Recognition performance after rule-based and information-integration categorization.

C. E. R. Edmunds, Lenard Dome, Andy J. Wills & Fraser Milton

June 25, 2020

Abstract

A large portion of the category learning literature is dedicated to discussing the stimulus representations underlying categorization. The SUSTAIN model has proved successful in accounting for categorization effects by proposing a middle ground between exemplar, prototype and rule representations. One exception to SUSTAIN's success may be in accounting for learning in information-integration tasks, which have been extensively argued to be learned implicitly (contrary to the predictions of SUSTAIN). However, recent work has failed to evidence for an implicit mechanism in this task. Therefore, in the following we investigated explicit recognition performance after learning either rule-based or information-integration tasks. We found that, contrary to the claims in the literature, that recognition memory was better following the information-integration task than the rule-based task and that this is predicted by the SUSTAIN model. Our results add to the growing literature that suggests that both of these categorization tasks are learned explicitly.

The category learning literature is abundant with formal models, all making disparate assumptions about stimulus representation, generalization, grouping and decision-making (Pothos & Wills, 2011). One of these, the SUSTAIN model (Supervised and Unsupervised STratified Adaptive Incremental Network; Love, Medin, & Gureckis, 2004), accounts for an impressive variety of categorization phenomena: learning rates (Love & Medin, 1998b); superior recognition memory for exception items (Davis, Love, & Preston, 2012); identification (Love & Medin, 1998a); inference learning (Love, Markman, & Yamauchi, 2000); developmental categorization trajectories (Gureckis & Love, 2004); and unsupervised learning (Gureckis & Love, 2002, 2003). This breadth is particularly admirable given modelers' tendency to focus on a small subset of effects (Wills, O'Connell, Edmunds, & Inkster, 2017).

One area has been firmly placed outside the explanatory scope of SUSTAIN is procedural learning (Davis et al., 2012). The category representations formed by SUSTAIN and by procedural learning mechanisms are fundamentally different. SUSTAIN represents category structures using clusters of stimuli (Love et al., 2004), where each category might be represented by one or many clusters. In contrast, procedural learning proceeds incrementally (Ashby, Alfonso-Reese, Turken, & Waldron, 1998): visual inputs are gradually associated with a particular motor response using reward prediction error to moderate the weights. Thus, the response acts as a proxy for the category label: there is no intermediate generalisation stage. These contrasting approaches mean that SUSTAIN predicts very different behaviour to that observed in tasks optimally learned by a procedural mechanism.

In the current work, we will consider a particular task widely argued to be learned procedurally: the information-integration (II) category structure (see Figure 1 Ashby & Valentin, 2018). Optimal performance on an II task is argued to require participants to combine (at least) two independent, non-commensurate stimulus dimensions at a pre-decisional stage (Ashby et al., 1998). As the optimal decision is difficult to describe in

this task[1] many have argued that II tasks are learned procedurally (for reviews, see Ashby & Maddox, 2005, 2011; Ashby & Valentin, 2017, 2018; Smith & Church, 2018).

The research often cited in support of the claim that II tasks are learned procedurally are studies that manipulate factors widely thought to affect procedural learning mechanisms. For instance, as prediction error is critical to this process, procedural learning mechanisms are thought to be sensitive to the nature and timing of feedback (Ashby et al., 1998; Ashby & Maddox, 2005, 2011). In agreement with this, learning of II tasks is poorer when feedback is absent (Ashby, Queller, & Berretty, 1999), or delayed (Maddox, Ashby, & Bohil, 2003; Maddox, Bohil, & Ing, 2004), or deferred (Smith et al., 2014). It is improved when more detailed feedback is given (Ashby & O'Brien, 2007) and there is an opportunity for error prediction (Ashby, Maddox, & Bohil, 2002). In contrast, SUSTAIN is much less sensitive to the type and format of feedback as it can account for unsupervised category learning, where there is no feedback at all (Gureckis & Love, 2002, 2003).

[1]One contorted way of describing Figure 1 might be "If the stimulus is darker than it is large, it is in Category A."

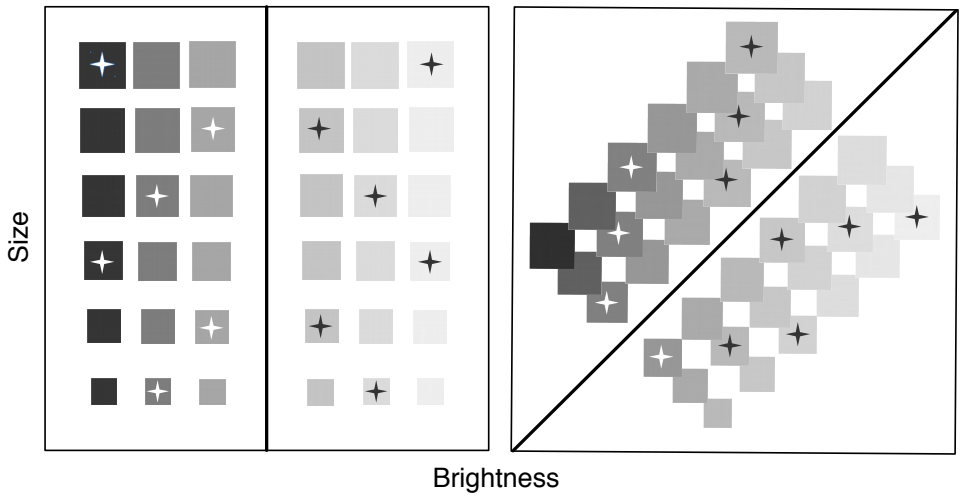


Figure 1. Representations of the category structures used in the following experiments. Left: Unidimensional rule-based, Right: Information-integration. The stars (colour irrelevant, not shown to the participants) mark a possible pattern of stimuli that were removed during the training phase of the experiment.

