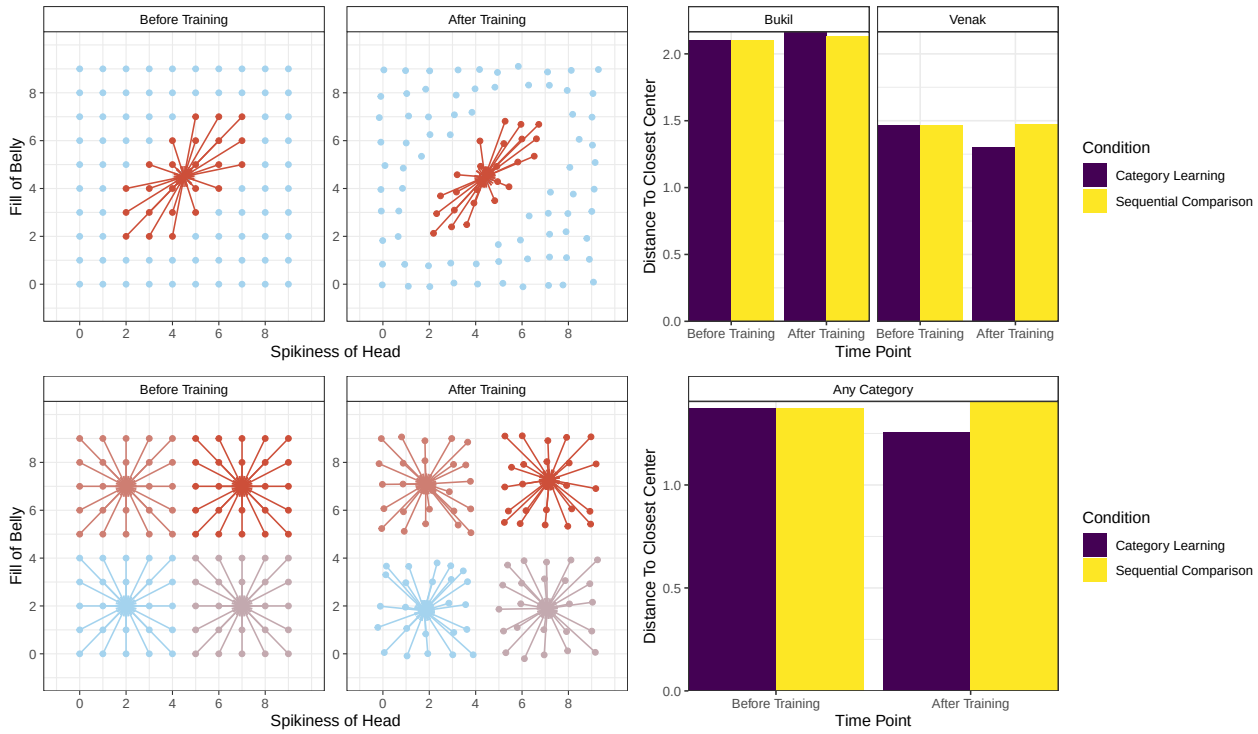


**Figure 3**

*Top Left: Prior distributions and posterior distributions of stimuli in the two-dimensional feature space before and after category learning with the ellipse category structure, respectively. Top Right: Average distance to the closest category center before and after the category learning task or the sequential comparison task plotted separately for the two categories. Bottom: Same as top row, but for experimental design with squared category structure. Note. Because representational change is predicted to be similar for all four categories, they are collapsed in the right plot.*

should allow us to observe at what level a potential representational change happens. In Experiments 1 and 2 we started by measuring visual-perceptual representations. For that purpose, we used a short-term memory task, because we assumed that the potential to detect an effect would be larger in such a task than in a fully perceptual task (Donkin et al., 2015; Hasantash & Afraz, 2020). We opted for the measurement of representations on a higher level in Experiment 3 because we did not observe a change in visual-perceptual representations in Experiments 1 and 2. In an attempt to do so, we measured

representations via similarity judgments of simultaneously presented stimulus pairs. Even though participants may primarily use the values on the feature dimensions to generate a similarity judgment, we assumed that category membership may provide additional information for that judgment. As we also did not observe any signature of representational change in Experiment 3, we made an attempt to further emphasize category membership in Experiment 4. Therefore, participants were required to say how likely they thought the two simultaneously presented stimuli belonged to the same category. Only in the latter case, when category membership was directly emphasized, we observed representational change.

### 4.1 General Method

All experiments used a two-groups between-subjects design. Participants were randomly assigned to one of two groups. Each group went through a set of three sequential stages. In the first stage, we attempted to get a baseline measurement of the representations of the full stimulus set. In the second stage, participants carried out the secondary task. One group carried out the category learning task as a secondary task, the other group carried out the sequential comparison task. In the third stage, we again measured representations of the full stimulus set. Essentially, representational change should be reflected in a detectable performance difference between stage three and stage one. Whereas the secondary tasks stayed the same across all four experiments (with one modification in the category structure, though), we varied the task to measure representations across the four experiments. Details are explained in the experiment-specific methods sections.

Participants for all experiments were recruited using the prolific platform (prolific.co). Fluent speakers of English aged between 18 and 35 with minimum approval rates of .9 and 1 for Experiments 1 and 2-4, respectively, and a minimal number of previous submissions of 5 and 20 for Experiments 1 and 2-4, respectively, were eligible to participate. By not further restricting participation in our experiments (e.g., to certain countries), we expect the results to generalize to young healthy adults in general. All

participants agreed to take part in the study and were informed about the general purpose of the experiment. Experiments were performed in accordance with the relevant guidelines and regulations approved by the ethics committee of the University of Tuebingen (protocol nr. 701/2020BO, study title: Experimente zum Sequenz- und Belohnungslernen). Experiments were presented to participants using a combination of HTML, javaScript, and CSS with custom code. After a presentation of the instructions, participants were required to complete a comprehension questionnaire. Only upon responding correctly to all questions, they could proceed to the main part of an experiment.

The stimuli, designed by Schäfer et al., 2022, were the same monsters in all four experiments, which differed from each other according to two continuous dimensions: the fill of their belly and the spikiness of their head (see Figure 1 for two examples). Belly fill was defined as the radius of the circle inside the body, which resulted in a blue area of varying size. Head spikiness was defined as the radial distance between ten inner and ten outer vertices. These vertices were located between ten equally spaced segments of two concentric circles with different radii. In conjunction, they formed the ten spikes of the head when they were connected by a line. Values in both dimensions could be parametrically varied according to 100 steps (e.g., from an almost empty belly to a full belly). Participants were only exposed to a subset of 100 combinations in the two-dimensional feature space. We created these 100 stimuli in the following way: in each dimension we cut 9 steps from both ends. Then, we cut the remaining values from 10 to 91 into 9 equally sized segments of 9 steps. The 10 values from 10 to 91 were fully crossed across the two dimensions yielding 100 stimuli in the feature space.

We used Bayesian statistics for data analysis to overcome some of the shortcomings associated with frequentist statistics (Wagenmakers, 2007). In particular, we rely on the Bayes factor (BF) to quantify the evidence in favor of an effect of interest. The BF reflects the posterior odds of two models if their prior odds are .5. Kass and Raftery, 1995 provide rough guidelines on how to interpret the strength of evidence of BFs: BFs ranging from

1-3.2 are not worth more than a bare mention, BFs between 3.2-10 provide substantial evidence, BFs between 10-100 are regarded as strong evidence, and BFs larger than 100 as decisive. To arrive at BFs for individual model parameters, we used the Savage-Dickey density ratio (e.g., Wetzels et al., 2009). This method provides BFs for nested models. In particular, it compares a model allowing the parameter of interest to vary freely to a model fixing the parameter to the null model. For hierarchical regression models, we first compared models varying in their random effects structure using the LOO method implemented in the loo R package (see Vehtari et al., 2017). We then used the winning model for inference about individual parameters (Matuschek et al., 2017). All Bayesian models were run in STAN (Carpenter et al., 2022) and accessed via the R programming language (R Core Team, 2022).

#### 4.1.0.1 Transparency and openness

All experimental scripts, raw data, and analysis scripts, as well as scripts for simulations with the computational model of representational change, are available on the following Open Science Framework webpage: https://osf.io/pgyr4/files. We pre-registered predictions, exclusion criteria, experimental designs, and more of Experiments 2-4 on the following OSF webpage: https://osf.io/pgyr4/registrations.

### 4.2 Experiment 1

In Experiment 1 we gauged representational change using a continuous reproduction task (Pertzov et al., n.d.; Souza et al., 2014). We wanted to measure a visual-perceptual representation of the stimuli to test whether category learning affects representations at a low cognitive level. Finding an effect on such a level would mean that representational change has a fundamental effect on how we mentally imagine objects.

