SUSTAIN captures category learning, recognition, and hippocampal activation in a unidimensional vs information-integration task

Anonymous CogSci submission

Abstract

There is a growing interest in alternative explanations to the dual-system account of how people learn category structures varying in their optimal decision bounds (unidimensional and information-integration structures). Recognition memory performance and hippocampal activation patterns in these tasks are two interesting findings, which have not been formally explained. Here, we carry out a formal simulation with SUSTAIN (Love, Medin, & Gureckis, 2004), an adaptive model of category learning, which had great success in accounting for recognition memory performance and fMRI activity patterns. We show, for the first time, that a formal single-system model of category learning can accommodate recognition performance after learning and is consistent with fMRI data obtained while participants learned these structures.

Keywords: categorization; recognition memory; formal model; SUSTAIN

Introduction

One commonly used pair of category structures in categorization research are the unidimensional (UD) and information-integration (II) category structures. The shared characteristic of UD and II is that they are perceptually the same, but differ in their optimal decision bounds. These properties were initially used for separating perceptual processes encoding the visual information from decision processes assigning a category response to the perceptual effects (Ashby & Gott, 1988). Figure 1 shows how stimuli varying in size and brightness are distributed within these two category structures on either side of the boundaries. UD category structures have a vertical decision bound: if the square is darker than x, then it is category A, otherwise it is category B. Figure 1 shows that this optimal decision bound is parallel to one of the dimensional axis in psychological space. II structures are defined by diagonal optimal decision bounds. Figure 1 shows that II decision bounds follow a linear function, where the gradient is neither zero, nor infinite.

Many experiments utilised these structures (e.g. Carpenter, Wills, Benattayallah, & Milton, 2016; Donkin, Newell, Kalish, Dunn, & Nosofsky, 2015; Dunn, Newell, & Kalish, 2012; Nomura et al., 2007; Kalish, Newell, & Dunn, 2017; Le Pelley, Newell, & Nosofsky, 2019) and many initial empirical results were taken as evidence for COVIS (Ashby, Paul, & Maddox, 2011) — one formalization of a dual-system theory of categorization. Traditionally, dual-system theories have two distinct architectures using functionally different mechanisms. In COVIS, the explicit system uses rules that can be easily verbalized, while the implicit system maps perceptual input onto category responses.

However, results from multiple labs pointed out flaws in the experimental designs (Newell, Moore, Wills, & Milton, 2013) with potential alternative explanations (Le Pelley et al., 2019; Donkin et al., 2015; Dunn et al., 2012) or problems with the decision-bound analysis themselves (Edmunds, Milton, & Wills, 2018; Edmunds, Wills, & Milton, 2019). The way COVIS explains how people learn II structures was also questioned by Carpenter et al. (2016) and Edmunds, Wills, and Milton (2016), who provided direct evidence for an involvement of similar processes in both II and UD problems.

Carpenter et al. (2016) found that the medial temporal lobe (MTL) and specifically the hippocampus (HPC) were more active in the category learning task involving II structures compared to the task with UD structures. As MTL and HPC is thought to be essential for explicit memory, this finding suggest that people should have explicit access to category knowledge in II to a greater extent than in UD structures. This prediction contradicts to the COVIS account of how II structures are learned. Edmunds et al. (2016) directly tested whether people have conscious access to information in II structures by including a recognition task in both UD and II problems. Edmunds et al. (2016) found better recognition memory after learning II compared to UD structures, essentially supplementing the neural data.

Building on these findings, we further supplement behavioral and neural data with evidence from computational modelling. Here, we provide a formal single-system explanation of these results. We do so by using SUSTAIN (Love et al., 2004).

SUSTAIN is a model in a single-system framework able to accommodate a wide range of behavioral and neural phenomena (e.g. Love & Medin, 1998b, 1998a; Love, Markman, & Yamauchi, 2000; Love et al., 2004; Gureckis & Love, 2003, 2004; Davis, Love, & Preston, 2012). This breadth is particularly admirable, because modelers tend to focus on a small subset of effects (Wills, O'Connell, Edmunds, & Inkster, 2017).

There are two reasons for using SUSTAIN. First, SUSTAIN can accommodate recognition memory performance in multiple tasks (Love & Gureckis, 2007; Davis et al., 2012; Mack, Love, & Preston, 2018). Second, SUSTAIN's concept-forming and -altering mechanism, adaptive clustering, has

been mapped to HPC and MTL functions and activations.