A further difference between procedural learning mechanisms and SUSTAIN is that they differ in their predictions about explicit knowledge or awareness of the information learned. SUSTAIN does not claim to learn implicitly. Further, it makes detailed predictions about explicit recognition memory performance following learning. In contrast, many have argued that the results of a procedural learning mechanism are not accessible to consciousness (e.g., Smith et al., 2014, 2015). Therefore, in the following we explore recognition performance following category learning to determine whether a cluster-based approach can successfully learn II category structures.

The rest of the paper proceeds as follows. First, we use formal modelling to specify what SUSTAIN would predict for recognition performance following category learning. We take this formal modelling step as skipping it would remove one of the great benefits of formal models (Wills & Pothos, 2012; Wills et al., 2017). Further, we have previously found that the verbal descriptions associated with formal models often imply conclusions or patterns of data that are subsequently falsified. For instance, we found that a model that have been extensively argued to be able to optimally learn overall similarity structures (such as the II task) actually struggled. Specifically, we looked to see whether the implicit Procedural learning mechanism from the COVIS model (Ashby et al., 1998) could learn an overall similarity structure (Edmunds & Wills, 2016). We found that without extensive changes to the model, the learning violated a key experimental feature: without changes, the learning curve was not monotonic. Therefore, in the following we looked to see whether SUSTAIN could learn the information-integration category structure. Second, we describe four experiments in which we investigated subsequent recognition performance following category learning of II category structures. We then use Bayesian modelling approaches to combine the results of these four studies, in order to better estimate whether recognition performance is improved or not. Finally, we discuss the implications of our findings.

SUSTAIN

Here, we describe the process of generating predictions from the SUSTAIN model for the learning and recognition of UD and II category structures. To begin, we describe the approximation of the formal architecture of SUSTAIN. Note that SUSTAIN is designed to incorporate both supervised and unsupervised learning but, given the nature of the task, here we only discuss supervised learning. Then, we describe the abstract category structures and the model fitting process. Finally, we describe how we generated predictions from the model as well as highlighting the key features of these predictions.

So, the SUSTAIN model. For a more detailed description of the formal model, we refer the reader to Love et al. (2004). Broadly, SUSTAIN learns category structures by assigning stimuli to clusters. The first cluster SUSTAIN creates is always centred on the first stimuli representation fed to the model. Thus, the category structures are represented in two levels: a single category can be represented by one or many clusters with each cluster containing one or many stimuli. SUSTAIN aims for the simplest possible solution - typically sorting stimuli on the basis of a single dimension. If a simple solution is not adequate, SUSTAIN adds more clusters as needed. This adaptiveness allows SUSTAIN to account for both rule-based and exemplar-based learning in a single framework.

Now looking more closely at the mechanism of the model on a trial-by-trial basis. Given a new stimulus, the model compares it to the features of all existing clusters. Similarity is defined as the distance between stimulus and cluster representations. Each cluster activates based on its similarity to the stimulus weighted by attention. This means that each dimension is differentially impactful in this stimulus-cluster comparison. Higher activations will belong to clusters, which are the most similar to the stimulus representation.

These activations will eventually compete to respond to the stimulus. If there are many competing alternatives, clusters will have lower output values after competition - the

model will be less confident in the choice. SUSTAIN selects the cluster with the highest output value and uses it to calculate the probability of making a response. It then updates the winning cluster to be more similar to the stimulus.

SUSTAIN only recruits a new cluster after a surprising event. In supervised learning, the model expands its architecture in response to a prediction error. If there is no surprising event, the model clusters together similar items. In unsupervised learning, the model recruits clusters when the stimulus mismatches existing clusters.

Recognition in SUSTAIN

The model was also supplemented to account for recognition performance of amnesic patients, infants, young adults, and older adults (Love & Gureckis, 2007). This was the addition of recognition scores, which were essentially the sum of output activations for all clusters (Equation A6 in Love & Gureckis, 2007). The smaller the cluster, the greater the recognition memory associated with that cluster (Davis et al., 2012). So the higher the activation score is, the better the recognition of an item will be.

For example, let us assume that SUSTAIN recruits one cluster that captures large sets of stimuli within category. This will produce lower recognition scores overall, because most stimuli will be further away from the prototypical cluster. But this will also produce higher recognition scores for a small subset of stimuli happen to be closer to the prototypical cluster.

On the other hand, SUSTAIN can also recruit many clusters that captures small sets of stimuli within category. This will produce higher recognition scores overall, because most stimuli will be close to its representative (exemplar) cluster. Interestingly, this suggests that more complex problems will result in better recognition, because SUSTAIN will recruit multiple clusters for each category.