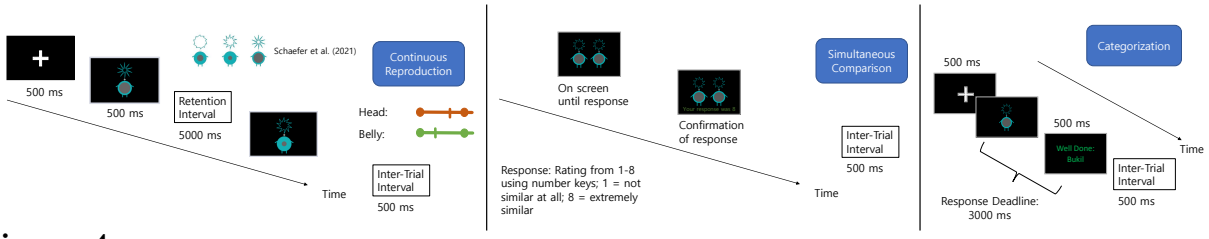


**Figure 4**

*Left: Procedure for the continuous reproduction task used in Experiments 1 and 2. The timing parameters were modified from Experiment 1. See text for details. Middle: Procedure for the simultaneous comparison task used in experiments 3 and 4. Right: Procedure for the category learning task used in all experiments. Note. The procedure for the sequential comparison task was the same as for the category learning task.*

### *4.2.1 Method*

#### 4.2.1.1 Participants

84 participants (10 unknown, 45 women, 29 men) completed one session lasting approximately 90 minutes. They received a base payment of 9.80 GBP and an additional performance-dependent bonus of up to 5.20 GBP. 12 participants quit the experiment in between or provided incomplete data. We excluded 9 participants because their average distance from the true stimulus in the continuous reproduction task was larger than three standard deviations above the mean. We excluded 3 additional participants because they restarted the experiment after they had already progressed through a substantial portion of the experiment. Thus, 60 participants remained in the experiment, 32 in the experimental group, 28 in the control group.

#### 4.2.1.2 Materials and procedure

All 100 stimuli were presented once in the first session of the continuous reproduction task and once in the second session. Presentation order of the stimuli was randomized. For the whole duration of the continuous reproduction task, the background color of the window

was white with the exception of an upward-facing rectangle in the center of the screen filled in black color. That setup was intended to increase the contrast between background and stimuli. A trial started by the presentation of a fixation cross in white color in front of the black rectangle for 500 ms. After that, a stimulus replaced the fixation cross and was displayed for 750 ms followed by a 2000 ms blank retention interval. Then, an average stimulus with a value of 50 on both feature dimensions was displayed on the screen. Participants could change the values in each dimension quasi-continuously (i.e., according to 100 steps) with two sliders that were presented below the stimulus. Once participants were satisfied with their response, they pressed a button to submit their response. The next trial started after an inter-trial interval of 500 ms. In the beginning of the experiment, participants completed two practice trials to get used to the procedure.

We used an ellipse category structure in Experiment 1 (see left panel of Figure 2). We used this type of information-integration category structure because it assures that participants have to pay attention to both dimensions to categorize stimuli accurately (Ashby & Gott, 1989). The background coloring of the screen was the same in the category learning task as in the continuous reproduction task. A trial started with presentation of a fixation cross for 500 ms followed by presentation of a stimulus. Participants were given a response window of 3000 ms to respond. When they responded within that period, a green message told them “Well done: Category X!” when they responded correctly, and a red message told them “Category would have been: Category X” when they responded incorrectly. When they responded after 3000 ms, a message appeared on screen, which indicated that they should respond faster. Responses were still collected, though. After participants had responded, the next stimulus was presented after an inter-trial interval of 500 ms. Responses were given by pressing digits 1 and 2 on the keyboard. Procedures of the set of tasks used in Experiment 1 are visualized in Figure 4.

The category learning task lasted for 640 trials. The first 40 trials consisted only of examples from the ellipse category. In the instructions displayed to participants, we labeled

the ellipse category as the target category called “Venak” and the residual category as the non-target category called “Bukil”. We presented 40 trials only from the target category to initially familiarize them with the more specific category. We created the set of 40 stimuli by first shuffling the stimuli from the ellipse category randomly, appending two sets of them and selecting the first 40 stimuli from that sequence. The remaining 600 stimuli were created in the same way, though, by appending several randomly shuffled sets of each category and selecting the first 300 stimuli from both resulting sequences.

The procedure in the sequential comparison task was the same as in the category learning task. The 640 stimuli were created in the same way as above, though, without the constraint of having the same number of stimuli from different categories. Participants responded with digits 1-4 reflecting similarity judgments from not similar to very similar. In both secondary tasks, stimuli were presented in four blocks with an equal number of trials. In between blocks, participants could recover for one minute or skip the break and continue immediately.

### *4.2.2 Results*

#### 4.2.2.1 Category learning

Our main goal is to show that participants learned to discriminate between the categories over the 640 trials. As can be seen in the left panel of Figure 5, the initial 40 examples from the ellipse category led to a head start for this category compared to the residual category. We binned the remaining 600 trials into blocks of 20 trials and analyzed the data with a hierarchical logistic regression. The head start in the ellipse category was reflected in decisive evidence for the main effect of category (BF $>$ 100). However, accuracy in the residual category approached accuracy in the ellipse category over the course of the experiment. While there was no clear evidence for a main effect of trial (BF = .7), the BF for the interaction of category x group was decisive ($>$ 100). A potential point of concern is that participants were not able to discriminate between the categories sufficiently well on average, even after 640 trials of practice. Even though categorization accuracy was well

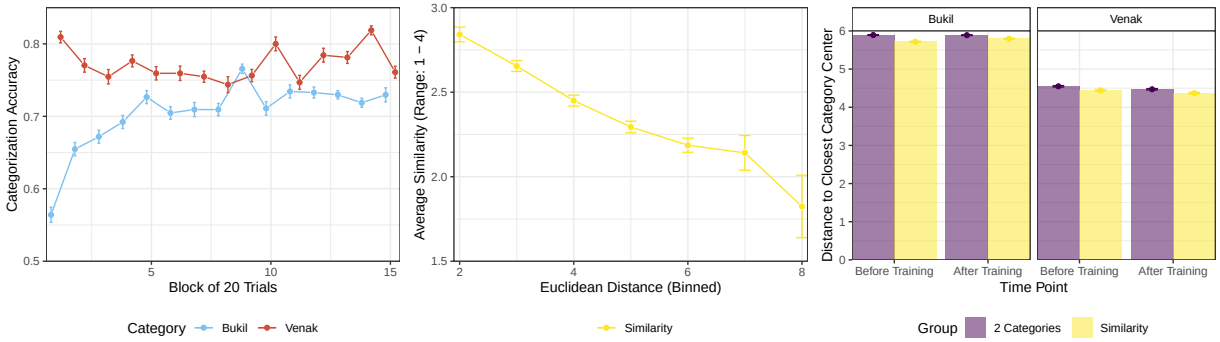


**Figure 5**

*Left: Average categorization accuracy plotted per block of 20 trials separately for the two categories in Experiment 1. Middle: Average similarity rating plotted against euclidean distance of subsequent stimuli in the sequential comparison task. Right: Average distance to the category center only for stimuli of the ellipse category (i.e., using the same stimuli for both groups) in the continuous reproduction task before and after the secondary task. Note: Error bars represent 95% within-subjects confidence intervals.*

above the chance level, it plateaued at .73 in the residual category and at .76 for the ellipse category in the last block of 20 trials.

#### 4.2.2.2 Sequential comparison

The main idea of the control condition was to expose participants to the same stimuli as in the experimental condition. If participants carried the task out as instructed, we would expect their similarity judgments to be predicted by the Euclidean distance between subsequently presented stimuli. To test that, we binned Euclidean distances into seven buckets and regressed participants' similarity ratings onto these binned distances in a hierarchical model (see middle panel of Figure 5). Subsequent stimuli were rated as more similar when their Euclidean distance in the two-dimensional feature space was smaller and vice versa, which was reflected in substantial evidence for the main effect of distance (BF = 4.1).

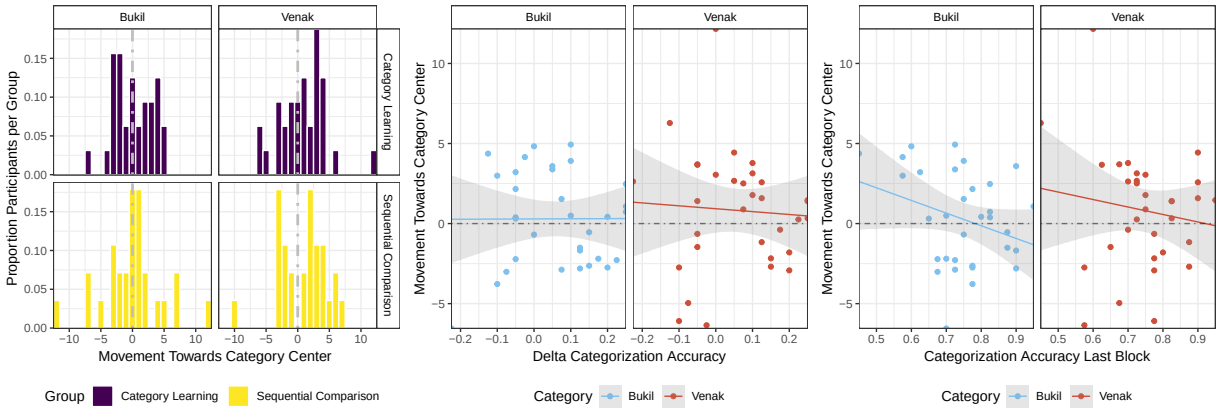


**Figure 6**

*Left: Histograms of by-participant averages of movements towards the category center plotted separately for the two categories (columns) and the two groups (rows) in Experiment 1. Middle: By-participant average movements towards the category center plotted against the average improvement in the category learning task . Right: By-participant average movements towards the category center plotted against the average final categorization accuracy in the category learning task.*

#### 4.2.2.3 Continuous reproduction

Our model predicted that distances to the category center should decrease for the ellipse category and increase for the residual category in the experimental group, but stay the same in the control group. That pattern is reflected in a three-way interaction between time point, category, and group on the distance measure. We broke it down to a two-way interaction on differences between distances after the secondary task and distances before the secondary task. Differences should become smaller for the ellipse category and larger for the residual category in the experimental group, but stay the same in the control group. The pattern is then reflected in the interaction between category and group. We tested the prediction in a hierarchical regression predicting difference scores using group and category as predictors. The BF in favor of the interaction was .28 providing substantial evidence against the hypothesis. There was also substantial evidence against the main effects of

category and group (BFs were .31 and .19, respectively). Together, the results suggest that representations were relatively stable. That stability is visible when inspecting the distribution of by-participant average movements towards the category center, which are scattered around zero in both groups for both categories (see left panel of Figure 6). Given that there was substantial variability in category learning success, we explored whether it was positively related to movements toward category centers in the continuous reproduction task. We measured categorization success using two different approaches: once as final accuracy in the last block and once as the amount of learning calculated via the difference between final accuracy in the last block and initial accuracy in the first block. The models were implemented as fixed-effects linear regressions. Movements away from the category boundary would be reflected in an interaction between categorization success and category. That is, representations from the ellipse category should be pulled towards the category center more for successful learners than for less successful learners, whereas representations from the residual category should be pushed away from that center more for successful learners than for less successful learners. However, posterior distributions of the interaction in both models (using final categorization accuracy and the amount of learning) were centered close to zero and the respective BFs provided evidence for the Null hypothesis (.48 and .48 for final accuracy and the amount of learning, respectively). The absence of these effects can also be seen by visually inspecting the middle and right panels of Figure 6. Regression lines with a slope of zero are well within the shaded region of 95% frequentist confidence intervals.