Cluster-specific model components in SUSTAIN have been directly connected to strong HPC activations present in early learning and low HPC functions in amnesic patients in a dot-pattern classification task (for a more exhaustive review, see Love & Gureckis, 2007). SUSTAIN views the hippocampus as the constructor and editor of clusters — binding information together into category representations, and views the MTL familiarity signals as indicators of cluster re-activations. These views have been reinforced by connecting computational modelling to neural activity patterns. For example, during rule-plus-exception learning, SUSTAIN makes predictions about item recognition. These predictions directly and consistently mapped to MTL activations (Davis et al., 2012). Furthermore, SUSTAIN's cluster-updating mechanism parallels HPC activity in response to changing task demands. SUSTAIN accommodates behavioral responses and HPC activity in subsequent learning tasks where the stimuli remains perceptually the same (Mack, Love, & Preston, 2016), while irrelevant features in the first task became essential in the new categorization problem. SUSTAIN is well matched with how the HPC binds together information into meaningful category representations and updates the stored representations to match with goal-oriented changes in task-demands (for a complete review, see Mack et al., 2018).

This logic leads to two predictions for SUSTAIN in II and UD problems: (1) increased HPC activity in II should correspond with a higher number of clusters being recruited in II, because the more difficult task requires the binding and storing of larger sets of information into clusters than the simpler task (Love & Gureckis, 2007); (2) higher MTL and HPC activity and better recognition memory in II than in UD should correspond with SUSTAIN recruiting higher number of clusters and recognizing more items. In this paper, we test these predictions by simulating Edmunds et al.'s (2016) experiment with SUSTAIN.

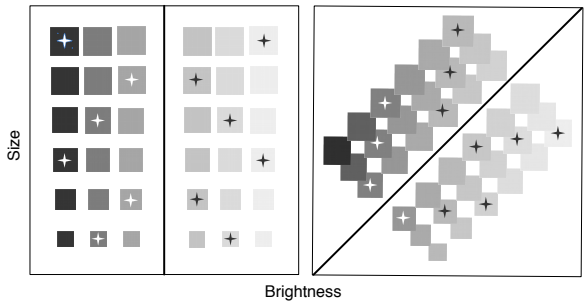

Figure 1: Representations of the category structures used in Edmunds et al. (2016). Left: Unidimensional, Right: Information-integration. The stars (colour irrelevant, not shown to the participants) mark possible stimuli that were removed during the training phase of the experiment.

Description of SUSTAIN

We refer the reader to Love et al. (2004) for the full description of the model's architecture and Love and Gureckis (2007) for the full description of the supplementing architecture capturing recognition memory.

Briefly, SUSTAIN is an adaptive clustering model, which proposes that clusters underlie category representations (Love, 2005). Clusters, from SUSTAIN's perspective, are single coordinates in the representation space. These coordinates are internal representations that connect to categories. SUSTAIN starts with one cluster, centered on the first input representation it encounters by default. When SUSTAIN encounters a stimulus, it computes similarity from all stored cluster representations in the psychological space. First, the distance is calculated for each dimension, then differentially weighted in the cluster activation function by attentional tunings. So similarity on dimensions with higher attentional tunings will be more impactful on which cluster is activated. The winning cluster will be the one with the highest activation. After this algorithm, clusters are laterally inhibited by each others' activations. Laterally inhibited activations are considered to reflect the models' overall familiarity with the current stimulus. The sum of these activations, Recognition score R, indexes this stimulus-specific familiarity. Lateral inhibition then ends in a winner-takes-all fashion — non-winning clusters' activations are muted for calculating further response probabilities.

Activations after lateral inhibition spread to the category output units by weighted connections. The activations of each output units are turned into response probabilities. If the model made the correct response, then the winning cluster's position is adjusted by moving it closer to the current input representation. In the event of a prediction error (an incorrect response) a new cluster centered on the current input representation is recruited. SUSTAIN prefers simple solutions, and only starts recruiting clusters in response to prediction errors. This means that more difficult tasks will cause SUSTAIN to densely populate the psychological space with clusters.