The formal description of this recognition mechanism have its limitations. SUSTAIN mathematically specifies recognition scores after lateral inhibition has taken

place. It will only suffice for small number of clusters. If the model recruits large number of clusters with a relatively low cluster competition parameter, the output activation scores (Equation 6 Love et al., 2004) will be inhibited. This represents the idea that many competing alternatives will reduce the confidence in the choice. So their sum will also be smaller compared to having two clusters with high cluster competition parameter. This contradicts to the way SUSTAIN incorporates recognition outlined above.

We decided to remedy this discrepancy between the theory and the exact, formal specification of the mechanism in two steps. First, we decided to sum activation scores before lateral inhibition takes place. This ensures that it reflects the model's overall familiarity with the stimulus and not its confidence. Second, we applied Shannon Entropy to the set of retrieved cluster activations. Shannon Entropy is a well-established way to quantify the information content of any stochastic random variable set. This represent the idea, that you encode and can retrieve more information by recruiting many exemplar clusters than creating one prototypical cluster. The more clusters are active, the more information you retrieve from memory. This recognition entropy, R_e, is specified as:

$$R_e = -\sum_{i=1}^{n} \Pr(H_j^{act}) \times \log_2 \Pr(H_j^{act}) \tag{1}$$

where n is the number of clusters stored in SUSTAIN. H_j^{act} is the activation of j cluster as output by Equation 5 in Love et al. (2004). $\Pr(H_j^{act})$ is the probability of j cluster. The probabilities are calculated by dividing the activation of cluster j with the sum of n cluster activations. When R_e is high, cluster retrieval results in more information available to make a recognition judgement. When R_e is low, cluster retrieval results in less information, so you are more likely to judge an item to be old or don't recognize an item. We are interested in overall performance. So a simple descriptive statistic of the trial-by-trial R_e will be sufficient for our current endeavour. But it is possible to map R_e to a choice axiom, where the model must judge whether it has seen the stimulus before or not.

Model fitting

To examine the performance of the SUSTAIN model in UD and II tasks, we used the `catlearn` package (Wills et al., 2018) implemented in R (R Core Team, 2015). This package includes implementations of prominent category and associative learning models along with canonical datasets (Wills et al., 2017).

In the following reported simulations, we had two aims: to show that SUSTAIN could indeed learn both UD and II tasks, and to find a measure that might discriminate learning between UD and II tasks. To start, we determined the parameters that would result in SUSTAIN best predicting the category labels from the stimuli for both the UD and II tasks. The category structures we used are shown in Figure 1. In addition, we also fitted SUSTAIN to the counterbalanced version of these category structures corresponding to a $\pi/2$ rotation. For the UD structure, this is where the decision boundary would discriminate based on size not brightness. For the II structure, the decision boundary has a negative slope rather than a positive one. We assumed that each stimulus would have been shown to participants 10 times each, resulting in 360 trials for each categorization task.

The model was fitted to these category structures using a differential evolution algorithm, as implemented in the `DEoptim` package (Mullen, Ardia, Gil, Windover, & Cline, 2011) with the default strategy that minimized the negative log-likelihood. The parameter search was limited according to the upper limits given in Table 1. This was to encourage the parameter values to be similar to previous simulations (e.g. Love et al., 2004). We systematically varied the parameter c which controls the speed of crossover adaptation in the differential evolution algorithm to find the value that resulted in the best fit.

Predictions

To generate predictions, for each category structure, we used the best-fitting parameters to simulate the responses to 360 trials for 30 hypothetical participants. Each participant received trials in a random order. In Figure 2 we can see that SUSTAIN can

Table 1. SUSTAIN parameter values, fitted to the category structures.

Parameters	Upper Limit	Category structures			
		UD^{length}	$UD^{orientation}$	$II^{positive}$	$II^{negative}$
Attentional focus r	20	20.0	20.0	11.523890	0.000004467
Cluster competition β	20	20.0	20.0	20.0	20.0
Decision constancy d	20	20.0	15.09278	20.0	20.0
Learning rate η	1	0.08772490	0.12087377	0.08579475	0.08739468

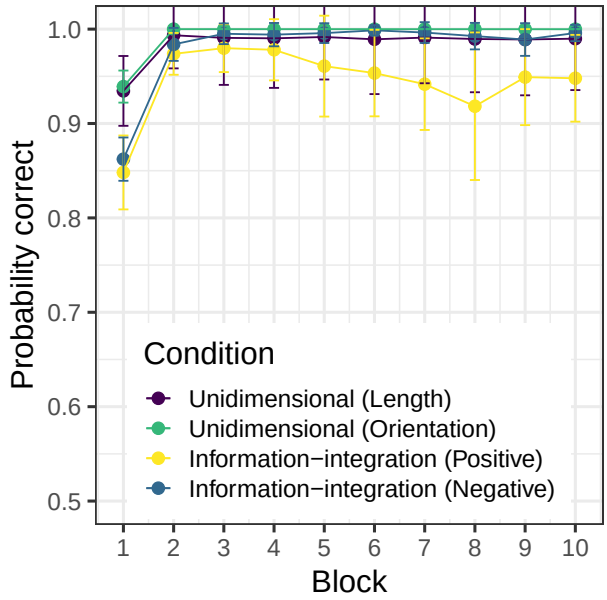


Figure 2. Learning curve graphs for the UD and II category structures predicted by SUSTAIN

easily learn the category structures. Note that these accuracies are much higher than are typically found in learning these category structures, especially for the II task. This is due to the lack of noise as we were fitting the category structure, not data from real participants.