#### *4.2.3 Discussion*

Experiment 1 tested the idea that visual-perceptual short-term memory representations become biased according to practice in a category learning task. We tested representational change on a low cognitive level because change on that level should feed forward and affect any cognitive operation using these representations. In addition, previous research showed that short-term memory representations are affected by

categorical knowledge (Hasantash & Afraz, 2020; Huttenlocher et al., 1991). The results provided, however, initial evidence against the idea of representational change. Representations did not change as a function of the secondary task at the group level. We additionally tested the idea that the amount of representational change is affected by the amount of learning in the categorization task. The results also provided evidence against that idea. Even participants who learned the categories well did not respond according to representations as predicted from the model.

One observation when analyzing individual differences in the amount of category learning was that a substantial proportion of the participants did not improve categorization performance at all over the course of approx. 550 trials (see middle panel of Figure 6; points are scattered around a delta of 0). That suggests that many participants only learned relatively imprecise representations of the categories because the category structure may have been too difficult to learn on average.

### 4.3 Conceptual Changes in Experiments 2-4

A prerequisite for representational change in our model is that the category structure is learned sufficiently well. One possible reason for the absence of representational change in Experiment 1, therefore, is that participants did not learn to discriminate the two categories well enough from each other. Categorization accuracy plateaued at roughly .75 on average. We, therefore, applied two changes to the category learning task. First, we replaced the ellipse category structure with a potentially simpler squared category structure (see right panel of Figure 2). Second, we motivated participants to perform well in the category learning task with two measures. Participants could only proceed to the third part of the experiment (i.e., the second measurement of representations) when they performed above one of two predefined performance thresholds. Additionally, they received a monetary reward when they surpassed one of the two thresholds. We also pre-registered the study designs, hypotheses, model predictions, analyses, exclusion criteria, and more in individual OSF pre-registrations for each experiment (https://osf.io/uvgc3).

## 4.4 Experiment 2

### *4.4.1 Method*

#### 4.4.1.1 Participants

192 participants (23 unknown, 69 women, 100 men) completed one session lasting approximately 80 minutes. They received a base payment of 9 GBP and an additional performance-dependent bonus of up to 6.50 GBP. 71 participants did not reach the predefined performance criterion for their secondary task. We further excluded 2 participants because their average distance from the true stimulus in the continuous reproduction task was larger than three standard deviations above the mean. Thus, 119 participants remained in the experiment, 58 in the experimental group, 60 in the control group.

#### 4.4.1.2 Materials and procedure

Materials and procedure of the continuous reproduction task were the same as in Experiment 1 with the following procedural changes: We reduced stimulus presentation duration to 500 ms and increased the duration of the retention interval to 5000 ms. We assumed that a potential influence from category knowledge on visual-perceptual representations is more likely with a shorter presentation duration and a longer retention interval (Donkin et al., 2015). We also changed the initial location of the two sliders to reproduce the two feature values after stimulus presentation to random locations. Additionally, we applied several changes to the category learning task. Instead of two, there were now four categories defined as the four quadrants of the feature space (see the right panel in Figure 2). We also reduced the number of trials in the secondary task (category learning and sequential comparison) to 400. Selection of the stimuli for the secondary task happened in the same way as in Experiment 1. Again, in both secondary tasks, stimuli were presented in four equally-sized blocks with the option of having a one-minute break. The initial 40 trials with only stimuli from the target category were dropped.

### *4.4.2 Results*

We used the same hierarchical models as in Experiment 1 for the category learning task and the sequential comparison task in all the remaining experiments. Performance in the category learning task improved over blocks of 20 trials (BF $>$ 100 for the main effect of block). The additional measures to increase performance in the category learning task also paid off. Final accuracy was at approximately .82 averaged over the four categories (see left panel of Figure 7) with chance performance being at .25. Similar to Experiment 1, participants in the control group engaged well in the sequential comparison task. Euclidean distance binned into seven buckets predicted similarity ratings (BF $>$ 100).

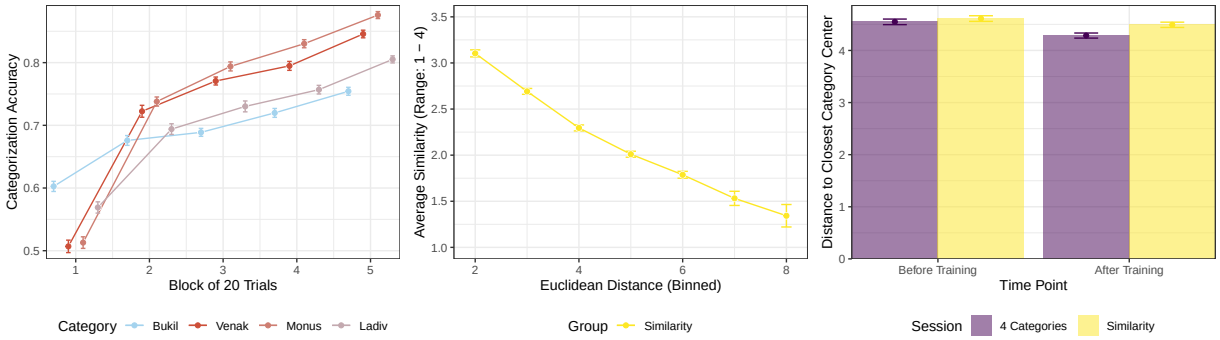


**Figure 7**

*Left: Average categorization accuracy plotted per block of 20 trials separately for the two categories in Experiment 2. Middle: Average similarity rating plotted against euclidean distance of subsequent stimuli in the sequential comparison task. Right: Average distance to the closest category center in the continuous reproduction task before and after the secondary task. Note: Error bars represent 95% within-subjects confidence intervals.*

We analyzed the representational change in the continuous reproduction task with a hierarchical model predicting distances (square-root transformed) to the closest category center using group and time point as predictors. The computational model predicts an interaction between time and group because the distances should become smaller for the category learning group, but not for the sequential comparison group. The results, however, provided substantial evidence against that interaction (BF = .13) reiterating the

null findings from Experiment 1. Visual inspection of the right panel of Figure 7 suggests that the pattern of distances is at least qualitatively as predicted by the model. Therefore, we explored individual differences in the expected pattern in two further analyses.

The first analysis again tested whether success in the category learning task was positively related to movements toward the closest category center in the continuous reproduction task. Because movements are predicted to be qualitatively similar across categories, success in the category learning task should be positively related to movements towards the closest category center in the continuous reproduction task. The analyses showed again no evidence for that idea (BF = 1.00 and BF = 1.00 for final accuracy and the amount of learning, respectively).

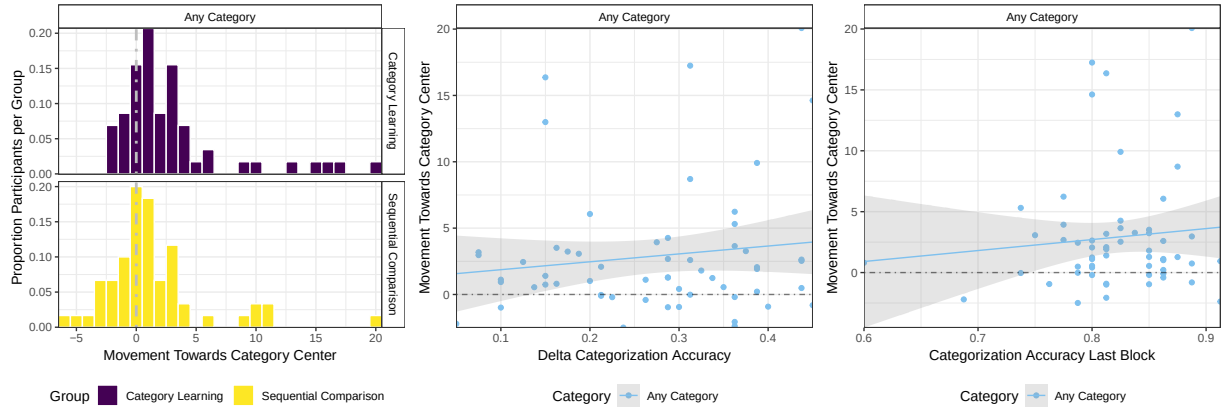


**Figure 8**

*Left: Histograms of by-participant averages of movements towards the category center plotted separately for the two categories (columns) and the two groups (rows) in Experiment 1. Middle: By-participant average movements towards the category center plotted against the average improvement in the category learning task. Right: By-participant average movements towards the category center plotted against the average final categorization accuracy in the category learning task.*

The second analysis asked the question of whether continuous reproduction responses are the result of a mixture of stable responses and changed responses. We implemented this idea in a model, in which the likelihood of the data comes from a mixture

between a normal distribution centered at zero and a gamma distribution with shape and rate parameters fixed across groups and participants. The normal distribution centered at zero represents responses coming from unchanged representations. The gamma distribution represents responses shifted towards the closest category center. This mixture model was compared to a hierarchical model with a normal likelihood, in which the group means of the normal were allowed to vary across groups. The latter model represents the idea that all representations change to the same degree when the average move toward the closest category center is numerically larger in the experimental group than in the control group. A formal model comparison of these two models favored the mixture model (LOO model weight for the mixture model was .772). The left two panels of Figure 9 show that the mixture model also qualitatively fits the data better by simultaneously accounting for the increased proportion of positive movements and the peak of responses at zero. Figure 10 shows the proportion of responses coming from the gamma component in the experimental group, and in the control group, and the difference between the two. Overall, the proportions were numerically small on the group level (approx. .02 and .01 in the experimental group and control group, respectively) and did not differ across groups (BF = .24 in favor of a group difference), even though a few participants showed a larger proportion of these responses (see right panel of Figure 9).

