Human Category Learning in Reversal Tasks: Dynamic Learning Rates Help Overcome Catastrophic Interference in Connectionist Networks

Santiago Castiello, Andrew R. Delamater

# Abstract (440 words)

Agents tend to group together physically distinct stimulus exemplars that have never directly been paired together but that share predictive value with respect to some outcome. This “common coding” phenomena was examined with a category learning task. We conducted two experiments, where human participants learned to classify arbitrarily chosen exemplars (different fractal images) into two distinct categories. Each participant completed two tasks that involved reversing the exemplar-category assignments after the initial learning phase. In the Total Reversal task, all exemplar-category assignments were reversed together, while in the Partial reversal task half the exemplar-category relations were reversed with the remaining ones maintained. We predicted that participants would learn Total reversals faster than Partial reversals to the extent that within-category grouping took place. Experiment 1a trained participants with 4 distinct exemplars mapping onto 2 separate categories, while Experiment 1b trained participants with 8 distinct exemplars. In Experiment 2, we evaluated whether the baseline perceptual distinctiveness of the exemplars within- and between-categories differed. We employed Multidimensional Scaling and Euclidean distance measures to analyse perceived similarities among exemplars both within and between categories in a 'psychological' space. Altogether, results showed that participants learned the Total reversal task faster and displayed interference not only with reversed items but also nonreversed items in the Partial reversal task; also, there were no significant differences in the perceived similarities among exemplars used within or between categories, suggesting that the learning effects we observed were not due to different baseline perceptual similarities among our exemplars. To explain the results, we used a standard 3-layered neural network to model category learning with an error backpropagation learning algorithm to adjust the weights between units. In these models, categories are represented by distinct unit activation vectors at the hidden layer, and different within-category exemplars are “grouped together” by activating the same hidden layer category representations. However, these models also suffer from “catastrophic interference” under reversal conditions. To preserve hidden layer category representations to enable more rapid Total than Partial reversal learning an additional mechanism, we observed that another mechanism is required to produce slower weight changes in the input-hidden versus the hidden-output layer. We incorporated the different dynamic learning rate rules inspired by Mackintosh (1975) and Pearce and Hall (1980) from the animal learning literature to implement this additional mechanism, and the modified model faithfully reproduced our empirical results. Overall, these findings underscore the significance of associative mechanisms in category formation and reversal learning, and they demonstrate how catastrophic interference in neural nets can be avoided with different dynamic learning rate adjustments made at different layers of the network as a function of how well the network learns to accurately classify stimuli.

# Introduction (484 words)

The New York Yankees have several skilled hitters. When two different players hit many home runs over the year, we often label those players as 'good hitters' or ‘home run hitters.’ This exemplifies the concept of common coding, where two cues (players) share the same outcome (hitting home runs), and this results in the formation of a common code (home run hitter) by which we relate the different players. During the off-season, for example, general managers looking to obtain a ‘home run hitter’ may then categorize different players interchangeably.

More broadly, common coding is a cognitive phenomenon in which an agent learns that one stimulus generalizes more to another when they have previously been associated with similar outcomes (e.g., Zentall, Sherburne, & Steirn, 1993). Common coding, also known as acquired equivalence, is often recognized as important for category formation, although its underlying mechanisms are not well understood (Delamater, 1998; Goldstone, 1994; Honey & Hall, 1989; Miller & Dollard, 1941). In this paper, we first establish a common coding effect empirically using a category reversal learning task (see also Zentall et al., 1991, 1992) and then offer a mechanistic explanation of category learning using an artificial neural network model of associative learning (Castiello et al., 2021; Delamater, 2012).

Common coding effects have been observed across various settings and species. It has been documented using instrumental conditioning procedures in pigeons (Kaiser et al., 1997; Urcuioli et al., 1995), in Pavlovian conditioning tasks with rats and pigeons (Bonardi & Hall, 1994; Bonardi et al., 1993; Hall, Rey, & Bonardi, 1993; Johns & Williams, 1998), and in matching to sample tasks with humans, even when the paradigm requires common consequences or common antecedents (Bonardi et al., 2005; Delamater & Joseph, 2000). Regardless of species or conditioning type, common coding effects may share similar underlying mechanisms.

An experimental approach to test whether common coding develops between stimuli categorized similarly is to reverse the category assignments either totally or partially following an initial training phase. If participants initially learn to group one set of exemplars in Category A and another set in Category B, then there is an opportunity for distinct common codes to develop among the exemplars within each category. During a Total reversal, all the exemplar-category assignments are switched. If common coding occurs, then learning the Total reversal would simply involve reassigning the internal common codes attached to each exemplar with the updated category response assignment and this should be relatively easy. In contrast, in the Partial reversal condition only half the original exemplar-category assignments are reversed. This requires a complete remapping of the exemplar-category common codes since exemplars from each original category are now treated differently than before. Faster Total than Partial reversal learning has been observed in pigeons, rodents, and humans in related tasks (Delamater & Joseph, 2000; Urcuioli et al., 1995; Zentall et al., 1991, 1992). The rationale for the present experiment is to investigate this Total/Partial reversal effect using a more standard human categorization task than has been reported in the literature. In two experiments we explored these effects by training participants on either a 4-exemplar (Experiment 1a) or 8-exemplar (Experiment 1b) categorization task. In Experiment 2 we tested for any potential differences in pre-experimental perceptual similarities among our exemplar sets both within and between the categories used in Experiments 1a and 1b. We then go on to discuss our data in the context of an artificial neural network approach to modelling category learning.

# Experiment 1a – Easy Categorization Task

Our categorization task consisted of exposing participants on a given training trial to one of a set of exemplars and asking them to learn to categorize it appropriately based on corrective feedback for each trial. A set of fractal images were used as exemplars that participants were to group into two distinct categories. Following an initial training phase long enough for participants to learn the task, the exemplar-category assignments were reversed using either the Partial or Total reversal protocols described above. Each participant performed in a Partial reversal task with one set of exemplars and in a Total reversal task with a second set of exemplars, with the order in which they experienced the two and their identities counterbalanced across participants.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1.** | | | | | | | | | |
| Categorical Reversal Learning Experiment 1a Design | | | | | | | | | |
| **Category Total Reversal** | | | | | **Category Partial Reversal** | | | | |
| Phase 1 | | Phase 2 | |  | Phase 1 | | Phase 2 | |  |
| Learning | | Reversal | |  | Learning | | Reversal | |  |
| Cues | Cat. | Cues | Cat. |  | Cues | Cat. | Cues | Cat. | Condition |
| A | N← | A | →S |  | A | N← | A | →S | reversal |
| B | N← | B | →S |  | B | N← | B | N← | nonreversal |
| C | →S | C | N← |  | C | →S | C | N← | reversal |
| D | →S | D | N← |  | D | →S | D | →S | nonreversal |
| *Note*: A, B, C, and D are cues in the form of fractal images; S← (“southern hemisphere”) and →N (“northern hemisphere”) are two distinct response categories (Cat.). | | | | | | | | | |

**Table 1** presents the experimental design in more detail. Participants were instructed that each fractal image represented the molecular structure of an object in the world, and it was their task to determine if the object came from the Northern (N) or Southern (S) hemisphere. Corrective feedback was given on each trial. Four exemplars (cues A-D) were used in Experiment 1a in which participants first learned to classify two exemplars (A, B) as belonging to the Northern and 2 (C, D) to the Southern hemisphere. During the Total Reversal phase, the two exemplars that previously were classified as belonging to the Northern hemisphere now were correctly classified as belonging to the Southern hemisphere, and vice versa. During the Partial Reversal task, one of the exemplars initially classified as belonging to the Northern hemisphere continued to be classified in that manner with the other one now being classified as belonging to the Southern hemisphere, and vice versa. During the reversal phases we were interested in determining if the Total Reversal condition would result in more rapid reversal learning relative to the reversed exemplars in the Partial Reversal condition, and whether categorization performance among nonreversed exemplars in the Partial Reversal condition might deteriorate as the new categorizations were learned. Both these effects would be expected if within-category exemplars developed a “common code” during the initial classification learning phase.

## Methods

### Participants

Thirty-two participants were recruited through the Prolific platform ([www.prolific.com](http://www.prolific.com)), with the constraints that they were at least 18 years old, performed the experiment on a computer, and resided in either the UK or USA. Their ages ranged between 20 and 63 with a mean of 36.29 and standard deviation (SD) of 10.86. Twenty participants were male, seven female, one non-binary, and four did not respond. The experiment lasted approximately 12 minutes and participants were paid $2.50 for their participation. All procedures posed no risk to participants and were approved by ARD’s CUNY IRB protocol.