In all conditions, SUSTAIN exhibits a variety of solutions. The average number of clusters in the II tasks are higher $M_{\text{Positive}} = 16.80$, $SD = 8.31$, $M_{\text{Negative}} = 5.77$, $SD = 0.89$, than those in the UD tasks, $M_{\text{Length}} = 3.30$, $SD = 6.56$, $M_{\text{Orientation}} = 2.23$, $SD = 0.50$.

In Figure 3 we show some examples of the clusters that SUSTAIN uses to learn these category structures. For the UD tasks, the simplest mapping from stimuli to clusters

are a cluster per category. The most complex mapping has four clusters to a category structure $UD^{orientation}$. On the other hand, SUSTAIN can present an unexpected, more exemplar-like, behaviour under some cases. For examnple, in the UD^{length} conditon, SUSTAIN recruits 2 clusters for 26 simulated participants, and 3 clusters for 3 other simulated participants. Simulation for one participant in the UD^{length} condition resulted in the model recruiting 38 clusters. This number is higher than the number of stimuli in our simulation, but organised along the grid-like arrangement of those stimuli.

SUSTAIN is highly sensitive to trial-order effects (Love et al., 2004) and this can result in some unexpected behaviour. SUSTAIN adjusts the position of the winning cluster, so the cluster will become more prototypical and drift away from stimuli that it already captured once. If the model doesn't encounter the stimulus for many trials, the stimuli will become more similar to other clusters. In that case, the model might be able to select the wrong cluster mapped to the wrong category, even though it already knows the right response. Then SUSTAIN would recruit a cluster in response to prediction errors. In response to prediction error, the modal solution would be to adjust the closest cluster mapped to the right category instead of recruiting a new category. We consider it to be the limitation of SUSTAIN, which is only present in edge cases.

For the II task, the simplest mapping involves four clusters per category. More complex clustering in this task becomes less ordered. Although it is definitely possible for SUSTAIN to learn both of these category structures, the II task is generally more complex.

Further, our Recognition Entropy measure predicts that overall mean recognition memory should be higher following II learning than UD, but the Recognition Score predicts more comparable recognition memory, see Table 2. Recognition Scores tend to overlap between UD and II conditions, see Figure 4. These scores can also become higher for small number of clusters than high number of clusters. These issues are not present with Recognition Entropy. Interestingly, both scores pedict very high recognition memory for the extreme case, when SUSTAIN recruited more clusters than stimuli.