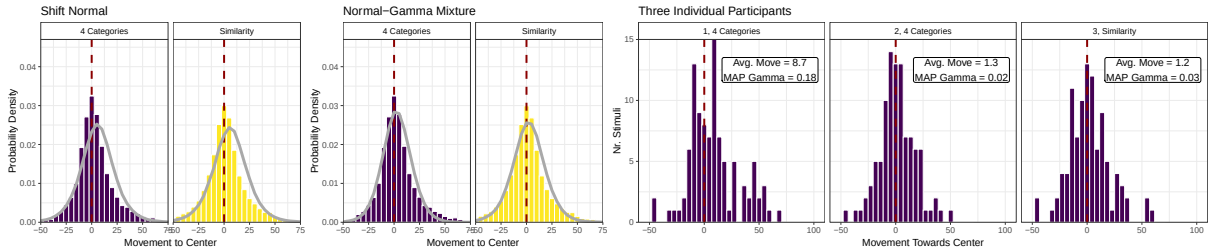


**Figure 9**

*Left and middle: Histograms of movements towards category centers overlaid by posterior predictions of the two models in Experiment 2. Right: Histograms of movements for three exemplary subjects.*

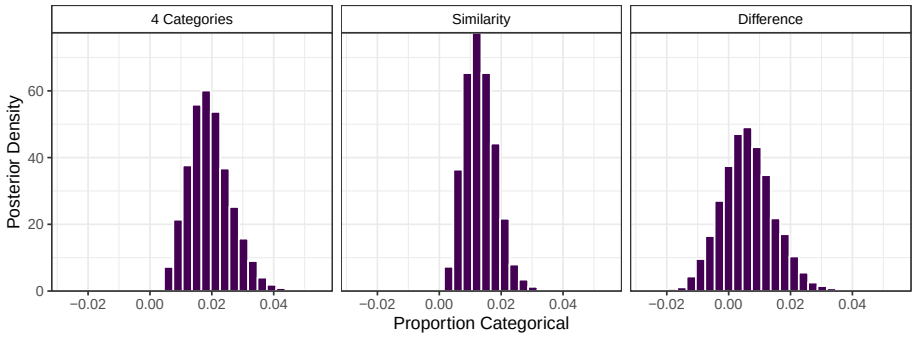


**Figure 10**

*Proportion of responses attributed to gamma component of mixture distribution for category learning group (left), sequential comparison group (middle), and difference of the two groups (right) in Experiment 2.*

#### 4.4.2.1 Combined Analysis

In a combined analysis of Experiment 1 and Experiment 2 we tested the idea that category learning affects perceptual representations of stimuli (e.g., Goldstone, 1994) over and above mere pre-exposure or pre-differentiation (e.g., Gibson and Walk, 1957). For example, Goldstone, 1994 found that participants became more sensitized to dimensions that were categorization relevant. That finding was not restricted to comparisons between categories (i.e., acquired distinctiveness), but was also observed for comparisons within categories, providing evidence against the idea of acquired equivalence. The upshot of the current analysis is that we can contrast any potential performance increases only due to exposure to the stimuli in the sequential comparison task to any potential performance increases due to category learning.

For that reason, we coded responses in the continuous-reproduction task as differences from the true stimulus values on both feature dimensions. The resulting distributions before and after secondary task practice are plotted in the left panel and the right panel of Figure 11, respectively. We modeled the data as coming from a bivariate normal distribution with means of zero. We placed independent hierarchical regression

models with random intercepts on the standard deviations. Pre-exposure to the stimuli predicts negative main effects of time point. If category learning increased perceptual sensitivity over and above pre-exposure, we would expect interactions between time point and group. The results show that standard deviations on both dimensions decreased over time (BF $>$ 100). We additionally observed evidence for the time point x group interaction on the belly dimension (BF = 25.72), but indecisive evidence in the head dimension (BF = .38). The mean of the posterior distributions of the interaction terms in the two dimensions were comparable, though. That analysis is therefore consistent with the idea that learning to categorize stimuli affects perceptual sensitivity over and above exposure to these stimuli alone.

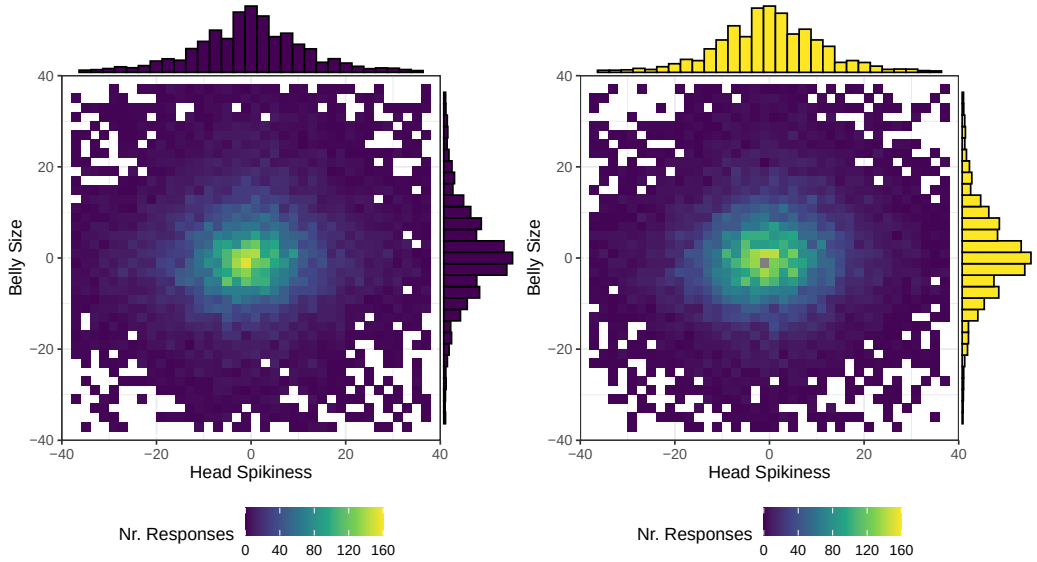


**Figure 11**

*Two-dimensional histograms of the responses in the continuous reproduction task before the secondary task (left) and after the secondary task(right).*

### *4.4.3 Discussion*

The main goal of Experiment 2 was to increase learning in the categorization task to better test our main hypothesis. To achieve that, we made an effort to render category learning easier by replacing the ellipse category structure with the squared category

structure and by additionally motivating participants with a monetary reward. Compared to Experiment 1, categorization accuracy in the last block increased by approx. .05 even though participants had about 200 fewer trials to practice. We conclude that the additional measures were successful. With regards to the main hypothesis of representational change, the pattern was qualitatively in agreement with the predictions on an average level (see right panel of Figure 7). A formal analysis, though, again provided evidence against the predicted effect. In an attempt to quantify representational shifts and to evaluate individual differences therein, we fit a Bayesian mixture model to the data. That allowed us to interpret a parameter indicative of responses moved toward category centers. The analysis showed that both groups only minimally tended to move responses closer to category centers (i.e., the location one quarter and the location three quarters on the two feature dimensions), without any between-group differences. Most people showed none or a very small amount of response shifts towards category centers. Only a few participants, either in the category learning group or in the control group, showed strong movements towards stereotypical response locations (i.e., 25 and 75 on both feature dimensions). As there was no difference across the two groups we interpret these movements as a fatigue effect: In an attempt to simplify the task, some participants discretized the continuous response scale into a discrete one with value 25 for small values on a feature dimension and value 75 for large values on a feature dimension.

A combined analysis of the data from Experiment 1 and Experiment 2 asked the question of whether improved perceptual sensitivity is only an effect of being exposed to a set of stimuli or whether learning to categorize these stimuli additionally increases it. The results supported the latter idea. Even though both groups responded more precisely in the continuous reproduction task after the secondary task than before, the increase in precision was larger for the category learning group than for the control group. As both dimensions were equally important for category learning, an account of dimensional attention (e.g., Nosofsky, 1991) cannot explain the current results. Such an account predicts that

participants pay more attention towards category-relevant dimensions than towards category-irrelevant dimensions. Any differences in perceptual sensitivity could then be explained as a consequence of the prioritized encoding of some dimensions. We can rule out this possibility.

To summarize, in combination with Experiment 1, we interpret the set of results as in disagreement with the hypothesis that visual-perceptual short-term memory representations change according to task-specific practice. In Experiment 3, we pursued the idea that representations change at a more abstract level. For that reason, we changed the continuous reproduction task to a simultaneous comparison task. We hypothesized that besides the feature dimensions of the stimuli, category information may be additionally used to respond to how similar two stimuli are to each other.