Simulation of Edmunds et al. (2016)

In the following, we present a formal simulation with the SUSTAIN model capturing human categorization accuracy in II and UD structures. In addition, we show how the model captures better recognition memory following the II compared to the UD recognition task (Edmunds et al., 2016). This is supplemented by more clusters recruited for II, which mirrors the higher hippocampal activation while learning the II structures compared to UD structures (Carpenter et al., 2016). We do so by fitting SUSTAIN to an abstract design similar to that of Edmunds et al. (2016). We decided on Edmunds et al. (2016), because this allowed us to present the model with a close approximation of the conditions present where the authors observed better recognition performance in II.

Edmunds et al. (2016) used 36 grey squares that varied in

brightness and size[1]. There were four conditions. UD structures included both horizontal and vertical category boundaries. II structures involved diagonal category boundaries with both positive and negative gradients. Figure 1 shows how these stimuli are distributed in the 2D physical stimulus space.

Each condition consisted of three phases. First, the training phase included 360 supervised training trials in blocks of 120. Each simulated participant received a 24 stimuli randomly picked from the 36 for their simulation. Those 24 stimuli were shown 5 times in each of the 3 blocks. This was followed by an OLD/NEW recognition phase. This phase consisted of 3 blocks of all 36 stimuli. The last phase was a categorization test phase. This phase was similarly made up of 3 blocks of the all 36 items. For a more detailed description of experimental procedure, see Edmunds et al. (2016).

Simulation

Our implementation of SUSTAIN is available in the R package catlearn (Wills et al., 2017). SUSTAIN's parameters were adjusted to minimise the sum of squared errors between the overall mean categorization test phase accuracy of SUSTAIN and humans. The trial-order was randomised on each iterations. The model was fitted with a differential evolutionary algorithm, as implemented in the DEoptim package (Mullen, Ardia, Gil, Windover, & Cline, 2011). The algorithm iterated 1000 times to find the best fitting parameters. The speed of crossover was set to $c = 0.5$, which gave larger weights to successful mutations. The top 30% best solutions were copied to the new iteration and was used in the new mutated population. These settings helped to find the single overall best parameter set for all category structures across different trial orders. The best fitting parameters are presented in Table 1.

Table 1: Best fitting parameters for SUSTAIN rounded to the 9th decimal place and their corresponding Upper Bounds set during the parameter search.

Parameters	Best Fitting
Attentional focus (r)	4.1301
Lateral inhibition (β)	8.3273
Decision consistency (d)	1.9883
Learning rate (η)	0.0626

Categorization Test Phase Accuracy SUSTAIN's categorization performance is qualitatively similar to what we observed from humans — II structures are harder to learn than UD. SUSTAIN matches human-level categorization test performance with a mean difference of 0.014, see Table 2.

[1]In our simulations, these values were put in a range $[0, 1]$ within each dimension. The values as specified by their respective coordinates are available in the supplementary material

Table 2: Categorization accuracy in SUSTAIN and humans.

Category Structures	SUSTAIN	Human
II	0.78	0.78
UD	0.85	0.87

Cluster Recruitment and Attentional Tuning The mean number of clusters recruited were $M_{ii} = 5.59$, $SD_{ii} = 1.20$ for II and $M_{ud} = 3.01$, $SD_{ud} = 1.18$ for UD. SUSTAIN solves II with a minimum of 4 clusters and a maximum of 12 clusters. SUSTAIN solves UD with a minimum of 2 and a maximum of 12 clusters. Example clusters populating the psychological space are shown in Figure 2.

The mean, and variation, in the number of clusters is the consequence of how trial-order interacts with the following mechanisms: similarity, attention and error-driven cluster recruitment. Simple problems on average result in fewer clusters, while harder problems require the recruitment of more clusters. This is attenuated by differentially weighing in relevant information from each dimension — by attentional tuning of perceptual inputs. Each dimension has its own attentional tuning, λ. For example, λ is higher for relevant dimensions in UD structures, but remains comparable across dimensions in II, see Table 3.