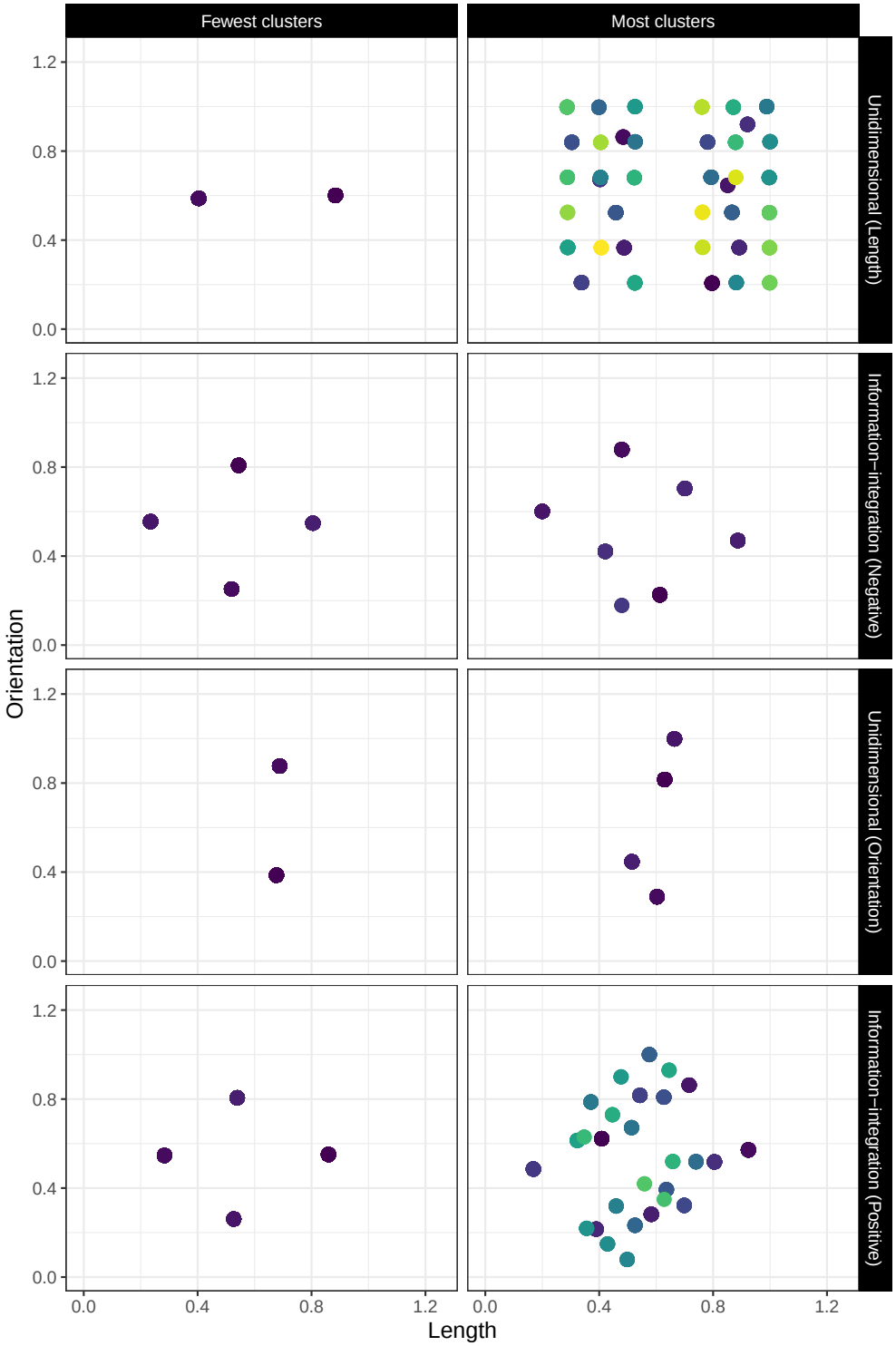


Figure 3. Example clusters predicted from SUSTAIN. Each coloured dot represents a different cluster.

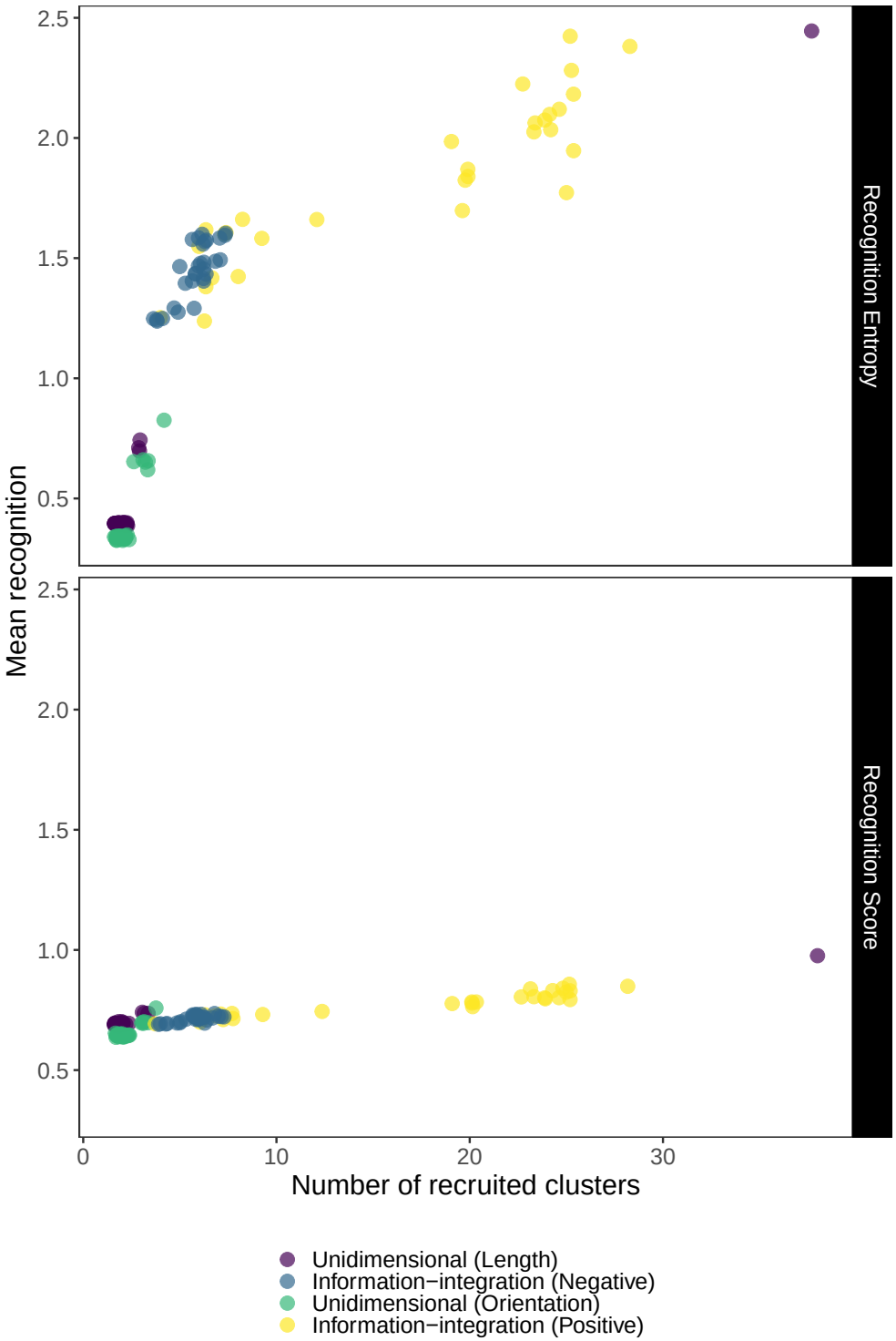


Figure 4. Example recognition performance for our Recognition Entropy and Recognition Score. Each coloured dot represents a mean recognition performance for a simulated participant.