### 4.5 Experiment 3

In Experiment 3, we tested the idea that representations change on a higher cognitive level than visual-perceptual representations in short-term memory. The predictions from the computational model can be inspected in Figure 12. The simulated data are the same as for the predictions in Experiments 1 and 2, but they were transformed into pairwise distances between stimuli from the same quadrant and between stimuli from different quadrants (side-by-side quadrants and cross quadrants were collapsed). Differences between model variants were again minor with the main expected pattern being distances between representations from the same category becoming smaller and distances between representations from different categories becoming larger. Whereas representations reflected visual-perceptual representations in Experiments 1 and 2, here, they reflect representations being used for similarity judgments in the simultaneous comparison task.

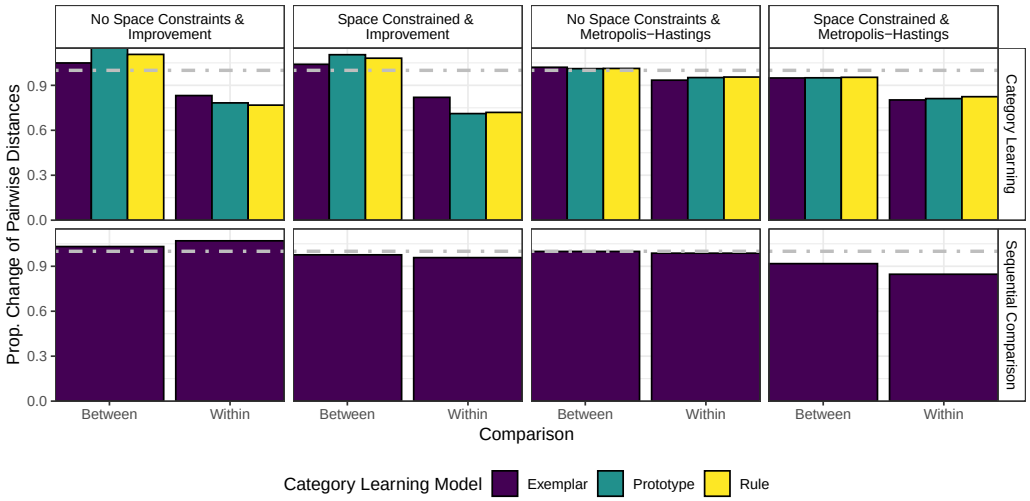


**Figure 12**

*Model predictions for the two secondary tasks (upper row: category learning, lower row: sequential comparison) when transformed to pairwise differences for stimuli from different categories (i.e., between categories) and for stimuli from the same category (i.e., within category). The y axis represents the proportion of the distance after secondary task training compared to the distance before that. That is, values below 1 represent distances to become smaller, and values above 1 represent distances to become larger.*

### *4.5.1 Method*

#### 4.5.1.1 Participants

160 participants (11 unknown, 77 women, 72 men) took part in one session lasting approximately 45 minutes. They received a base payment of 5 GBP and an additional performance-dependent bonus of up to 2.50 GBP. 57 participants did not reach the predefined performance criterion for their secondary task. We further excluded 4 participants because the correlation between their similarity judgments and Euclidean distance was larger than three standard deviations above the mean correlation. We excluded 3 additional participants because they restarted the experiment after they had already progressed through a substantial portion of the experiment. Thus, 96 participants remained in the experiment, 50 in the experimental group, 46 in the control group.

#### 4.5.1.2 Materials and procedure

We changed the task to measure representations from a continuous reproduction task to a pairwise simultaneous comparison task. In every trial, participants observed two monsters presented side by side and were instructed to "rate the monsters by how much you think they look similar to each other" using digits 1-8 (1: not similar at all, 8: very similar) on the keyboard. Stimuli were presented on screen until participants gave a response, which was followed by a message indicating their given response within a 500 ms inter-trial interval. Presentation of the next pair followed immediately. Similar to Experiments 1-2, the two stimuli were presented on black backgrounds on a white screen. We adapted the response scale in the sequential comparison task of the control group to match the scale in the simultaneous comparison task in order to avoid response confusions between the two conceptually similar tasks.

In order to avoid ceiling and floor effects in similarity ratings, we strategically sampled pairs for every participant according to the following procedure: First, we binned Euclidean distances of stimulus pairs into five buckets with thresholds 0, 30, 50, 70, 90, and $\infty$. We then assured that the distances of pairs to be rated for comparisons within and between quadrants of the feature space were on average not too small and too large, respectively. We did so by assuring that pairs sampled for every participant satisfied the following distributions over distance bins: [.5, .5, 0, 0, 0], [.1, .4, .3, .2, 0], and [0, .2, .3, .4, .1] for comparisons within the same quadrant (x4), between quadrants touching side by side (x4) and between quadrants touching only with their corners (x2), respectively. We randomly sampled 10 pairs for every quadrant comparison.

### 4.6 Results

The evidence was again decisive that participants in the experimental group improved in the category learning task over blocks (BF for main effect of block $> 100$). Similarly, those in the control group performed as expected in the sequential similarity task and rated more distant stimuli as more dissimilar and vice versa (BF for main effect of

binned distance > 100).

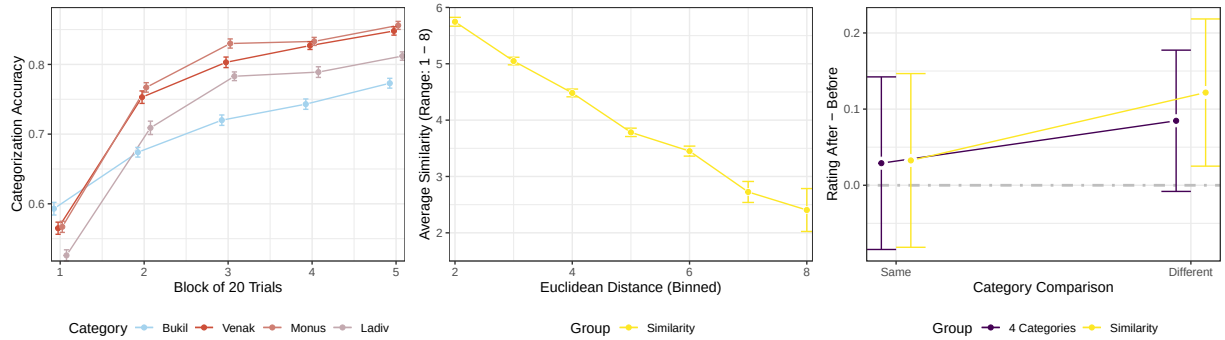


**Figure 13**

*Left: Average categorization accuracy plotted per block of 20 trials separately for the two categories in Experiment 3. Middle: Average similarity rating plotted against euclidean distance of subsequent stimuli in the sequential comparison task. Right: Difference between the similarity ratings given to the same pairs of stimuli after vs. before. Note: Error bars represent 95% within-subjects confidence intervals.*

We hypothesized that participants use category membership as an additional variable to generate similarity judgments. Any effect of category membership should reinforce the predictions of our model presented in Figure 12, which displays the proportionate change of pairwise distances between stimuli after the secondary task compared to pairwise distances before the secondary task. It can be seen that the model predicts an interaction between category comparison (i.e., stimuli from the same category or stimuli from different categories) and group. Pairwise distances of stimuli within the same category should decrease and pairwise distances of stimuli from different categories should increase for the experimental group, but not for the control group. That pattern would be reflected in an interaction between group and category comparison on difference scores. Because people responded on a similarity scale, the direction of the effect should be mirrored. We tested this prediction in a hierarchical model predicting differences in similarity judgments using group and category comparison as predictors. The BF in favor of the group x category comparison interaction was .05, therefore providing strong evidence

for the absence of the effect.

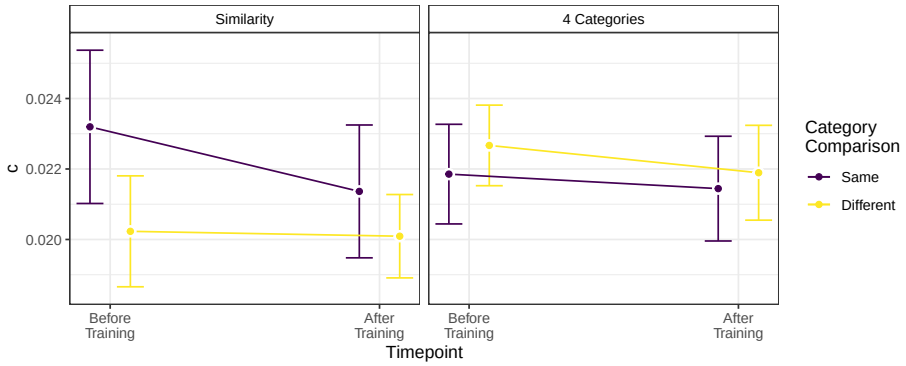


**Figure 14**

*Estimated c parameters for the model relating physical distance to psychological similarity for the two groups in Experiment 3.*

In an attempt to account for individual differences in the mapping of physical distance to similarity ratings, we related physical distance to similarity ratings with a negatively accelerated exponential function. Individual differences can be accounted for by a generalization parameter c. We placed a hierarchical regression model with random by-participant intercepts on the c parameter in a hierarchical Bayesian model using stan. Besides the random intercept, there were main effects for group, category comparison, and time point. We additionally added the three-way interaction of these variables into the model. Representational change would be reflected in an increased c parameter for comparisons between categories (i.e., less generalization) and a decreased c parameter for comparisons within categories (i.e., more generalization) for the experimental group, but not for the control group. This prediction would be reflected in evidence in favor of the three-way interaction. The results (see Figure 14) showed, however, the absence of such an effect (BF = 0.01).