We exclude participants that―by combining both tasks―did not show adequate learning in the first 96 trials of the learning phase (before reversal, see below). We define adequate learning when the probability of correct categorical classification, averaged over all phase 1 training trials, exceeded chance performance using a binomial test where p(correct) at chance was 0.5. Thus, participants must have, for both tasks combined, at least 57 correct trials over 96 trials in phase 1. Two participants failed to meet this criterion and were excluded.

### Task and Procedures

The experiment consisted of two tasks: Total and Partial Reversal. Every participant performed in each one with the order of tasks being randomized. Each task, in turn, included two training phases: learning and reversal. The tasks had six 8-trial blocks of training trials in each phase. In each block there were 2 training trials with each one of the 4 fractal images, and the trials were fully randomized. This gave rise to a Total of 6 (blocks) \* 8 (trials/block) \* 2 (phases) = 96 trials for each task.

Participants were given a cover story indicating that fractal images represented the molecular structure of different objects in the world and were asked to classify the four fractals into two categories – objects coming from the “Northern Hemisphere” (N) or “Southern Hemisphere” (S; see instructions in the **Appendix**). In every trial, participants saw one fractal exemplar and selected which category it belonged by pressing the left or the right arrow keys (**Figure 1A**). All fractals were squares, were cantered at the screen, and the size was proportional to the vertical axis of the. The size was ±30% of vertical screen size. Left was linked with the Northern hemisphere (i.e., N←; see **Table 1**) and right with Southern hemisphere (i.e., →S). Between trials there was an Inter-Trial Interval (ITI) with a white fixation cross centered on a grey background. The ITI was a random variable distributed uniformly between .5 and 1.5 seconds rounded to the first decimal place. Immediately following the participant’s choice, feedback was presented for 1000 msec. This took the form of a visual stimulus (i.e., the word “Correct” or “Incorrect”) replacing the display in the center of the screen together with a 200 msec auditory sound (high pitch for correct, low pitch for incorrect). The exact instructions are provided in the **Appendix**. When a participant completed one task they were given the option of starting the second task after a short break.

### Stimuli, Software, and Open Science

The task was programmed in PsychoPy® (Peirce et al., 2019) and the experiment was run online via Pavlovia through Prolific. All the statistical analysis were conducted in R (R Core Team, 2021). The PsychoPy code and R scripts are available in GitHub, along with the analysis scripts: <https://github.com/santiagocdo/categoricalReversalLerning>

All the fractal images used as stimuli in these experiments were kindly provided by the artist Cory Ench (<https://www.enchgallery.com/index.htm>). We chose 2 sets of 4 images to use in Experiment 1a and 2 sets of 8 images to use in Experiment 1b. There were two counterbalancing schemes used in which each unique image was assigned a role, across participants, to a Total and Partial reversal task.

### Statistical Analysis

We used Linear and Logistic Mixed Models instead of classical approaches given the following advantages: *i)* modelling all the trials instead of averages of blocks, *ii)* handling of unbalanced and missing data (we remove trials in which RT < 200 ms and RT > 7000 ms), *iii)* flexibility in modelling correlational structures (no need of sphericity assumption), *iv)* incorporation of continuous or integers covariates (such as blocks), and *v)* modelling multiple sources of random variation (such as within-subjects variation; (Krueger & Tian, 2004; Yu et al., 2022).

To test our hypotheses, we used only the experimental data for the second phase, i.e., post-reversal. We modelled correct classification (correct=1, incorrect=0) with a Logistic Mixed Model (LogMM):

, (*eq. 1*)

Where was a dummy variable encoding Total-reversed with 1 and Partial-reversed with 0, and were used as an integer. The model displayed in *eq. 1* had Blocks as a random slope and participants as a random intercept. To confirm whether the effect is due to the reversal, we ran another LogMM with only data from the first block in the second phase.

, (*eq. 2*)

No random slope was used, and participants were incorporated as random intercept.

To test category interference in the Partial Reversal task, we used a three-step approach. First step, we tested whether the probability of correct for the Partial-nonreversed exemplars was reduced between the first block after and last block. We tested this with the next model:

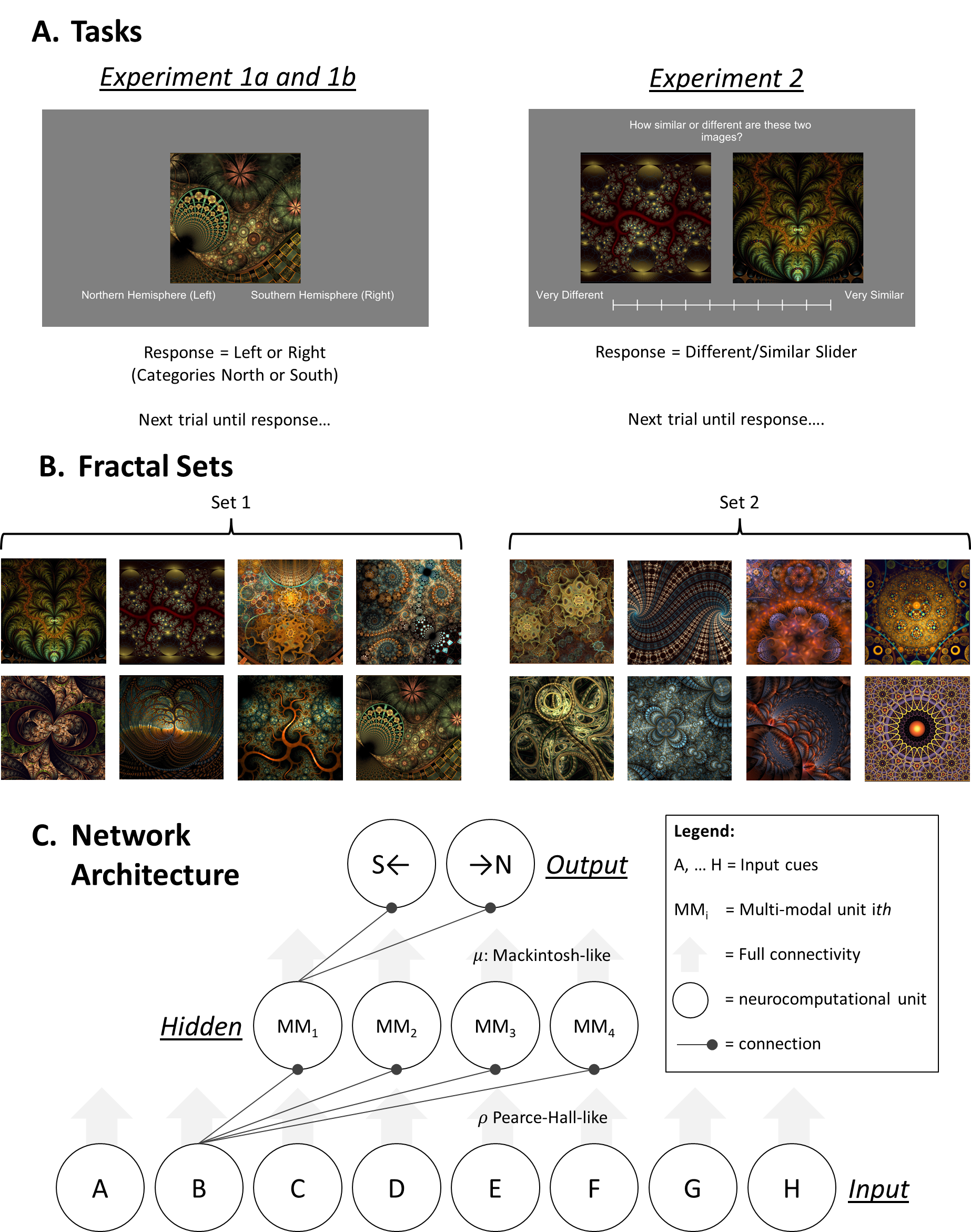
, (*eq. 3*)

The model used participant as random intercept, and is a dummy variable that encodes the first block after reversal as 1 and the last block before reversal as 0. Second step, we used a LogMM similar to *eq. 1*, but only with the Partial reversal conditions (reversed and nonreversed), and with the first two blocks of Phase 2, hence after reversal. We explore the significant interaction by comparing the conditions for each block. Third and final, like the previous step, we ran a LogMM butonly used the Total-revered and the Partial-nonreversed conditions, we explored interactions by testing individual blocks.