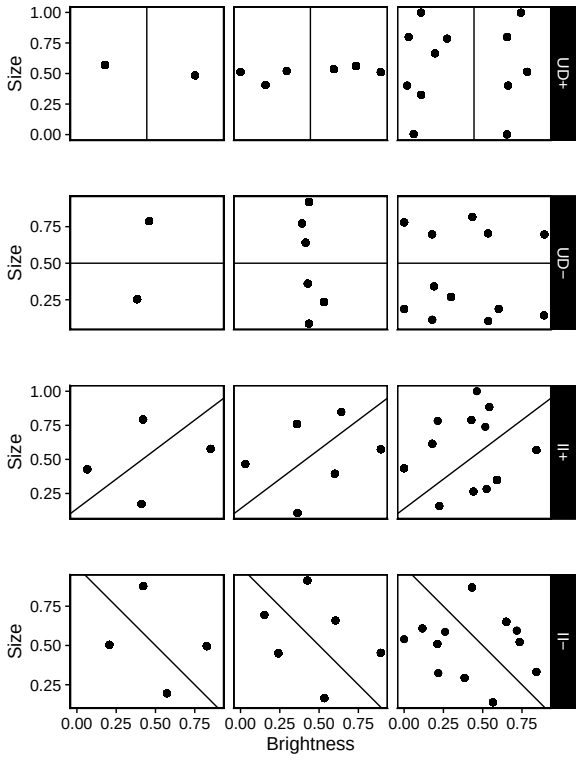

Figure 2: Example clusters recruited by SUSTAIN for three simulated participants across conditions. The juxtaposed black lines are the optimal decision bounds. UD = unidimensional; II = information-integration; +/- = vertical/horizontal for UD and positive/negative for II.

Table 3: Mean λ values for each dimension across all category structures. UD = unidimensional; II = information-integration; +/- = vertical/horizontal for UD and positive/negative for II.

Stimuli Dimensions	II+	II-	UD+	UD-
X (Brightness)	13.52	13.47	13.38	7.80
Y (Size)	13.46	13.51	7.80	13.39

Overall, SUSTAIN requires higher numbers of clusters to solve II due to its difficulty. This will result in clusters more perfectly matching training items, so the matching clusters will dominate the activation function. From the model's point of view, we see higher HPC activations in II, because the category structure requires more representations to be encoded by the HPC. MTL has been shown to be responsible for similarity-based familiarity signals, and the HPC is specifically responsible for encoding new clusters after surprising events (Love & Gureckis, 2007). HPC activations have been shown to positively relate to cluster activations, updates and recruitments in SUSTAIN (Mack et al., 2016, 2018).

Recognition To get an approximate d' measure from R, Recognitions Score, we applied Equation A11 from Love and Gureckis (2007) to turn stimulus-specific R values during the categorization test phase into choice probabilities: $P(old) = R/(R+k)$ where k is a response threshold parameter. We calculated the mean probability of a hit $(P(H) = P(old \mid item_{old})$ and false alarm $P(F) = P(old \mid item_{new})$ for each participant. We continued to determine d' for each participant using the z-transformed $P(H)$ and $P(F)$. Then we calculated group-level averages. This algorithm (including Equation 1 and the group-level d' calculations) were fitted against human performance in the recognition phase as indexed by d'. Similarly, we used DEoptim and reitereted the paramater search 50 times. More details are included in the code available in the supplementary material. We found that the best-fitting parameter k was 0.571. This parameter will not change the ordinal pattern of the recognition performance ($II > UD$) SUSTAIN shows given our previous fit to the categorization test data, but simply brings the values closer to human data. While this difference is rather small, it is statistically present in the between-groups comparison.

Table 4 shows the performance of humans and SUSTAIN. The model shows better recognition performance observed after learning II structures by Edmunds et al. (2016), but predicts higher false alarm rates for UD structures. A comparison of d' between SUSTAIN and human data yields a mean difference of 0.0071. This difference of d' and II and UD results from the difference in the number of recruited clusters between the two structures.

Recognition in SUSTAIN is based on similarity-driven cluster activation and lateral inhibition. Higher number of clusters recruited will be located more closely to the input representations encountered during test. This means that the stored represention will match better to the model's previous experience in II than in UD problems. The more densely populated the psychological space with clusters, the more clusters neighbouring the input representation will activate. These activations then compete and will diminish as a result of lateral inhibition. The better recognition memory performance in II results from the higher sum of activations in regions neighbouring the input representations.

This benefit parallels HPC activation patterns. Better recognition memory performance follows not just from the modelling perspective, but also from a neural point-of-view. HPC has been long identified as crucial for memory (O'Reilly & Rudy, 2001; Schlichting & Preston, 2015). Love and Gureckis (2007) predicted this relationship, where high number of clusters mirror higher levels of HPC involvement. This prediction strongly aligns with Carpenter et al. (2016), who observed higher HPC involvement in the II compared to UD task, and our simulation, where SUSTAIN recruits more clusters for the II task.