### *4.6.1 Discussion*

In Experiment 3, we replaced the continuous reproduction task, measuring visual-perceptual short-term memory representations, with a simultaneous comparison task

intended to measure higher-level representations. It has been argued that pairwise similarity ratings are affected by context (Hebart et al., 2020). We therefore hypothesized that information about category membership, which arguably was important within the context of the current experiment, may be used as additional information to respond to that task. The results were clear-cut. Even though participants learned to categorize stimuli well and responded as expected in the sequential comparison task, their responses did not change as a function of the secondary task. The conclusion stayed the same when we accounted for individual differences in the mapping from stimulus space to psychological space with a cognitive model.

## 4.7 Experiment 4

Whereas the previous experiments showed evidence against the idea that visual-perceptual short-term memory representations and representations used to rate the similarity of two stimuli change, we modeled the task to measure representations in Experiment 4 as even closer to the category learning task. To achieve that, we asked people to judge how likely two simultaneously presented stimuli belong to the same category. Evidence for representational change in such a task in combination with the previous null findings would suggest that representational change is tightly linked to the practiced task and generalizes only very narrowly and in a domain-specific fashion.

### *4.7.1 Method*

#### 4.7.1.1 Participants

113 participants (5 unknown, 40 women, 68 men) took part in one session lasting approximately 45 minutes. They received a base payment of 5 GBP and an additional performance-dependent bonus of up to 2.50 GBP. 26 participants did not reach the predefined performance criterion for their secondary task. We further excluded 3 participants because the correlation between their similarity judgments and Euclidean distance was larger than three standard deviations above the mean correlation. We

excluded 2 additional participants because they restarted the experiment after they had already progressed through a substantial portion of the experiment. Thus, 82 participants remained in the experiment, 42 in the experimental group, 40 in the control group.

#### 4.7.1.2 Materials and procedure

Materials and procedure were the same as in Experiment 3 with the following exception in the instruction: in the simultaneous comparison task participants were instructed to respond by how likely they thought that the two stimuli belonged to the same category.

### *4.7.2 Results*

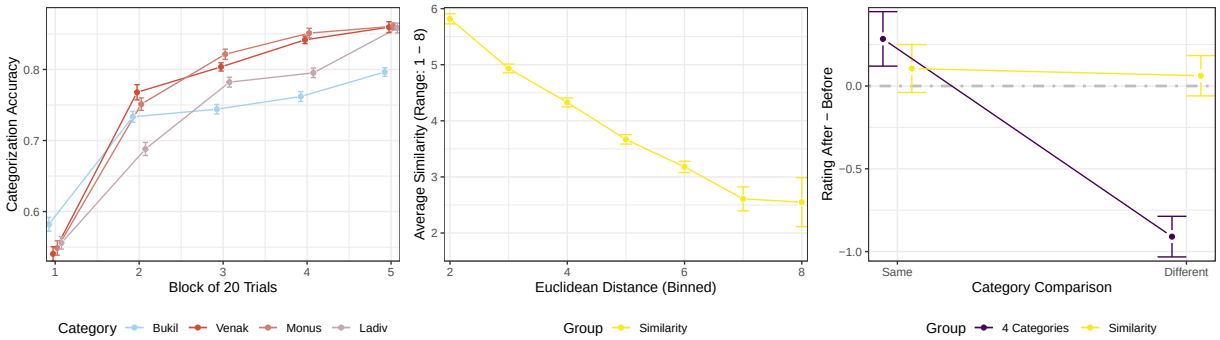


**Figure 15**

*Left: Average categorization accuracy plotted per block of 20 trials separately for the two categories in Experiment 4. Middle: Average similarity rating plotted against Euclidean distance of subsequent stimuli in the sequential comparison task. Right: Difference between the similarity ratings given to the same pairs of stimuli after vs. before. Note: Error bars represent 95% within-subjects confidence intervals.*

As in the three previous experiments, (a) categorization performance improved substantially over blocks (BF $>$ 100) and (b) similarity ratings in the sequential comparison task were a negative linear function of Euclidean distance between subsequent stimuli (BF $>$ 100). Model predictions for the simultaneous comparison task were the same as for Experiment 3; the pattern may even be quantitatively amplified, because participants were explicitly instructed to rate the likelihood of category membership. The

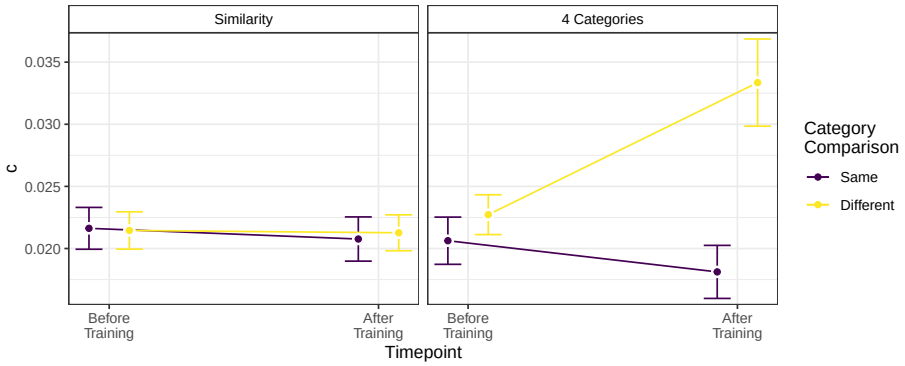


**Figure 16**

*Estimated c parameters for the model relating physical distance to psychological similarity for the two groups in Experiment 4.*

results confirmed the predictions. Similarity ratings for stimuli from the same category slightly increased whereas similarity ratings for stimuli from different categories decreased for the experimental group, but not for the control group (BF $>$ 100 for the interaction between category comparison and group; see right panel in Figure 15). The conclusion remained the same when accounting for individual differences in the mapping between physical space and similarity ratings in the c parameter. Generalization between stimuli was only modulated by the secondary task in the experimental group, but not in the control group. The BF in favor of the three-way interaction effect was decisive (BF $>$ 100).

### *4.7.3 Discussion*

Experiment 4 provided evidence that participants' responses in a task aimed at measuring representations changed according to task-specific practice. Whereas stimuli were perceived as similar across the two groups before the secondary task, only category learning but not sequential comparisons changed their interpretation. After the secondary task, participants in the category learning group rated stimuli from different categories as less likely to come from the same category and tended to rate stimuli from the same category as more likely to come from the same category. Responses in the control group remained unchanged. The conclusion remained the same when we accounted for individual

differences in the mapping from physical space to psychological space with a computational model.

## 5 General Discussion

Research from several areas suggests that object representations are affected by knowledge about the same objects accumulated in a different task. In the current paper, we brought the idea of representational change to a systematic test. We proposed a computational model of representational change based on principles of noisy stimulus encoding akin to population coding (Bays, 2014; Pouget et al., 2000) and preferential storage of helpful information in long-term memory. We tested the model's predictions in a series of four experiments. The main qualitative prediction of the model was that representations of stimuli close to a decision boundary drift away from that boundary. For closed categories (as opposed to the residual category in Experiment 1), the shift should direct towards the respective category center. Only the last experiment showed signatures of representational change. In that experiment the measurement of representations was modeled close to the task mapping in the category learning task. The continuous reproduction task, which we used in Experiments 1 and 2, was closely modeled according to previous research that found influence from categorical knowledge on behavior (Hasantash & Afraz, 2020; Huttenlocher et al., 1991). Despite achieving acceptable levels of category learning in Experiment 2, the results showed evidence against representational change. We assumed that category knowledge would assist as additional information in the pairwise similarity judgments of Experiment 3 because pairwise similarity ratings have been argued to be context-specific (Hebart et al., 2020). However, similarity judgments in Experiment 3 were left completely unaffected by learned category knowledge. To test whether the results depended on the specific method we used to measure representational change, we also analyzed representational change with the representational behavioral similarity analysis (Karagoz et al., 2022). The results, though, showed the same pattern as the analyses reported in the main part of the manuscript (see Appendix).

Our results shed new light on findings from neuroscience that showed that representations as measured via activation patterns in the hippocampus change according to concept learning (Mack et al., 2016; Theves et al., 2019; Theves et al., 2020; Theves et al., 2021). The current findings suggest that it is not the low-level representation of an object that changes, but representations that result after applying a task mapping to the object. In other words, it is the interpretation of objects in a given task context, which changes. Given the current results, we would predict that presenting the same stimulus in different task contexts would elicit different patterns of activation in the hippocampus once the different tasks have been learned sufficiently. That could be tested, for example, by presenting a random pre-cue indicating one of two category structures to be used for an upcoming categorization. The current results also add to the discussion about what mechanisms repetition suppression reflects. In their review, Barron and colleagues listed four mechanisms of repetition suppression proposed in the literature (Barron et al., 2016). All of them emphasize the identity of the stimulus to be crucial for a repetition suppression effect to emerge. Our findings permit an interpretation of repetition suppression effects not purely based on stimulus properties. We propose that repetition suppression measures adaptation to a context-specific representation of a stimulus. The same stimulus may not give rise to a repetition suppression effect if it was used in a different task.

Our results are partially in line with the literature on categorical perception. On the one hand, the combined analysis of Experiment 1 and Experiment 2 showed that category learning sharpened representations over and above mere pre-exposure to the same stimuli, a similar effect already shown previously by Goldstone, 1994. On the other hand, we did not observe any differential effects from category learning on similarity judgments for stimulus pairs between categories and stimulus pairs within categories. Such an effect would be predicted if stimuli from the same category were harder to differentiate (i.e., acquired equivalence) and stimuli from different categories easier to differentiate (i.e., acquired distinctiveness).