Lastly, we also collected Reaction Time (RT) data (measured from the time of exemplar presentation to the classification response) and modelled it with a Linear Mixed Model (LMM), equivalent to *eq. 1*. Although the model residuals from the raw data were not normally distributed, nor with log nor root-square transformations, we decided to run the model with that limitation—we also conducted a sensitivity analysis with the log transformation and root-square, see **Supplementary Information**. The model had the following form:

, (*eq. 4*)

with the same random structure as *eq. 1*. The Type I error, or significance threshold, alpha, for the models’ parameters tests was set to .05. All models’ coefficients were presented with their standardized coefficient (Std. Coef.) ―an equivalent to effect sizes (Makowski et al., 2021)―and their corresponding 95% confidence intervals.



**Figure 1**. ***A****. present one trial for both tasks used in Experiments 1a and 1b, and 2.* ***B****. Sets of fractals used for each of the two tasks.* ***C****. Architecture used with all models: BP (Castiello et al., 2021; Delamater, 2012). 𝜌 controls learning rate changes from Input-Hidden, and 𝜇 controls learning rate changes between Hidden-Outputs. The Hidden layer always consisted of four multimodal units (i.e., fully connected).The fractal stimuli were kindly provided and authorized by artist Cory Ench* (<https://www.enchgallery.com/index.htm>)*.*

## Results

### Superior Learning After the Total Reversal

The average probabilities of correct classification for all blocks in both tasks are presented in **Figure 2A**. Blue circles represent average Total Reversal task performance across all exemplars from both categories for each training phase. The Partial task in yellow has two conditions during the reversal phase: reversed (square) and nonreversed (triangle) trials. Although these distinctions did not apply during phase 1 training, that data are segregated in this manner as well. The critical comparisons assess phase 2 performance on the reversed exemplars between the Total and the Partial task (**Figure 2C**), and phase 2 performance on nonreversed trials in the Partial task relative to reversed trials in the Partial and Total tasks.

The LogMM revealed a positive main effect for Total vs Partial [], suggesting that in phase 2, participants had higher correct scores in the Total than Partial Reversal task. There is also a positive main effect of Blocks [], which suggest that participants improved across reversal trials. However, the interaction between task and Blocks was not significant []. Just to corroborate that Partial reversal impaired reacquisition of the categories, we tested the first block between Total vs Partial, and we found a significant and large effect [].

### Within-Category Interference in the Partial Reversal task

To the extent that category learning involves a common coding mechanism, we expect that in the Partial Reversal task learning a new category assignment on reversed trials would interfere with performance on nonreversed trials. Nonreversed exemplars suffered a reduction in performance even though their category assignments remained unchanged (**Figure 2A**). Nonreversed exemplar performance declined on block 7 relative to block 6, i.e., comparing the block just prior to and after reversal []. In addition, we might expect that as reversed exemplar performance improves this could result in further performance reductions on nonreversed trials. The largest improvements on reversed exemplars during the reversal phase came from blocks 7 to 8. However, as performance on those trials improved across these blocks, performance on nonreversed trials did not get worse, but also improved. We compared the Partial reversed versus Partial nonreversed conditions, and although we found a significant interaction with blocks (7 and 8) and conditions [] this simply indicates that reversed trials improved at a faster rate. Finally, if reversal learning in the Partial task interferes with performance on nonreversed trials, this could be reflected by poorer performance on nonreversed trials relative to performance in the Total Reversal condition. To test this idea we compared performance in the Partial nonreversed and Total reversed conditions in blocks 7 and 8 and found a significant interaction []. However, this interaction indicates that while accuracy for Partial nonreversed exemplars was superior to Total reversed exemplars in block 7, [] Total Reversal performance did not significantly exceed Partial nonreversed performance in block 8 []. Overall, this pattern of data provides some evidence in support of within-category interference.

### No Differences in Reaction Time After Reversal

We also tested the RT for the second phase with an LMM. We found that RTs did not differ between tasks (Total versus Partial) [], nor was there an effect of Blocks [], or an interaction []. We also ran the same model but after transforming the RT data with the natural logarithm and the root-squared. The results were similar (see **Supplementary Information**).

### No Differences in Phase 1 Acquisition

As a control measure, we corroborated that learning in Phase 1 did not differ between the two tasks only testing the items that would be reversed in Phase 2. We fit a LogMM with only data from Phase 1. We found a positive significant effect of Blocks [] but no main effect of Task [] or an interaction []. This is evidence that participants similarly learned both tasks and conditions in Phase 1. Thus, differences in Phase 2 are only explained by the tasks and no different learning trajectories in the previous phase.

A group of graphs showing different types of data

AI-generated content may be incorrect.

**Figure 2**. *Left panels Experiment 1a, right panels Experiment 1b.* ***A*** *and* ***D*** *probability of correct as a function of blocks.* ***B*** *and* ***E*** *mean of median reaction times as a function of blocks.* ***C*** *and* ***F*** *same as* ***A*** *and* ***D*** *but only plotted for the reversal part of the task, after the vertical line between bocks 6th (easy task) and 8th (hard task). We also removed the partial-nonreversed (yellow triangle) conditions. Error bars are the within participants standard errors of the means (SEM) using the Cousineau–Morey method* (Cousineau & O’Brien, 2014)*.*

# Experiment 1b – Hard Categorization Task

Experiment 1b was run in an identical manner to Experiment 1a except we required participants to categorize 8 exemplars into two categories of 4 exemplars each. This 8-exemplar task made it more difficult for participants to learn, and so the initial training phase was extended to 8 training blocks and since most of the effects in Experiment 1a appeared during the early reversal blocks we decreased the number of reversal blocks to 4.

## Methods

### Participants

Thirty-two additional participants were recruited through Prolific ([www.prolific.com](http://www.prolific.com)). The age ranged between 22 and 65 with a mean of 40.16 (SD 13.80). Eleven participants were male, 15 female, one non-binary, and 5 did not reply. The present experiment lasted approximately 20 minutes and participants were paid $4.00 for their participation. All procedures posed no risk to participants and were approved by ARD’s CUNY IRB protocol.

We excluded participants using the same criterion as in Experiment 1a. Thus, participants’ inclusion was based on a binomial test in which the probability of correct phase 1 performance for both tasks had to be higher than random accuracy (.5). This implies at least 142 correct trials over 256 trials in phase 1 (128 trials per task). Six participants were excluded for not reaching this criterion. We rerun all the analysis including those participants and the results were similar, supporting the same conclusions.

### Task and Procedures

The task structure was the same as in Experiment 1a. The only difference in this experiment was the task difficulty manipulated by adding more stimuli (fractals; **Table 2**). Also, the number of blocks per phase varied, where Phase 1 included 8 training blocks with 4 reversal blocks in Phase 2. Furthermore, each block had 16 trials (2 presentations of each exemplar). Within each block the trials were fully randomized for a Total of 8 (phase1: blocks) \* 16 (trials) + 4 (phase 2: blocks) \* 16 (trials) = 192 trials for each of the two tasks (Total and Partial reversal).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 2.** | | | | | | | | | |
| Categorical Reversal Learning Experimental 1b Design | | | | | | | | | |
| **Category Total Reversal** | | | | | **Category Partial Reversal** | | | | |
| Phase 1 | | Phase 2 | |  | Phase 1 | | Phase 2 | |  |
| Learning | | Reversal | |  | Learning | | Reversal | |  |
| Cues | Cat. | Cues | Cat. |  | Cues | Cat. | Cues | Cat. | Condition |
| A | S← | A | →N |  | A | S← | A | →N | reversal |
| B | S← | B | →N |  | B | S← | B | →N | reversal |
| C | S← | C | →N |  | C | S← | C | S← | nonreversal |
| D | S← | D | →N |  | D | S← | D | S← | nonreversal |
| E | →N | E | S← |  | E | →N | E | S← | reversal |
| F | →N | F | S← |  | F | →N | F | S← | reversal |
| G | →N | G | S← |  | G | →N | G | →N | nonreversal |
| H | →N | H | S← |  | H | →N | H | →N | nonreversal |
| *Note*: A, B, C, D, E, F, G, and H are cues in form of fractals; S← (“southern hemisphere”) and →N (“northern hemisphere”) are two distinct correct response categories (Cat.). | | | | | | | | | |

### Analysis

The same analysis plan was used as in Experiment 1a.