Table 4: Mean recognition scores and d' for each category structure. Standard deviations are in parentheses.

	SUSTAIN d'	Human d'
II	0.040 (0.056)	0.01 (0.02)
UD	-0.016 (0.135)	0.00 (0.01)

Discussion

We have presented a formal account of empirical results (Edmunds et al., 2016; Carpenter et al., 2016) concerning the acquisition of unidimensional (UD) and information-integration (II) category structures. In so doing, we have shown - for the first time - that both the behavioral and neuroimaging data obtained in these tasks can be accomodated by a single-system model, SUSTAIN. The increased number of clusters recruited by SUSTAIN for the II structure served as a base for better recognition memory performance, and larger HPC activation, than in the UD structure. According to SUSTAIN, this is because the differing task demands of the two structures requires a larger amount of information to be encoded in the HPC for II structures.

Previously, Davis et al. (2012) speculated that tasks like the II category learning were not suitable to model with SUSTAIN. This sentiment was based on the idea that II category learning is a procedural learning task (Nomura et al., 2007) — and hence characterized by mechanisms not specified within SUSTAIN. However, procedural accounts of II problems are based on a range of experiments that received considerable scrutiny which turn out to have alternative explanations (Newell, Dunn, & Kalish, 2011; Stanton & Nosofsky, 2007).

The only formal model — before SUSTAIN — that has been argued to accomodate both structures was COVIS. COVIS posited a procedural account of how people learn II struc-

tures. COVIS solves II with a procedural learning mechanism conceptualized as a three-layer network: the first layer calculates the exponent of the distance between activated input units and and sensory units; the second layer attenuates these similarities by weighted connections between sensory units and striatal units before spreading to the striatal units; in the third layer, a decision rule responds with the most activated striatal unit; and then the weights are updated. It is a distributed-representation connectionist network, where input node activations are supplied by a distance between sensory unit coordinates and input representation in the psychological space. COVIS solves UD by a different, rule-based system, which establishes a decision bound dominating responding. Therefore, at limit COVIS predicts no recognition memory for either category structures. This still doesn't allow better recognition in II than UD. One approach would be to create an architecture that converts similarity derived from sensory input and weighted connections to activations of memory traces. A similar approach has been used to describe recognition by multiple-trace memory models (Hintzman, 1986), but this can require the assumption that rule-based and procedural systems are able to access the representation space where these values are stored.

Conclusion

We formally show that a single-system adaptive clustering model, SUSTAIN, can accommodate categorization and recognition performance in two frequently used category structures, information-integration and unidimensional. The behavior of the model is also consistent with MTL and HPC activity involved in learning these structures. Our simulation not only provides a formal account of how people learn these structures, but also contributes to the literature bridging formal models of category learning, behavior and the brain.

Open Science Statement

All simulation code is available in the Open Sciences Framework at `https://osf.io/jc9xs/?view_only=c00913447a7d43deb1d61471c91b11ae`.

References

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33.

Ashby, F. G., Paul, E. J., & Maddox, W. T. (2011). COVIS. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 65–87). New York: Cambridge University Press.

Carpenter, K. L., Wills, A. J., Benattayallah, A., & Milton, F. N. (2016). A comparison of the neural correlates that underlie rule-based and information-integration category learning. *Human Brain Mapping*, *37*, 3557–3574. doi: 10/gg58ps

Davis, T., Love, B. C., & Preston, A. R. (2012). Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. *Cerebral Cortex*, *22*(2), 260–273. doi: 10/c42v8p

Donkin, C., Newell, B. R., Kalish, M., Dunn, J. C., & Nosofsky, R. M. (2015). Identifying strategy use in category learning tasks: A case for more diagnostic data and models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 933–948. doi: 10/gg58pv

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 840–859. doi: 10/gg58pt

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2018). Due process in dual process: Model-recovery simulations of decision-bound strategy analysis in category learning. *Cognitive Science*, 1–28. doi: 10/gdqzc5

Edmunds, C. E. R., Wills, A. J., & Milton, F. N. (2016). Memory for exemplars in category learning. In A. Papfragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2243–2248). Austin, TX: Cognitive Science Society.