Previous studies showed that responses in short-term memory tasks are affected by categorical knowledge (Hasantash & Afraz, 2020; Huttenlocher et al., 1991). Experiments 1 and 2 could not experimentally induce a tendency to rely on categorical representations. The observation that some participants used stereotypical responses in Experiment 2 adds a new flavor to some of the previous studies. For example, responses shifted to stereotypical angles and locations on a circle in Huttenlocher et al. (1991) might also be partially due to stereotypical responding. That is, participants just preferentially responded at a few stereotypical locations from the whole space of possible responses.

More broadly, the observation of only context-specific representational change more broadly aligns with research about cognitive training. An important question in the training literature is whether practice effects in a single task also transfer to different tasks, which would be evidence that a more general cognitive ability is trained. Transfer can be divided up into near transfer to similar tasks and far transfer to less similar tasks. Despite initial findings in favor of far transfer (e.g., Jaeggi et al., 2008), recent work showed that transfer is limited to near transfer when controlling for unspecific training effects as measured using an active control group engaging in a different cognitive task (Melby-Lervåg et al., 2016; Ripp et al., 2022; Sala et al., 2019; Soveri et al., 2017). Context-specific representational change and only limited transfer of cognitive training contrast to some degree with the widespread opinion that humans are excellent at generalization across tasks, a point often made in artificial intelligence research (e.g., Mishra et al., 2022). Given the current results, it seems plausible that what makes us good at generalization across tasks are not representations or basic cognitive abilities that are shared across tasks but general knowledge about the world (Lake et al., 2017).

### 5.1 Conclusion

In the current work, we introduced a computational model of representational change and tested its qualitative predictions in a series of four experiments. Based on these, we can exclude that representational change is a general, context-free phenomenon.

The results suggest much more that representational change describes a re-interpretation of a perceptual representation within a given task context. Only tasks measuring representations very close to that task context will show signatures of representational change.

## 6 Appendix

We also analyzed representational change using the behavioral similarity approach proposed by Karagoz et al., 2022, called behavioral representational similarity analysis (bRSA). Therefore, we calculated the change in distance for every pair of stimuli when comparing responses before the secondary task and responses after the secondary task. In Experiments 1 and 2, we calculated the mean Euclidean distances between all pairs of stimuli before and after the secondary task. In Experiments 3 and 4, we calculated the mean difference in simultaneous comparison judgments for all the rated pairs, also before and after the secondary task. We then calculated the difference for every pair after vs. before. We proceeded in the same way for the model predictions and then correlated differences according to the model with with the observed differences in the four Experiments. Correlations between model matrices and responses should be positive for the experimental group in Experiments 1 and 2 because predictions and data reflect distances, but negative for Experiments 3 and 4 because data are similarity judgments, but predictions are distances.

bRSA Experiment 1

r = .016 (sequential comparison), .017 (category learning)

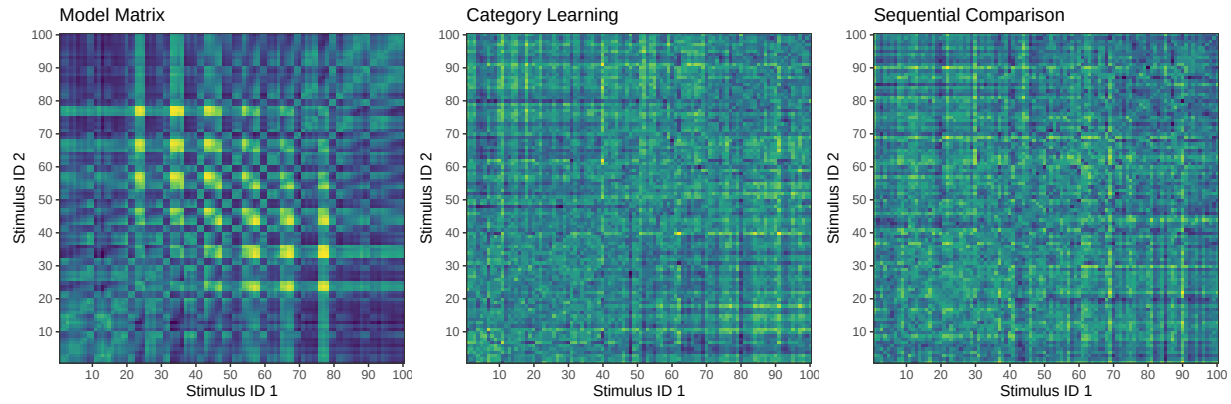


**Figure 17**

*Left: bRSA Model matrix of movements from individual stimulus representations derived from model predictions for Experiment 1. Middle: Average observed movements in category learning group. Right: Average observed movements in sequential comparison group.*

bRSA Experiment 2

r = .013 (sequential comparison), .062 (category learning)

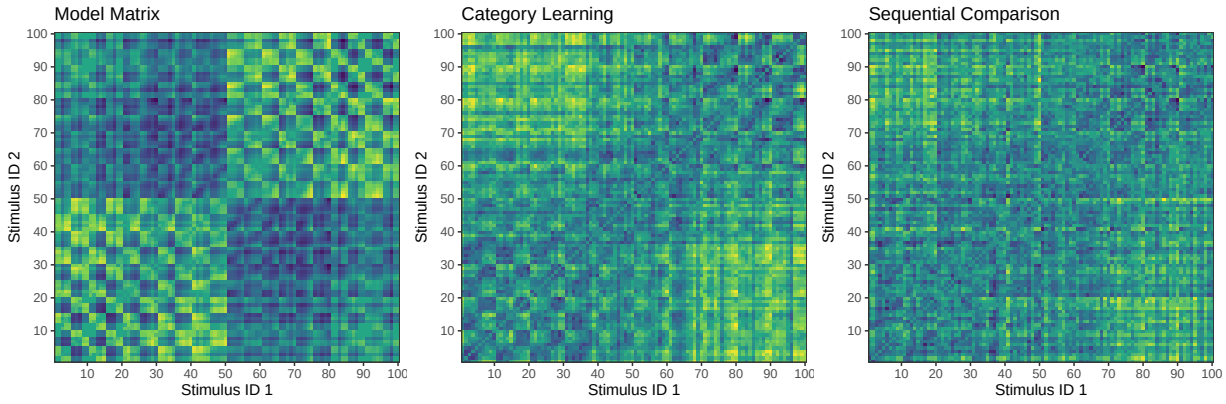


**Figure 18**

*Left: bRSA Model matrix of movements from individual stimulus representations derived from model predictions for Experiment 2. Middle: Average observed movements in category learning group. Right: Average observed movements in sequential comparison group.*

The RSA plots for E3 and E4 show some empty cells, as not all 10k pairs were observed by participants.

bRSA Experiment 3

r = -.030 (sequential comparison), -.013 (category learning)

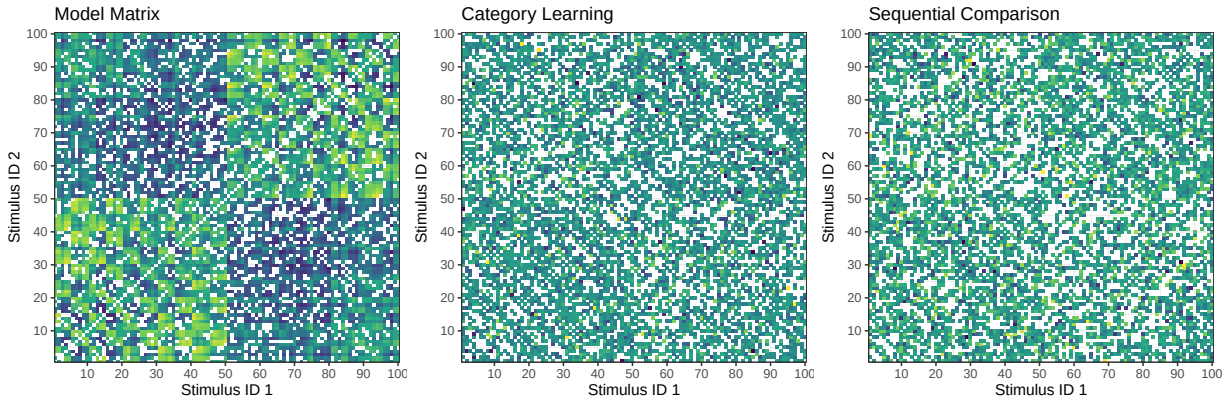


**Figure 19**

*Left: bRSA Model matrix of movements from individual stimulus representations derived from model predictions for Experiment 3. Middle: Average observed movements in category learning group. Right: Average observed movements in sequential comparison group.*

bRSA Experiment 4

r = -.013 (sequential comparison), -.158 (category learning)

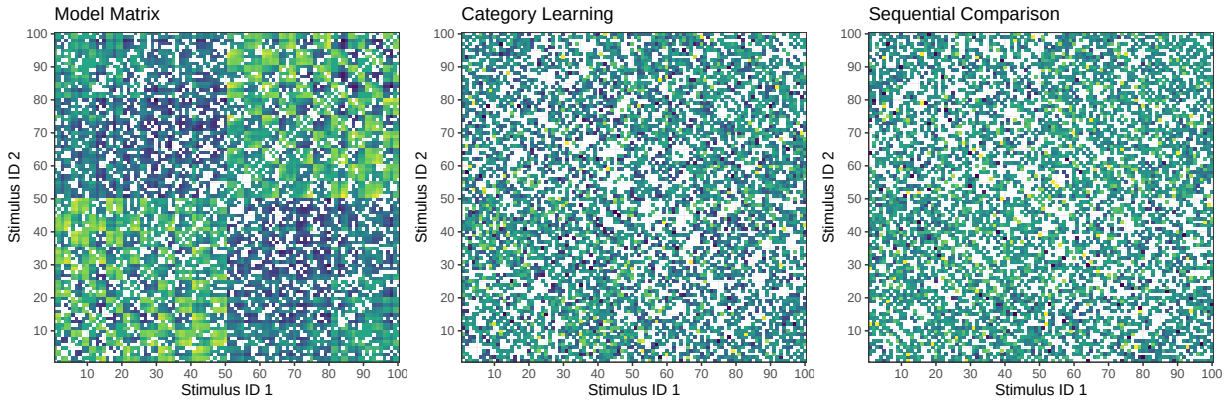


**Figure 20**

*Left: bRSA Model matrix of movements from individual stimulus representations derived from model predictions for Experiment 4. Middle: Average observed movements in category learning group. Right: Average observed movements in sequential comparison group.*