Table 2. Mean and standard deviation of the 'Recognition Entropy' and 'Recognition Score' output from SUSTAIN

		Entropy		Score	
	Condition	Mean	SD	Mean	SD
UD	Length	0.40	0.14	0.71	0.05
	Orientation	0.50	0.38	0.66	0.03
II	Positive	1.83	0.33	0.77	0.05
	Negative	1.44	0.12	0.71	0.01

Conclusions from modelling

So, we can see here that SUSTAIN can learn information-integration category structures at least as well as researchers have found participants can in the literature. This suggests that, contrary to the intuitive predictions of those that promote the model, SUSTAIN can indeed learn the II task.

However, along with that comes an additional prediction: SUSTAIN and our Recognition Entropy measure predicts that there should be higher recognition memory for items following learning the II task compared to those learning the UD task. This is contrary to the predictions of those who most frequently use II tasks as they would predict that participants learn this task procedurally, and therefore, that there would be poorer recognition memory following II learning than UD learning.

In summary, SUSTAIN predicts that a) it is possible to learn II tasks explicitly and b) that recognition performance following that will be higher after II than UD tasks. Therefore, in the following experimental work, we looked to see whether participants a) could report the strategy they used in training and b) whether they had higher recognition performance following II learning than UD learning.

Experimental work

Here we report the combined estimates from four experiments. This was necessary given that the effect we expected to find was very small. This was due to a) the predicted

difference by SUSTAIN and b) because the stimuli in perceptual categorization experiments tend to be very similar and thus, very confusable. In these four experiments, we attempted to make the stimuli more distinct whilst still maintaining their consistency with the previous literature. This was not successful (see Appendices for results from individual experiments), hence combining the information from all particiants across the four experiments using hierarchical Bayesian techniques.

Participants

The participants were 155 undergraduate psychology students recruited from the Plymouth University participation pool. They were compensated with partial course credit for their participation.

Category structures and stimuli

The abstract category structures were identical in all these experiments. Half the participants were randomly assigned to learn a unidimensional rule-based category structure and the other half to learn an information-integration category structure (see Figure 1). The orientation of the category boundaries in abstract stimulus space were counterbalanced within conditions resulting in two unidimensional category structures–where the optimum boundary was vertical or horizontal–and two information-integration category structures–where the optimum boundary had either a positive or negative gradient.

These category structures were an adaption of the positive information-integration category structure used in Experiment 1 of Spiering and Ashby (2008). To create the positive information-integration category structure, we added an additional row of 6 stimuli that were perpendicular to the category boundary. This brought the total number of stimuli up to 36, which facilitated the random selection of a third of stimuli as "new" items for the recognition task. The remaining category structures were rotations ($\pi/4$, $\pi/2$, $3\pi/4$

radians) of this adapted structure such that 'centre of gravity' (i.e. the mean of both stimulus dimensions) of the points remained the same. The abstract stimuli coordinates were log-scaled so that all adjacent stimuli were approximately equally perceptually discriminable.

The stimuli varied between experiments, please see Appendices for detailed specifications. Experiment 1 used ; Experiment 2 used ; Experiment 3 used ; and Experiment 4 used.

Procedure

The experiment was split into four phases: category training, recognition test, category test and verbal report questionnaire.

Category training. First, participants were trained on two thirds of the available stimuli. These training stimuli were selected randomly for each participant subject to several constraints: 1) that those stimuli selected were symmetrical around the category boundary and 2) that no adjacent stimuli of similar difficulty were removed (for an example see Figure 1). In total there were 360 training trials, split into 3 blocks of 120 trials. In each block, 24 stimuli were each shown 5 times in a random order. On each trial, the participants looked at the stimulus until they made a response using either the "Z" key for Category A or the "/" key for Category B. Participants were unable to respond until at least 500ms had passed. Following their response, either "Correct" in green or "Incorrect!" in red was displayed for 500ms. A blank white screen was displayed between each trial for 500ms. Throughout the experiment, the labels "Category A" and "Category B" were displayed on the bottom left and right of the screen respectively. If participants took longer than 5000ms to respond, no corrective feedback was given, instead the message "PLEASE RESPOND FASTER" was displayed for 500ms.

Recognition test. Second, participants judged whether each stimulus was "old" and appeared in the training phase, by pressing the "O" key, or was "new" and had not been shown in the training phase, by pressing the "N" key. The words "New" and "Old" were presented on the bottom left and right of the screen respectively. After this, participants judged the confidence they had in their old-new judgement on a Likert scale that varied from 1 (=guessed) to 5 (=certain) by pressing the corresponding number key. Following previous work (Palmeri & Nosofsky, 1995), each of the 36 stimuli were presented three times in a randomised order. No feedback was given.

Category test. Third, participants were asked to judge the category membership of all 36 stimuli, not just those they had seen in the category training phase. No corrective feedback was given in this phase. Otherwise, the procedure was identical to that of the training phase. Each of the 36 stimuli were presented three times in a random order.