## Results

### Faster and Better Learning After the Total Reversal

**Figure 2D** shows the average p(correct) category classifications for each of the training and reversal blocks in the Total and Partial reversal tasks. The Total reversal task is again represented by one line (blue circles) collapsing across the two categories, but the Partial task in yellow is represented separately for reversed exemplars (square) and nonreversed exemplars (triangle). The critical comparisons involve comparing Phase 2 data for reversed exemplars in the Total and Partial tasks (**Figure 2F**) as well as nonreversed trials in the Partial task relative to reversed trials in the Total task.

We fit a LogMM to the reversed exemplars in Phase 2 (**Figure 2F**), and the model revealed two significant main effects, Total vs Partial [] and Blocks []. The former suggests that, in Phase 2, participants were more accurate in the Total than in the Partial task; the latter effect suggests that overall participants improved over the reversal phase. In contrast to the previous experiment, the interaction between task and Blocks was significant [], this suggests that the rate of readjusting to the new phase (i.e., the slopes difference between reversed tasks as a function of Blocks) was different. We also tested the first block between Total vs Partial, and we found a significant and large effect [].This suggests that Total reversal was acquired quickly than Partial reversal.

### Within-Category Interference in the Partial Reversal task

In this experiment, we found stronger evidence supporting within-category interference in the Partial Reversal condition (**Figure 2D**). First, we compared the first block of Phase 2 against the last block of Phase 1 (i.e., block 9 versus block 8) in nonreversed exemplars of the Partial task and found a significant reduction in accuracy []. Second, we also compared Partial reversed and Partial nonreversed exemplars for the first two blocks of Phase 2 (reversal training). This analysis revealed a significant interaction [] supporting the visual impression that whereas performance improved on reversed trials (with a positive slope) across blocks 9 and 10 performance on nonreversed trials failed to improve across these blocks (and even appeared to get worse). Higher accuracy was found in the nonreversed condition in block 9 [], but not in block 10 []. Finally, to test the degree to which within-category interference affected performance to Partial nonreversed exemplars we compared them against the Total reversal condition, also in those first two reversal blocks. The interaction was also significant []. Interestingly, we observed no differences between these conditons in block 9 [], but Total Reversal accuracy was superior to Partial Nonreversal accuracy on block 10 []. Altogether this pattern of results suggest substantial within-category interference in this experiment pointing to the presence of an underlying common coding process in category learning.

### Larger Reaction Time in the Total Reversal

In contrast to the previous experiment, we found RT differences between tasks during the reversal phase. Overall the RTs in the Total task were higher than on reversed trials of the Partial task [], but there was no main effect of Blocks [], and no significant interaction was found []. Thus, overall, in Phase 2 participants were slower to respond in the Total reversal task. The results were similar when we used the log-transformed and root-squared RT (see **Supplementary Information**).

### No Differences in Phase 1 Acquisition

As in the previous experiment, it is important to test that the learning trajectories in Phase 1 were not different between the two tasks. We ran a similar model with Blocks, Task, and its interaction. We only found a significant effect of Blocks suggesting adequate learning in Phase 1 [], but no difference between Total and Partial reversed exemplars as main effect [] or an interaction []. Thus, similar learning between reversed exemplars was observed in Phase 1

# Experiment 2 – Pre-experimental Exemplar Similarities

In Experiments 1a and 1b we assigned fractal images to the two categories using two sets of exemplars and made sure that across the Total and Partial reversal conditions the same physical set of exemplars underwent reversal training (or nonreversal training in the Partial task) across participants. However, because there was no obvious way to categorize the various fractal images based on any particular set of perceptual features (e.g., animate/inanimate, spiky/stubby, red/green, circular/triangular shapes, etc), we did not attempt to randomize our assignments. Instead, we chose assignments that appeared to us to cut across any perceptual dimension we could think of. Therefore, we thought it would be important to collect normative data on an independent set of participants to test whether our fractal exemplar-to-category assignments resulted in different inherent pre-experimental similarities within vs between the two categories. We suspected there would be none, but, nevertheless, thought it important to establish that fact.

To accomplish this goal, we asked new participants to judge the similarities between every pair of images used in Experiments 1a and 1b. We then subjected this similarity rating data to multidimensional scaling procedures and obtained an overall distance measure within multidimensional space among those exemplars that appeared within and between categories. As will be seen below, our intuitions about the stimuli were supported by this multidimensional analysis.

## Methods

### Participants

Sixteen additional participants were recruited through Prolific ([www.prolific.com](http://www.prolific.com)). The age ranged between 28 and 61 with a mean of 42.69 (SD 9.91). Six participants were male and 7 were female. Again, all procedures were in accordance with ARD’s IRB protocol. The task took approximately 10 min and participants were paid $2.50 for their participation.

### Task and Procedures

Participants rated each pair of exemplars for each set of fractals (**Table 3**) on a 10-point sliding scale varying from 'very similar' to 'very different'. Each trial consisted of a fixation cross followed by the rating screen. The fixation duration, serving as the inter-trial interval (ITI), was sampled from a uniform distribution ranging between 0.5 and 1.5 seconds. After the ITI, participants saw the rating screen (see **Figure 1A**), which remained until they provided their similarity or difference rating. The slider appeared 1 second after participants viewed the two exemplars. This delay ensured that participants had adequate visual exposure to both exemplars before responding. They chose anywhere on the line (i.e., on or in between markings) and the computer registered their rating numerically as a real number between 1 and 10.

In Experiments 1a and 1b, participants completed two tasks, each involving distinct sets of four (E1a: A–D, E–H, I–L, or M–P) or eight (E1b: A–H, or I–P) exemplars. In this experiment participants rated the similarity of all possible comparisons per set of exemplars. Thus, for each set, there were 64 possible pairings (8 left positions × 8 right positions) minus the 8 self-comparisons. This resulted in 56 unique comparisons per set of stimuli (**Table 3**). This number of trials enabled each comparison to be made twice, with left and right positions counterbalanced. Consequently, participants were expected to rate a Total of 112 exemplar pairs for each set of stimuli (56 trials per task \* 2 sets of stimuli). However, due to a programming error, the final comparison for each task (e.g., left H versus right G) was omitted.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 3.** | | | | | | | | | |
| Categories Perceptual Similarity: exemplar comparisons | | | | | | | | | |
|  |  | Right | | | | | | | |
|  |  | A/I | B/J | C/K | D/L | E/M | F/N | G/O | H/P |
| Left | A/I |  | **x** | **x** | **x** | *x* | *x* | *x* | *x* |
| B/J | **x** |  | **x** | **x** | *x* | *x* | *x* | *x* |
| C/K | **x** | **x** |  | **x** | *x* | *x* | *x* | *x* |
| D/L | **x** | **x** | **x** |  | *x* | *x* | *x* | *x* |
| E/M | *x* | *x* | *x* | *x* |  | **x** | **x** | **x** |
| F/N | *x* | *x* | *x* | *x* | **x** |  | **x** | **x** |
| G/O | *x* | *x* | *x* | *x* | **x** | **x** |  | **x** |
| H/P | *x* | *x* | *x* | *x* | **x** | **x** |  |  |
| *Note:* 56 fractals similarity trials and ratings. 8 x 8 = 64, but 64 - 8 = 56, given the 8 stimuli diagonal, where it does not make sense to compare the same stimulus against itself. However, the comparison Left H and Right G was not tested due to a programming mistake. “/” separate letters used for each set: A–H and I–P. **Bold** xs represent the within category comparisons, and *italics* xs represent the between category comparisons. | | | | | | | | | |

### Statistical Analysis

The intuition behind the analysis is that the perceived similarity between exemplars within one category (e.g.*,* S←) should not differ from the perceived similarity within the other category (e.g., →N). Similarly, perceived similarity within categories should not differ from the perceived similarity between categories. The perceived similarity was obtained by the distance between each pair of exemplars in in a 'psychological' space (described below).

We first averaged the similarity ratings for each exemplar comparison. For example, we combined the similarity ratings for Left A vs. Right B with those for Left B vs. Right A. For each participant and each task, we used the 56 averaged similarity ratings between pairs of exemplars and implemented Multidimensional Scaling (MDS) with 2 dimensions (i.e., *k=2*). We chose two dimensions because prior research suggests that visual perception can be meaningfully reduced to this number of dimensions (Bracci & De Beeck, 2023). This provided us with 2 coordinates in “psychological” space for each exemplar. This procedure was repeated for each task and each participant. We then used the Euclidean distance to compare each exemplar against the others in the psychological space. For each participant in each task, this yielded 28 distances, resulting from the combination of 2 elements in a set of 8, per task. We conducted a sensitivity analysis with three dimensions (MDS with *k*=3) and the results were comparable (see **Supplementary Information**).