## References

Ashby, F. G., & Gott, R. E. (1989). Decision rules in the perception and categorization of multidimensional stimuli. [Publisher: US: American Psychological Association]. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33. https://doi.org/10.1037/0278-7393.14.1.33

Barron, H. C., Garvert, M. M., & Behrens, T. E. J. (2016). Repetition suppression: A means to index neural representations using BOLD? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1705), 20150355. https://doi.org/10.1098/rstb.2015.0355

Bays, P. M. (2014). Noise in Neural Populations Accounts for Errors in Working Memory. *Journal of Neuroscience*, *34*(10), 3632–3645. https://doi.org/10.1523/JNEUROSCI.3204-13.2014

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives [Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2022). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology*, *33*(3), 386–404. https://doi.org/10.1016/0022-0965(82)90054-6

Donkin, C., Nosofsky, R., Gold, J., & Shiffrin, R. (2015). Verbal labeling, gradual decay, and sudden death in visual short-term memory. *Psychonomic Bulletin & Review*, *22*(1), 170–178. https://doi.org/10.3758/s13423-014-0675-5

Dubova, M., & Goldstone, R. L. (2021). The Influences of Category Learning on Perceptual Reconstructions [eprint:

https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12981]. *Cognitive Science*, *45*(5), e12981. https://doi.org/10.1111/cogs.12981

Gibson, E. J., & Walk, R. D. (1957). The effect of prolonged exposure to visually presented patterns on learning to discriminate them. [Publisher: US: American Psychological Association]. *Journal of Comparative and Physiological Psychology*, *49*(3), 239. https://doi.org/10.1037/h0048274

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. [Publisher: US: American Psychological Association]. *Journal of Experimental Psychology: General*, *123*(2), 178. https://doi.org/10.1037/0096-3445.123.2.178

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27–43. https://doi.org/10.1016/S0010-0277(00)00099-8

Hasantash, M., & Afraz, A. (2020). Richer color vocabulary is associated with better color memory but not color perception. *Proceedings of the National Academy of Sciences*, *117*(49), 31046–31052. https://doi.org/10.1073/pnas.2001946117

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements [Number: 11 Publisher: Nature Publishing Group]. *Nature Human Behaviour*, *4*(11), 1173–1185. https://doi.org/10.1038/s41562-020-00951-3

Homa, D., Sterling, S., & Trepel, L. (1982). Limitations of exemplar-based generalization and the abstraction of categorical information. [Publisher: US: American Psychological Association]. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(6), 418. https://doi.org/10.1037/0278-7393.7.6.418

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and Particulars: Prototype Effects in Estimating Spatial Location. [Publisher: US: American Psychological Association]. *Psychological Review*, *98*(3), 352. https://doi.org/10.1037/0033-295X.98.3.352

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory [Number: 19]. *Proceedings of the National Academy of Sciences*, *105*(19), 6829–6833. https://doi.org/10.1073/pnas.0801268105

John, G. H., & Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifers. *Proceedings of the Eleventh Conference on Uncertainty in Arti*.

Karagoz, A., Reagh, Z., & Kool, W. (2022). *The construction and use of cognitive maps in model-based control* (tech. rep.) [type: article]. PsyArXiv. https://doi.org/10.31234/osf.io/ngqwa

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors [Publisher: Taylor & Francis eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1995.10476572]. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253. https://doi.org/10.1017/S0140525X16001837

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1959). The discrimination of speech sounds within and across phoneme boundaries. [Publisher: US: American Psychological Association]. *Journal of Experimental Psychology*, *54*(5), 358. https://doi.org/10.1037/h0044417

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *113*(46), 13203–13208. https://doi.org/10.1073/pnas.1614048113

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*(1), 49–70. https://doi.org/10.3758/BF03211715

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working Memory Training Does Not Improve Performance on Measures of Intelligence or Other Measures of "Far Transfer": Evidence From a Meta-Analytic Review. *Perspectives on Psychological Science*, *11*(4), 512–534. https://doi.org/10.1177/1745691616635612

Milton, F., & Pothos, E. M. (2011). Category structure and the two learning systems of COVIS [eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-9568.2011.07847.x]. *European Journal of Neuroscience*, *34*(8), 1326–1336. https://doi.org/10.1111/j.1460-9568.2011.07847.x

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. [Publisher: US: American Psychological Association]. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775. https://doi.org/10.1037/0278-7393.27.3.775

Mishra, S., Khashabi, D., Baral, C., & Hajishirzi, H. (2022). *Cross-Task Generalization via Natural Language Crowdsourcing Instructions* (tech. rep. arXiv:2104.08773) [arXiv:2104.08773 [cs] type: article]. arXiv. Retrieved March 1, 2023, from http://arxiv.org/abs/2104.08773
Comment: ACL 2022

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. [Publisher: US: American Psychological Association]. *Journal of Experimental Psychology: General*, *115*(1), 39. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory [Place: US Publisher: American Psychological Association].

*Journal of Experimental Psychology: Human Perception and Performance*, *17*, 3–27. https://doi.org/10.1037/0096-1523.17.1.3

Nosofsky, R. M., Stanton, R. D., & Zaki, S. R. (2005). Procedural interference in perceptual classification: Implicit learning or cognitive complexity? *Memory & Cognition*, *33*(7), 1256–1271. https://doi.org/10.3758/BF03193227

Oberauer, K. (2009). Chapter 2: Design for a Working Memory. In *Psychology of Learning and Motivation* (pp. 45–100). Elsevier. https://doi.org/10.1016/S0079-7421(09)51002-X

Pertzov, Y., Bays, P. M., Joseph, S., & Husain, M. (n.d.). Rapid forgetting prevented by retrospective attention cues. [Publisher: US: American Psychological Association]. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(5), 1224. https://doi.org/10.1037/a0030947

Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes [Bandiera_abtest: a Cg_type: Nature Research Journals Number: 2 Primary_atype: Reviews Publisher: Nature Publishing Group]. *Nature Reviews Neuroscience*, *1*(2), 125–132. https://doi.org/10.1038/35039062

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ripp, I., Emch, M., Wu, Q., Lizarraga, A., Udale, R., von Bastian, C. C., Koch, K., & Yakushev, I. (2022). Adaptive working memory training does not produce transfer effects in cognition and neuroimaging [Number: 1 Publisher: Nature Publishing Group]. *Translational Psychiatry*, *12*(1), 1–13. https://doi.org/10.1038/s41398-022-02272-7

Sala, G., Aksayli, N. D., Tatlidil, K. S., Gondo, Y., & Gobet, F. (2019). Working memory training does not enhance older adults' cognitive skills: A comprehensive meta-analysis. *Intelligence*, *77*, 101386. https://doi.org/10.1016/j.intell.2019.101386

Schäfer, T., Schulz, E., Theves, S., & Doeller, C. (2022). Effects of prototype abstraction on pattern completion and inference in concept space. Retrieved June 21, 2023, from https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_3380402

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *6*(1), 8151. https://doi.org/10.1038/ncomms9151

Souza, A. S., Rerko, L., Lin, H.-Y., & Oberauer, K. (2014). Focused attention improves working memory: Implications for flexible-resource and discrete-capacity models. *Attention, Perception, & Psychophysics*, *76*(7), 2080–2102. https://doi.org/10.3758/s13414-014-0687-2

Soveri, A., Antfolk, J., Karlsson, L., Salo, B., & Laine, M. (2017). Working memory training revisited: A multi-level meta-analysis of n-back training studies. *Psychonomic Bulletin & Review*, *24*(4), 1077–1096. https://doi.org/10.3758/s13423-016-1217-0

Theves, S., Fernandez, G., & Doeller, C. F. (2019). The Hippocampus Encodes Distances in Multidimensional Feature Space. *Current Biology*, *29*(7), 1226–1231.e3. https://doi.org/10.1016/j.cub.2019.02.035

Theves, S., Fernández, G., & Doeller, C. F. (2020). The Hippocampus Maps Concept Space, Not Feature Space. *The Journal of Neuroscience*, *40*(38), 7318–7325. https://doi.org/10.1523/JNEUROSCI.0494-20.2020

Theves, S., Neville, D. A., Fernández, G., & Doeller, C. F. (2021). Learning and Representation of Hierarchical Concepts in Hippocampus and Prefrontal Cortex [Publisher: Society for Neuroscience Section: Research Articles]. *Journal of Neuroscience*, *41*(36), 7675–7686. https://doi.org/10.1523/JNEUROSCI.0657-21.2021

Vanpaemel, W., & Navarro, D. J. (2007). Representational Shifts During Category Learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 7.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values [Number: 5]. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wammes, J., Norman, K. A., & Turk-Browne, N. (2022). Increasing stimulus similarity drives nonmonotonic representational change in hippocampus (M. Barense, T. E. Behrens, & T. Brown, Eds.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *11*, e68344. https://doi.org/10.7554/eLife.68344

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review*, *16*(4), 752–760. https://doi.org/10.3758/PBR.16.4.752