Analysis

All analyses were conducting using R (R Core Team, 2015). The trial-level raw data, verbal reports and analyses are available at `www.willslab.org.uk/pu037` with mds checksum `MDS checksum`.

Results

Three participants (two from the unidimensional condition, one from the information-integration) were removed because they failed to score over 50% on the final categorisation test leaving 75 and 77 participants in the unidimensional and information-integration conditions respectively.

Recognition performance was estimated using adjusted d-prime d_a, as the estimates of d-prime varied across confidence rating levels (Macmillan & Creelman, 2005). This is

calculated as follows

$$d_a = \left(\frac{2}{1+s^2}\right)^{\frac{1}{2}} [z(H) - sz(F)] \tag{2}$$

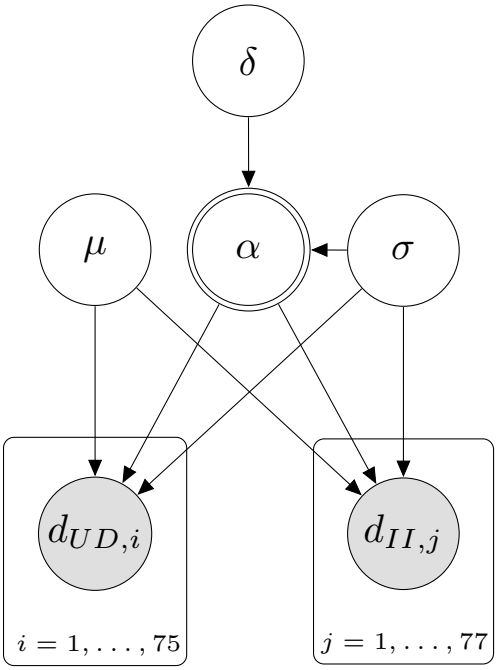
where s is the slope of the receiver operating characteristic (ROC) curve, H is the hit rate and F is the false alarm rate. We also calculated the adjusted bias rate c_a as follows

$$c_a = \left(\frac{-\sqrt{2}s}{(1+s^2)^{\frac{1}{2}}(1+s)}\right) [z(H) + z(F)] \tag{3}$$

To determine whether information-integration training resulted in superior recognition memory than rule-based training, we estimated a Bayes Factor using hierarchical Bayesian estimation techniques (Kruschke, 2015; Rouder, Haaf, & Vandekerckhove, 2018). A Bayesian approach is ideal here for two reasons. First, it allows us to combine data from several experiments in a principled way (Kruschke, 2015). Second, unlike null-hypothesis significance testing, it allows us to find evidence for the null (Dienes, 2011).

Our hierarchical model was specified as follows and is displayed graphically in Figure 5. For both category structure conditions, the data ($d_{UD,i}$ and $d_{II,j}$) were assumed to be drawn from group-level Normal distributions. Denoting the grand mean by μ, and the group difference in means by α and the group-level standard deviation by σ, the group-level Normal distribution for the unidimensional condition is defined as $N(\mu - \alpha/2, \sigma^2)$ and for the information-integration condition is defined as $N(\mu + \alpha/2, \sigma^2)$.

For the parameters that are not subject to statistical test (i.e., μ and *sigma*) we specified uninformative priors. The prior for the group mean μ was the standard Normal and for *sigma* was a uniform distribution between 0 and 5. The key aspect of our model is the parameter δ that quantifies effect size, $\delta = \alpha/\sigma$. This means that positive δ represents an effect of information-integration training resulting in greater memory than rule-based training, with negative δ indicating the opposite. We used the spike-and-slab prior on *delta* (Rouder et al., 2018). A spike-and-slab prior is a mixture of a null model (the *spike*) and



$$\theta_{UD,i} \sim Normal(\mu - \alpha/2, \sigma^2)$$

$$\theta_{II,j} \sim Normal(\mu + \alpha/2, \sigma^2)$$

$$\mu \sim Normal(0, 1)$$
$$\sigma \sim Uniform(0, 5)$$
$$\alpha = \delta * \sigma$$
$$\delta \sim M * slab + (1 - M) * slab$$
$$slab \sim Normal(0.3, 0.15)$$
$$spike \sim Normal(0, 1x10^{-5})$$

Figure 5. Graphical representation of the model used to estimate Bayes Factor.

an effect model (the *slab*). The spike was represented a Normal distribution, with mean 0 and a standard deviation of 1×10^{-3}. The slab was distributed as a Normal distribution, with mean 0.3 and standard deviation 0.15, representing a small effect size greater than zero (Dienes, 2011). The spike model was weighted as 0.9 and the slab model as 0.1, as previous runs indicated a bias towards sampling from the slab model (Kruschke, 2015).

We drew MCMC samples from the posterior distribution for δ, we ran 4 chains for 50000 iterations each, discarding the first 1000 trials as burn-in. We visually inspected the chains and calculated the Gelman and Rubin (1992) $\hat{R}$ statistic to check that the chains had converged. When the posterior was compared to the prior, this resulted in a Bayes Factor of $B_{10} = 12.02$. This provides positive support for better recognition performance post information-integration training than after rule-based training, compared to the null.