For each set of exemplars, we conducted two relevant comparisons: (1) a mean within-category distance measure was determined for each participant for each of the two categories; comparing exemplars A to D versus E to H and I to L versus M to P; and (2) mean between-category versus within-category distance measures were determined for each participant (e.g., by comparing the mean distances between any given exemplar and the others within its category verses its distances to other category exemplars). As the distance distributions did not pass the Kolmogorov-Smirnov test for normality, we used the Wilcoxon rank-sum test with continuity correction for the comparisons. The Type I error rate, α (alpha), was set to .05.

## Results & Discussion

The comparisons of distances are shown in **Figure 3**. We first tested whether there was a difference within each category (S← vs →N) for the set 1 (exemplars A to H). This comparison involved only the distances within each category between A to D (A vs B, A vs C, …, C vs D) against E to H (E vs F, E vs G, …, G vs H). We did not find any differences between the within distances in A to D and E to H, (**Figure 3A**). Then, we conducted the same test for the set 2, comparing the two categories, I to L versus M to P. We did not find differences, (**Figure 3B**).

Next, in order to reject the possibility that our choice of exemplar-category assignments might have been biased pre-experimentally—A to D and E to H for set 1, and I to L and M to P for set 2—we compared the mean distances within the categories against the mean distances between the categories we used in our studies (see the bold and italics note for **Table 3**). To exemplify this, we take as reference category A to D for set 1. Here, a distance between A and B is considered within-category, but a distance between A and E is considered between-category because A belongs to the A to D category but E to the E to H category. We found no differences between within-category distances and between-category distances for set 1, (**Figure 3C**), and for set 2, (**Figure 3D**).

These results suggest that participants did not perceive one set of category exemplars as more alike than the other. This provides evidence that the effects observed in Experiment 1 are not due to perceptual resemblance of the exemplars but due to the underlying learning of those categories.

A diagram of a distance in perceptual space

AI-generated content may be incorrect.

**Figure 3**. *Category Similarity Analysis, using two psychological dimensions (multidimensional scaling with k=2) in the psychological space.* *Left panels (****A*** *and* ***C****) represent Experiment 1a results for the exemplars for Set 1, and the right panels (****B*** *and* ***D****) represent the comparisons conducted with exemplars in Set 2. Each dark point represents a comparison for each participant, i.e., for each of the 28 stimulus comparisons. Boxplots represent the median, interquartile (IQR) range, and the whiskers are 1.5 the IQR. \*\*\*: p < .001, \*\*: p < .01, \*: p < .05*

# General Discussion

The results of our experiments demonstrate that category learning results in a grouping process whereby the cognitive machinery enhances the functional similarity among exemplars within a category and differentiates them from exemplars assigned to different categories. This effect is reminiscent of so-called acquired equivalence and distinctiveness of cue effects, common coding effects, and equivalence class effects investigated in the animal learning literature for some years (Delamater, 1998; Delamater & Joseph, 2000; Honey & Hall, 1989; Johns & Williams, 1998; Sidman, 2000; Zentall et al., 1991, 1992). We provided evidence for such an effect using a Total/Partial reversal learning paradigm in which reversed exemplar-category assignments were not only learned more rapidly in a Total than Partial reversal task. Apparently, once the category task is learned, new learning to any one exemplar in the category can generalize more effectively to other within-category exemplars. This would result in both an enhancement of category reversal learning in the Total reversal task, as well as impaired performance on reversed and nonreversed exemplars in the Partial reversal task until new category structures emerge. In addition, our results in Experiment 2 suggest that there was very little basis for the two arbitrarily chosen sets of exemplars to display these effects other than their functional relationships defining their category assignments (i.e., objects belonging to the “northern” or “southern hemisphere”).

Our findings also are consistent with other results that have used a transfer procedure. Once two sets of exemplars have been categorized appropriately, researchers have shown that learning new responses to a subset of these exemplars transfer more readily to other exemplars within than between categories (Delamater & Joseph, 2000; Johns & Williams, 1998; Urcuioli et al., 1995). Both Total/Partial reversal and transfer effects, then, can be taken to support the basic finding of greater functional similarity among exemplars within and reduced similarity among exemplars between categories.

These conclusions are supported by our accuracy data in both Experiments 1a and 1b. However, the RT data where not as clear. We only found a difference in Experiment 1b, where people were slower to respond in the Total reversal condition than in the Partial condition. Perhaps this was not observed in Experiment 1a because that task was easier, and it has been shown that cognitive load modulates reaction times (Nilsson et al., 2018). It is not clear what accounts for slower RTs in our Total reversal task. On *a priori* grounds, one could have expected faster RTs in this condition. In any case, while the Experiment 1b data could be described in terms of a speed-accuracy trade-off across Partial and Total reversal conditions, the Experiment 1a data cannot be described in this way. Apparently, slower RTs in the difficult Total reversal condition reflects greater cognitive processing prior to response execution, but the nature of that processing remains to be determined.

Another finding we observed was a category interference in the nonreversal exemplars. In the Partial Reversal task, when only half of exemplars' category assignments were reversed, the accuracy for nonreversed exemplars worsened. We interpret this as evidence indicating that exemplars within a category were commonly coded. If people would have learn unique exemplars associations in contrast to a commonly code association, there would be no reason for an impairment in performance for the nonreversed exemplars. Nevertheless, the performance for nonreversed exemplars decreased just after the reversal, in Experiment 1b this interference lasted for two blocks post-reversal. Where participants had difficulty to reach their baseline performance for two blocks and even in the second block after performance their performance was worse than the completely new learnt category learnt from the Total Reversal task.

Others have commented on the possibility that associative processes might play a role in category learning (Delamater, 1998; Goldstone, 1994; Honey & Hall, 1989; Miller & Dollard, 1941; Turner & Wasserman, 2023; Wasserman, Brooks, & McMurray, 2015). However, our results demonstrate that any such effort would require an associative framework that goes beyond the formation of simple binary exemplar-category associations. For example, if 8 distinct exemplar-category associations had been established, then there is no direct way in which Total reversals should be learned more rapidly than Partial reversals, for the simple reason that separate binary associations should not generalize any more to exemplars within versus between categories (when the inherent similarities among them are equated as our scaling data suggest they were). What is needed is an associative framework that provides a mechanism whereby greater within- than between-category generalization can occur.

## Simulation

We explored this idea by considering connectionist models that include a “hidden layer” in between an input layer that codes for individual exemplars and an output layer that codes for the different category responses used in the task (Castiello et al., 2021; Delamater, 2012). Within such a network, learning consists of weight changes between the input-hidden and hidden-output layers in such a way that exemplars come to activate the appropriate output layer category response. Importantly, such networks accomplish this by forming strong connections among all the exemplars within a category and the same set of hidden units that, in turn, come to form strong connections with the appropriate category output response unit. This hidden layer convergence within-category and between-category divergence is exactly the type of mechanism that would result in “common coding” amongst within-category exemplars and “distinctive coding” amongst between-category exemplars. This can readily be seen by considering the input-hidden (or first layer, k=1) and hidden-output (or last layer, k=2) weight matrices (see **Figure 4**) that commonly emerges in a simple 8-4-2 connectionist network (see **Figure 1C)**.

The weights matrices in **Figure 4C** and **4D** illustrates a very typical solution to the category learning problem when 8 distinct exemplars are categorized into two categories. In essence, through the application of a standard backpropagation learning algorithm (for details see Castiello et al., 2021; Delamater, 2012), each within-category exemplar forms strong excitatory connections to hidden layer units that themselves form strong connections with one of the category output units. In addition, exemplars within each category develop inhibitory connections to hidden units that strongly activate the other category output unit. In essence, all within-category exemplars are “commonly coded” at the hidden layer which develops category representations that come to activate the appropriate output category response.

This sort of connectionist modelling approach has the potential to reveal the sorts of empirical results we report here. If within-category items are commonly coded by activating a similar set of hidden layer units, then learning a Total reversal should be relatively easy because it merely would involve altering connections between the “category representations” at the hidden layer and the appropriate output responses. In contrast, a Partial reversal task can only be solved once the hidden layer category representations have been realigned, since they no longer apply during the reversal phase. That deconstruction of the internal task representation should, intuitively, be more difficult to achieve.