Edmunds, C. E. R., Wills, A. J., & Milton, F. N. (2019). Initial training with difficult items does not facilitate category learning. *Quarterly Journal of Experimental Psychology*, *72*(2), 151–167. doi: 10/gg58pj

Gureckis, T. M., & Love, B. (2004). Common mechanisms in infant and adult category learning. *Infancy*, *5*(2), 173–198. doi: 10/dg6x54

Gureckis, T. M., & Love, B. C. (2003). Towards a unified account of supervised and unsupervised category learning. *Journal of Experimental and Theoretical Artificial Intelligence*, *15*(1), 1–24. doi: 10/fhmz7z

Hintzman, D. L. (1986). " schema abstraction" in a multiple-trace memory model. *Psychological review*, *93*(4), 411.

Kalish, M. L., Newell, B. R., & Dunn, J. C. (2017). More is generally better: Higher working memory capacity does not impair perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(4), 503. doi: 10/gcx82s

Le Pelley, M. E., Newell, B. R., & Nosofsky, R. M. (2019). Deferred feedback does not dissociate implicit and explicit category-learning systems: Commentary on smith et al.(2014). *Psychological science*, *30*(9), 1403–1409.

Love, B. C. (2005). Environment and goals jointly direct category acquisition. *Current Directions in Psychological Science*, *14*(4), 195–199. doi: 10/dfvrbg

Love, B. C., & Gureckis, T. M. (2007). Models in search of a brain. *Cognitive, Affective & Behavioral Neuroscience*, *7*(2), 90–108. doi: 10/ftn2cn

Love, B. C., Markman, A. B., & Yamauchi, T. (2000). Modelling classification and inference learning. In H. Kautz & B. Porter (Eds.), *Proceedings of the seventeenth national conference on artificial intelligence* (pp. 136–141). Cambridge, MA: MIT Press.

Love, B. C., & Medin, D. L. (1998a). Modeling item and category learning. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the 20th annual conference of the cognitive science society* (pp. 639–644). Mahwah, NJ: Erlbaum.

Love, B. C., & Medin, D. L. (1998b). SUSTAIN: A network model of human category learning. In C. Rich & J. Mostow (Eds.), *Proceedings of the fifteenth national conference on artificial intelligence* (pp. 671–676). Cambridge, MA: MIT Press.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332. doi: 10/fm394p

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, *113*(46), 13203–13208. doi: 10/f9fvzr

Mack, M. L., Love, B. C., & Preston, A. R. (2018). Building concepts one episode at a time: The hippocampus and concept formation. *Neuroscience Letters*, *680*, 31–38. doi: 10/gd5m53

Mullen, K., Ardia, D., Gil, D., Windover, D., & Cline, J. (2011). DEoptim: An R package for global optimization by differential evolution. *Journal of Statistical Software*, *40*(6), 1–26. doi: 10/ggt446

Newell, B. R., Dunn, J. C., & Kalish, M. (2011). Systems of category learning: Fact or fantasy? In *Psychology of learning and motivation* (Vol. 54, pp. 167–215). doi: 10 .1016/B978-0-12-385527-5.00006-1

Newell, B. R., Moore, C. P., Wills, A. J., & Milton, F. (2013). Reinstating the Frontal Lobes? Having More Time to Think Improves Implicit Perceptual Categorization: A Comment on Filoteo, Lauritzen, and Maddox (2010). *Psychological Science*, *24*(3), 386–389. doi: 10/gg58pk

Nomura, E. M., Maddox, W. T., Filoteo, J. V., Ing, A. D., Gitelman, D. R., Parrish, T. B., ... Reber, P. J. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cerebral Cortex*, *17*(1), 37–43. doi: 10/dd3rdr

O'Reilly, R., & Rudy, J. (2001). Conjunctive Representations in Learning and Memory: Principles of Cortical and Hippocampal Function. *Psychological Review*, *108*(2), 311–345.

Schlichting, M. L., & Preston, A. R. (2015). Memory integration: Neural mechanisms and implications for behavior. *Current Opinion in Behavioral Sciences*, *1*, 1–8. doi: 10/gg775s

Stanton, R. D., & Nosofsky, R. M. (2007). Feedback interference and dissociations of classification: Evidence against the multiple-learning-systems hypothesis. *Memory & Cognition*, *35*(7), 1747–1758.

Wills, A. J., O'Connell, G., Edmunds, C. E. R., & Inkster, A. B. (2017). Progress in modeling through distributed collaboration. In *Psychology of learning and motivation* (pp. 79–115). Elsevier.