To our surprise, this was not the case. Our simulations with standard backpropagation ―Model 1 (Delamater, 2012)― revealed more rapid Partial than Total reversal learning (**Figure 4A**). Evaluation of the weight matrices after initial learning and again after reversals reveals why the network fails. In both Partial and Total reversal tasks, the category representations at the hidden layer were remapped during reversal training by updating the input-hidden weights, whereas the hidden-output weight matrices were little changed (see the matrices for the Total reversal task, **Figure 4C**, and the Partial task, **Figure 4D**). Given that remapping the categories in the Partial task requires fewer weight matrix changes (for only half the exemplars) learning this task is quicker for the network (**Figure 4A**). However, for the Total task, this is not efficient because the network has already learned to represent the categories at the hidden layer after the initial training phase (**Figure 4C**). So, there should be no need to remap the category representations in the first layer, but instead only change the output matrix layer preserving the already formed common coding. This does not occur in Model 1 because there is an output error, initially, on every single Total reversal trial. This produces “catastrophic interference” throughout the network because the backpropagation of error occurs throughout the entire network on every trial. That produces rapid changes in the input-hidden weights, and, thus, a destruction of the initially learned category representations. To achieve more stability in the hidden layer category representations and overcome this widespread catastrophic interference effect, we implemented an additional mechanism within the network –dynamic learning rates at each of the two layers.

One reason for faster remapping categories in the first layer (k=1) instead of the second (k=2) is that learning rates are the same for every layer within the network in Model 1. Moreover, the standard backpropagation learning rule implements weight changes as a function of unit activation. Since input units are active, by definition, at either 1 (stimulus present) or 0 (stimulus absent), this will ensure relatively quicker input-hidden than hidden-output weight changes on each reversal trial. Recall that hidden unit activation levels are limited to the range of 0-1, and very often the unit would have an activity level that is less than 1. Thus, when a category is already formed in first layer weights, it is easier to remap the weights in the first layer than only changing the hidden to output layer.

To overcome this difference between network and human performance in our task, we incorporated two new learning rate change rules in our network at the two different layers. Mackintosh (1975) and Pearce & Hall (1980) suggested that a major component of associative learning was that the associability of a cue, i.e., attention directed to the cue, varies over the course of training. In particular, Mackintosh (1975) hypothesized that stimuli highly predictive of an outcome should command more attention, whereas poor predictors should lose attention. In contrast, Pearce and Hall (1980) assumed that cue associability (attention) declines as critical outcomes become well-predicted. Other authors have attempted to reconcile these seemingly opposing views on how learning rates to individual cues might change across conditioning (e.g., Esber & Haselgrove, 2011; Haselgrove et al., 2010; Le Pelley, 2004; Pearce & Mackintosh, 2010). Here we implement both ideas by assuming that each applies at a different layer in the network. We apply the “Mackintosh-like” rule to describe changes in the learning rates between hidden-output layer connections and the “Pearce-Kaye-Hall-like” rule (Kaye & Pearce, 1984; Pearce et al., 1984) to learning rates between input-hidden layer connections. In our modified model, Model 2, Mackintosh-like dynamic learning rates are controlled by and dynamic Pearce-Kaye-Hall-like dynamic learning rates are controlled by (**Figure 1C**). These parameters are related to the prediction errors () generated at the output layer on each training trial and they dictate how learning rates dynamically change over the course of learning to affect input-hidden and hidden-output weight changes. For the case of the first layer (k=1, input to hidden) the Pearce-Kaye-Hall learning rate, , changes as a function of the output layer unit errors () on the current and previous trials, weighted by :

(Eq. 1)

Thus, on any given training trial , the higher the prediction error the higher the assigned to the input stimulus . This rule implicitly encodes the (Pearce & Hall (1980) intuition, where is larger when output surprise is larger. However, in order to slow the rate of change in these learning rates, the rule weights prediction errors occurring on the current and previous trials with the parameter . If = 1 then only the current trial would contribute to the computation of the updated learning rate.

In the case of the second layer (hidden to output), changes in the following way, weighted by :

(Eq. 2)

where incorporates Mackintosh’s intuition (Mackintosh, 1975) that higher s apply to stimuli accompanied by less prediction error, i.e.,small . The parameter, , similarly weights the contributions of the current and previous trials.

By applying these two rules to the different layers of the network, input-hidden weight changes should occur slowly during reversal training but hidden-output weight changes should occur rapidly. These two dynamic learning rate change rules should achieve more stability in the hidden layer category representations once learned. Our intuition was that doing so would result in less catastrophic interference when Total reversals are introduced. Essentially, the network should shift its “focus” from learning about input stimuli to learning about hidden layer category representations as the network becomes more successful at predicting the outcome. Different weight parameters (, and ; **Figure 1C**) were chosen to reflect differing contributions of current and recent trial values. We observed that this modified model faithfully reproduced our human category learning reversal data (**Figure 4B**). Higher along with a small by the end of the initial category learning phase resulted in rapid rates of hidden-output weight changes during reversal training. This, together with slow weight changes in input-hidden connections meant that the network relied heavily on hidden unit category representations to learn the reversal with mainly hidden-output weights changing and input-hidden weights being largely preserved (see **Figure 4E**). Because of these dynamic weight change rules, the network learning the Partial reversal task only with great difficulty because input-hidden weights need to change more than hidden-output weights as new exemplar-category mappings are learned at the hidden layer. The low at this layer at the end of training slows down the speed of these weight changes.

One caveat is that our simulation results are parameter-dependent. Since more successful Total than Partial reversal learning requires stability in the hidden layer category representations, more reliance on current trial prediction errors will result in more rapid input-hidden weight changes that will just result in catastrophic interference in the Total reversal task. Thus, our simulation results depend on to avoid this form of interference. When both parameters are equal, the network learns the Partial reversal faster than the Total reversal, as in Model 1 (the standard backpropagation model without dynamic learning rate changes). Nevertheless, an interesting feature of our modified model is that stability in hidden layer representations can be shown to be a function of the relative dynamic learning rates applied to the different layers of our network. It remains unknown whether other artificial network models of conditioning (Aguayo-Mendoza & Dos Santos, 2025; Sánchez et al., 2010) can provide an alternative explanation.

A screenshot of a graph

AI-generated content may be incorrect.

**Figure 4***. Simulation of the two conditions (Total and Partial reversals) in the Category Learning experiments.* ***A****. Model 1: backpropagation as is shown here cannot emulate findings with living organisms.* ***B*** *Model 2: almost the same algorithm but with dynamic learning rates applied to input-hidden and hidden-output layers. This shows faster learning to the Total reversal condition. Y-axis correct score represents the adequate unit activation divided by the sum of both unit activations.* ***C****,* ***D****,* ***E****, and* ***F*** *are the weight matrices at the end of each phase (learning or reversal), for both models (model 1:* ***C*** *and* ***D****; model 2 with dynamic learning rates:* ***E*** *and* ***F****) and for both conditions (Total:* ***C*** *and* ***E****; Partial:* ***D*** *and* ***F****). Green denotes excitatory and red inhibitory connections.*

## Conclusions

Our study contributes to the broader understanding of common coding as a general phenomenon in humans, other animals, and in artificial neural networks. The findings from our experiments provide compelling evidence for a common coding phenomenon that emerges in category learning. This was evident given that participants learnt faster the task where all exemplar categorizations were switched together in a way that preserved the initially learned exemplar-category mappings, in comparison to the Partial reversal task that forced participants to reconfigure the underlying exemplar-category mappings appropriate to new reversal phase assignments. Across both the easy (Experiment 1a) and hard (Experiment 1b) versions of our tasks, participants consistently showed better learning and faster reacquisition in Total reversal conditions compared to Partial reversals.

Our analysis through Multidimensional Scaling and psychological space measures revealed no significant preexisting differences in perceived similarities within and between categories used in Experiments 1a and 1b. This suggests that the learning effects observed in Experiments 1a and 1b were rooted in the associative mechanisms at work in these tasks such as common coding rather than perceptual resemblances of the exemplars. The multidimensional space approach allowed us to map out and quantify the 'psychological' distances between exemplars, providing a robust framework to understand the cognitive processes underlying category learning.

The computational model we employed further supports our experimental findings. By incorporating dynamic learning rates inspired by the Macintosh (Mackintosh, 1975) and Pearce-Kaye-Hall (Pearce et al., 1984) associative learning models, our artificial neural network was able to mimic the Total reversal advantage seen in living organisms. However, it is important to note that without applying these dynamic learning rate rules to the different layers of our network, reversal learning produces catastrophic interference that prevents it from learning in a more cognitively efficient manner. Future research should explore how this model can be used to provide inside in a more diverse range of category learning tasks (Navarro et al., 2019).

In summary, the research presented here advances our knowledge of category learning and common coding. By integrating behavioural experiments with computational modelling, we provide a more comprehensive perspective on how associative mechanisms drive cognitive processes such as formation of categories. Future research could build on these findings by exploring the neural correlates of common coding.

# Research Transparency Statement

***General Disclosures:***

Conflicts of interest: None. Artificial intelligence: No artificial intelligence assisted technologies were used in this research or the creation of this article. Ethics: This research received approval from the Institutional Review Board (IRB) at the City University of New York.

***Experiment 1a and 1b Disclosures:***

Preregistration: No preregistration. Materials: The code task is publicly available. Data: All primary data are publicly available. Analysis scripts: All primary scripts are publicly available. See <https://github.com/santiagocdo/categoricalReversalLerning>.

***Experiment 2 Disclosures:***

Preregistration: No preregistration. Materials: The code task is publicly available. Data: All primary data are publicly available. Analysis scripts: All primary scripts are publicly available. See [https://github.com/santiagocdo/categorySimilarity](https://github.com/santiagocdo/categorySimilarityicalReversalLerning).

**CRediT**

* Conceptualization: SC, ARD
* Methodology: SC, ARD
* Software: SC, ARD
* Validation: SC, ARD
* Formal Analysis: SC, ARD
* Investigation: SC, ARD
* Resources: SC, ARD
* Data Curation: SC, ARD
* Writing – Original Draft: SC, ARD
* Writing – Review & Editing: SC, ARD
* Visualization: SC, ARD
* Supervision: SC, ARD
* Project Administration: SC, ARD
* Funding Acquisition: SC, ARD

**Acknowledgements:**

NA

# References

Aguayo-Mendoza, M., & Dos Santos, C. V. (2025). DDM-UI: A user interface in R for the discrepancy diffuse model in behavioral research. *Behavior Research Methods*, *57*(5), 1–16. https://doi.org/10.3758/S13428-025-02648-9/FIGURES/8

Bracci, S., & De Beeck, H. P. O. (2023). Understanding Human Object Vision: A Picture Is Worth a Thousand Representations. *Annual Review of Psychology*, *74*(Volume 74, 2023), 113–135. https://doi.org/10.1146/ANNUREV-PSYCH-032720-041031/CITE/REFWORKS

Castiello, S., Zhang, W., & Delamater, A. R. (2021). The retrosplenial cortex as a possible “sensory integration” area: A neural network modeling approach of the differential outcomes effect in negative patterning. *Neurobiology of Learning and Memory*, *185*, 107527. https://doi.org/10.1016/J.NLM.2021.107527

Cousineau, D., & O’Brien, F. (2014). Error bars in within-subject designs: a comment on Baguley (2012). *Behavior Research Methods*, *46*(4), 1149–1151. https://doi.org/10.3758/S13428-013-0441-Z/METRICS

Delamater, A. R. (1998). Associative mediational processes in the acquired equivalence and distinctiveness of cues. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*(4), 467–482. https://doi.org/10.1037/0097-7403.24.4.467

Delamater, A. R. (2012). On the nature of CS and US representations in Pavlovian learning. *Learning & Behavior*, *40*(1), 1–23. https://doi.org/10.3758/s13420-011-0036-4

Delamater, A. R., & Joseph, P. (2000). Common coding in symbolic matching tasks with humans: Training with a common consequence or antecedent. *Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, *53*(3), 255–273. https://doi.org/10.1080/027249900411182,

Esber, G. R., & Haselgrove, M. (2011). Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1718), 2553–2561. https://doi.org/10.1098/RSPB.2011.0836

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178–200. https://doi.org/10.1037//0096-3445.123.2.178,

Haselgrove, M., Esber, G. R., Pearce, J. M., & Jones, P. M. (2010). Two Kinds of Attention in Pavlovian Conditioning: Evidence for a Hybrid Model of Learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(4), 456–470. https://doi.org/10.1037/A0018528

Honey, R. C., & Hall, G. (1989). Acquired Equivalence and Distinctiveness of Cues. *Journal of Experimental Psychology: Animal Behavior Processes*, *15*(4), 338–346. https://doi.org/10.1037/0097-7403.15.4.338

Johns, K. W., & Williams, D. A. (1998). Acquired equivalence learning with antecedent and consequent unconditioned stimuli. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*(1), 3–14. https://doi.org/10.1037/0097-7403.24.1.292

Kaye, H., & Pearce, J. M. (1984). The strength of the orienting response during Pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *10*(1), 90–109. https://doi.org/10.1037/0097-7403.10.1.90

Krueger, C., & Tian, L. (2004). A Comparison of the General Linear Mixed Model and Repeated Measures ANOVA Using a Dataset with Multiple Missing Data Points. *Biological Research For Nursing*, *6*(2), 151–157. https://doi.org/10.1177/1099800404267682

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, *57*(3), 193–243. https://doi.org/10.1080/02724990344000141

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), 276–298. https://doi.org/10.1037/h0076778

Makowski, D., Ben-Shachar, M. S., Patil, I., & Lüdecke, D. (2021). Automated Results Reporting as a Practical Tool to Improve Reproducibility and Methodological Best Practices Adoption. *CRAN*. https://github.com/easystats/report

Miller, N. E., & Dollard, J. (1941). *Social learning and imitation*. Yale University Press.

Navarro, V. M., Jani, R., & Wasserman, E. A. (2019). Pigeon category learning: Revisiting the Shepard, Hovland, and Jenkins (1961) tasks. *Journal of Experimental Psychology: Animal Learning and Cognition*, *45*(2), 174–184. https://doi.org/10.1037/XAN0000198

Nilsson, E. J., Aust, M. L., Engström, J., Svanberg, B., & Lindén, P. (2018). Effects of cognitive load on response time in an unexpected lead vehicle braking scenario and the detection response task (DRT). *Transportation Research Part F: Traffic Psychology and Behaviour*, *59*, 463–474. https://doi.org/10.1016/J.TRF.2018.09.026

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552. https://doi.org/10.1037/0033-295X.87.6.532

Pearce, J. M., Kaye, H., & Hall, G. (1984). Predictive accuracy and stimulus associability: Development of a model for Pavlovian learning. In M. Commons, R. Herrnstein, & A. R. Wagner (Eds.), *Quantitative analysis of behavior* (Vol. 3). Ballinger.

Pearce, J. M., & Mackintosh, N. (2010). Two theories of attention: a review and a possible integration. In C. Mitchell & M. E. Le Pelley (Eds.), *Attention and Associative Learning: From Brain to Behaviour* (pp. 11–39). Oxford University Press.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/S13428-018-01193-Y/FIGURES/3

R Core Team. (2021). *R A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. httpswww.R-project.org

Sánchez, J. M., Galeazzi, J. M., & Burgos, J. E. (2010). Some structural determinants of Pavlovian conditioning in artificial neural networks. *Behavioural Processes*, *84*(1), 526–535. https://doi.org/10.1016/J.BEPROC.2010.01.018

Sidman, M. (2000). Equivalence relations and the reinforcement contingency. *Journal of the Experimental Analysis of Behavior*, *74*(1), 127. https://doi.org/10.1901/JEAB.2000.74-127

Urcuioli, P. J., Zentall, T. R., & DeMarse, T. (1995). Transfer to Derived Sample-comparison Relations by Pigeons Following Many-to-one versus One-to-many Matching with Identical Training Relations. *The Quarterly Journal of Experimental Psychology Section B*, *48*(2), 158–178. https://doi.org/10.1080/14640749508401445

Yu, Z., Guindani, M., Grieco, S. F., Chen, L., Holmes, T. C., & Xu, X. (2022). Beyond t test and ANOVA: applications of mixed-effects models for more rigorous statistical analysis in neuroscience research. *Neuron*, *110*(1), 21–35. https://doi.org/10.1016/J.NEURON.2021.10.030

Zentall, T. R., Sherburne, L. M., Steirn, J. N., Randall, C. K., Roper, K. L., & Urcuioli, P. J. (1992). Common coding in pigeons: Partial versus total reversals of one-to-many conditional discriminations. *Animal Learning & Behavior*, *20*(4), 373–381. https://doi.org/10.3758/BF03197960

Zentall, T. R., Steirn, J. N., Sherburne, L. M., & Urcuioli, P. J. (1991). Common Coding in Pigeons Assessed Through Partial Versus Total Reversals of Many-to-One Conditional and Simple Discriminations. *Journal of Experimental Psychology: Animal Behavior Processes*, *17*(2), 194–201. https://doi.org/10.1037/0097-7403.17.2.194

# Appendix

## *Connectionist Model*

The model was implemented with code written in R and is publicly available at the following website: <https://github.com/santiagocdo/ALANN> (check folder *training\_files*: *category\_total\_reversal.csv* and *category\_partial\_reversal.csv*). The connectionist model consists of a three-layer feedforward neural network, where activations go bottom-up (from CS to US representations) and learning top-down. The layers are: input (L = 0), hidden (L = 1) and output (L = 2). This model can be described with two rules: the activation rule (that calculates unit activations) and the learning rule (that determines changes in weights or connections between units).

The connectivity constraint between the input and the hidden layer given the visual, multimodal, and auditory pathways (Delamater, 2012) is encoded in the **C** matrix (size: number of input units x number of hidden units), where 1 represents an existent connection and 0 a non-existent connection; hence, connections between visual input units to auditory hidden layer pathway units will be equal to 0, and similarly for auditory inputs and visual pathways.

The model computes weight matrices for layers 1 and 2, i.e., between input and hidden (L = 1), and between hidden and output (L = 2), at every trial *t.* The values of at *t* = 0 are random variables from a uniform distribution between -0.5 and 0.5. In order to maintain the partial connectivity between input and hidden  is multiplied (via Hadamard product or elementwise multiplication) with **C**. On the one hand, maps the connections’ weights between input and hidden units, thereby, the matrix size is the number of inputs x number of hidden units. On the other hand, represents the weights between hidden and output units, hence, the matrix size is the number of hiddenx number of outputs. At every *t*, the CS information is encoded in the activation of the input layer, , where this vector contains 0s and 1s, representing presences and absence of CSs). The USs are encoded in a vector (size is equal to the number of USs)[[1]](#footnote-1).

*Activation Rule*. The first rule of this model specifies how units are activated. The activation of the hidden and output layers is given by a sigmoidal function that receives the net input (**n**) vector for the current layer L, i.e., :

(1).

The activations are a real number between 0 and 1. We subtract 2.2. from to shift the sigmoidal curve to the right and obtain low activation levels when elements is 0. The net input for the hidden layer is:

(2),

where is a matrix multiplication between a transpose () column vector (i.e., a row vector), and the matrix.

*Learning rule*. The second rule of the model specifies how changes at every *t*. This rule is an adaptation of backpropagation (Rumelhart, Hinton, & Williams, 1986). In general terms, the weights are the sum of the current weights and the change of those weights at *t*. Hence:

(3),

Where is the Hadamard product or elementwise multiplication, and the partial connectivity only applies to input layer (L = 0) to hidden layer (L = 1) connections within the weight matrix . This ensures that weights between units with impermissible connections (e.g., auditory inputs to visual pathway hidden units) are fixed at 0. Changes in weights are determined by:

(4),

where and are free parameters, representing the learning rate (assumed to be 0.3) and the momentum decay (assumed to be 0.9), respectively. Finally, is a vector of delta values, and encodes the prediction errors that are back-propagated throughout the network, following the next equation:

(5),

where is a vector which encodes with 0s (absent) and 1s (present) the USs at every *t*.

## Instructions

### Experiment 1a and 1b

In this experiment, you will be presented with a series of abstract images that represent the molecular structure of various natural objects. Your task will be to learn from which of two regions in the world these objects come (Northern Hemisphere or Southern Hemisphere).

More specifically, you will see 1 of 8 different abstract images at a time and be asked to indicate whether you think that image reflects an object taken from the Northern or Southern Hemisphere. Choose the Left Arrow Key for Northern or the Right Arrow Key for Southern Hemisphere. At first, you will need to guess, but you will be provided with feedback after your answer to help you learn which objects come from Northern or Southern Hemispheres.

Your response times are also important. Please make your response choices as quickly, but also as accurately, as you can. Your feedback will display the time (in sec) that it took for you to reach your decision, and also if your choice was correct (with a high pitch sound) or not (low pitch sound).

There will be a break halfway through.

Press the space bar when you are ready to begin.

### Experiment 2

[categorySimilarity]

In this experiment, you will be shown pairs of abstract images. Your task is to rate how similar or different the two images appear to you.

The experiment consists of two parts. In the first part, you will see and rate 56 pairs of images, and in the second part, another 56 pairs. Below each pair, there will be a slider ranging from "very different" to "very similar." Use your mouse to select a position on the slider that best represents how similar or different you find the two images are relative to one another.

There will be a break between both parts.

Press the space bar when you are ready to begin.

[fractal\_similarity]

In this experiment, you will be shown pairs of abstract images. Your task is to rate how similar or different each pair appears to you.

The experiment consists of two parts. In the first part, you will see 56 pairs of images, and in the second part, another 56 pairs, individual images may repeat between pairs. Below each pair, there will be a slider ranging from "very different" to "very similar." Use your mouse to select a position on the slider that best represents how similar or dissimilar you find the images.

There will be a break between both parts.

Press the space bar when you are ready to begin.

## Informed Consent

This research investigates the psychological processes used when people learn to identify objects in the world. You are being asked to participate in this research study because you are a normal healthy adult, and we wish to better understand basic learning processes in your population. The purpose of this research is to gain more knowledge about the cognitive processes involved in simple forms of associative learning.

If you agree to participate, we will ask you to perform in a simple computer task that will last approximately 12 minutes. In this task, you will see a series of abstract images presented individually on the screen and your task will be to learn to choose one of two response options for each image. Also, you will be asked to respond quickly and accurately on your computer keyboard when the image appears.

• Risks/Discomforts: There are no risks for participating in this study beyond those associated with normal computer use over a 15 min period.

• Benefits: This research is not designed to directly benefit you, but your help with this study will advance basic science on the cognitive processes involved in predictive learning in normal healthy individuals. Ultimately, this research could lead to a better understanding of some of the associative learning processes that are negatively impacted by various psychological conditions (such as aging, dementia, etc).

• Confidentiality: This study does not collect identifying information, and all data collected will remain anonymous. We will ask about your gender, age, and nationality, and we will record your performance in the task itself. However, this information will not be linked directly to any individual participant. The data we obtain will be stored indefinitely and may be shared publicly via online repositories for findings that are ultimately published.

Your participation in this research is completely voluntary, and you will be able to stop at any time without penalty. If you have any questions, you can contact: Andrew R. Delamater (andrewd@brooklyn.cuny.edu). If you have any questions about your rights as a research participant or if you would like to talk to someone other than the researcher, you can contact CUNY Research Compliance Administrator at 646-664-8918 or HRPP@cuny.edu.

If you wish to participate in the study, please press the spacebar for additional instructions.

# Supplementary Information

## Experiment 1a: Sensitivity Analysis of the Reaction Time

*RT transformed with Natural Logarithm*. We ran the same LMM with condition (total versus partial), Blocks, and its interaction. We found no main effects in Total reversed versus Partial reversed [] nor Blocks [], and no Interaction [].

*RT transformed with Root Square*. We ran the same LMM with condition (total versus partial), Blocks, and its interaction. We found no main effects in Total-Condition [] nor Blocks [], and no Interaction [].

## Experiment 1b: Sensitivity Analysis of the Reaction Time

*RT transformed with Natural Logarithm*. We ran the same LMM with condition (total versus partial), Blocks, and its interaction. We found the main significant effects in Total reversed versus Partial reversed [] and Blocks [], but no Interaction [].

*RT transformed with Root Square*. We ran the same LMM with condition (total versus partial), Blocks, and its interaction. We found a main effect in Total reversed versus Partial reversed [], but not in Blocks [] nor the Interaction [].

## Experiment 2: Multidimensional Scaling (*k*=3)

**A diagram of a distance

AI-generated content may be incorrect.**

**Figure S1**. *Category Similarity Analysis, using two psychological dimensions (multidimensional scaling with k=3) in the psychological space.* *Left panels (****A*** *and* ***C****) represent Experiment 1a results for the exemplars for Set 1, and the right panels (****B*** *and* ***D****) represent the comparisons conducted with exemplars in Set 2.*

1. All vectors are considered to be column vectors. Hence to get a row vector we transpose them, e.g. **x** (column vector), **x**T (row vector). [↑](#footnote-ref-1)