General Discussion

The SUSTAIN model has been successful in accounting for many categorization effects (Love et al., 2004). However, proponents of the SUSTAIN model put procedural

learning well outside its explanatory domain (Davis et al., 2012). Further, Davis et al. (2012) did not directly test whether or not this was the case. Therefore, in the current paper, we more carefully investigated whether the SUSTAIN model could account for the patterns of learning in a task argued to be learned procedurally (Ashby & Valentin, 2017).

By fitting the formal model to the category structure, we showed that SUSTAIN could learn an II category structure at least as well as found in the literature. We further used this model fitting to make predictions that could discriminate between SUSTAIN and a procedural learning account. We found superior recognition performance following the II task than the UD task. This is consistent with the predictions of the SUSTAIN model, but not consistent with the assumption that participants learning II category structures procedurally.

Resolving the conflict with the literature

So, if our results fail to find evidence of procedural learning during an II task, why are II tasks widely argued to be learned procedurally? Part of the reason may perhaps be because reviews of the literature can be biased (Wills et al., 2019). In actuality, there is a great deal of experimental and modelling literature that casts doubt on whether II tasks are learned procedurally.

First, many of the experiments that showed learning to affected in a way that was consistent with procedural learning, have been re-interpreted by other, subsequent work. For instance, in our lab we re-examined work that showed poor learning of II tasks with an observational procedure (Ashby et al., 2002). We found that performance in the II task was very similar to another, rule-based, two-dimensional categorization task. Therefore, it appears that the previous dissociation might be better understood as a dissociation between the number of relevant stimulus dimensions for categorization, rather than between a rule-based and procedural learning mechanism.

Further, participants learning II tasks are well able to describe the strategies they

used to complete the task. In several of our previous experiments, we have asked participants to describe the strategies they used to complete the learning task (Edmunds, Milton, & Wills, 2015; Edmunds, Wills, & Milton, 2018)[ANALOGICAL TRANSFER]. Participants who learned an II task tend to report using a two-dimensional, rule-based strategy rather than reports that would be consistent with procedural learning such as an overall similarity strategy or going with their gut.

Finally, another key assumption in this literature has shown to be flawed. Experiments using an II task often use a decision-bound strategy analysis to check that participants are learning the category structures optimally. In this approach, evidence that participants learned the task procedurally comes from the fact that participants appeared to use the optimum strategy for the task. However, this argument is rather circular: II tasks are learned procedurally, the participants learned the task using the optimum solution, therefore they learned the task procedurally.

Unfortunately, this approach has been shown to be unreliable (Donkin, Newell, Kalish, Dunn, & Nosofsky, 2015; Edmunds, Milton, & Wills, 2018). In decision-bound modelling, strategies are determined by fitting several different strategies to the responses from each participant. Then, a participant's strategy is argued to be the one that best fits. However, the evidence suggests that successful recovery of the correct strategy depends on whether you include the correct strategy in the set (Donkin et al., 2015; Edmunds, Milton, & Wills, 2018) and the category structure under consideration. Most importantly, the analysis is biased towards finding that participants used the optimal strategy for the category structure, no matter which strategy they actually used. Indeed, we were able to apparently recover that participants used the optimal strategies but were simulated as using rule-based strategies, at the same level of accuracy in published work (Edmunds, Milton, & Wills, 2018; Smith et al., 2014). This suggests that many of the participants that were argued to learn procedurally using this approach could have been using multi-dimensional rule-based strategies. Indeed, this would be more consistent with the

reports that participants have given (Edmunds et al., 2015; Edmunds, Wills, & Milton, 2018)[ANALOGICAL TRANSFER]. Thus, it seems like there is much evidence that suggests that II tasks are not learned procedurally.

Conclusion

Here, we investigated the predictions of the SUSTAIN model of category learning (Love et al., 2004). Contrary to a prior claim, we found that SUSTAIN can account for learning of II category structures. Further, we argue that this adds to evidence that II category structures are not learned procedurally.

Open practices statements

All data and analysis code is available in the Open Sciences Framework at `www.osf.com/`.

Author contributions

C. E. R. Edmunds ran the experiments, analyzed them, aided in the modelling and took the lead on writing the paper. Lenard Dome finalised the modelling and wrote the modelling section. Fraser Milton advised on experimental design and provided valuable insight on analysis and the paper. Andy J. Wills advised as his rule as Ph.D. supervisor of C. E. R. Edmunds and Lenard Dome.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481. doi: 10.1037/0033-295X.105.3.442

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*(1), 149–178. doi: 10.1146/annurev.psych.56.091103.070217

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*, 147–161.

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, *30*(5), 666–677. doi: 10.3758/BF03196423

Ashby, F. G., & O'Brien, J. B. (2007). The effects of positive versus negative feedback on and information-integration category learning. *Perception & Psychophysics*.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, *61*(6), 1178–1199. doi: 10.3758/BF03207622

Ashby, F. G., & Valentin, V. V. (2017, January). Multiple systems of perceptual category learning: Theory and cognitive tests. In *Handbook of categorization in cognitive science* (pp. 157–188). Elsevier.

Ashby, F. G., & Valentin, V. V. (2018). The Categorization Experiment: Experimental Design and Data Analysis. In E. J. Wagenmakers & J. T. Wixted (Eds.), *Stevens' handbook of experimental psychology and cognitive neuroscience, fourth edition, volume five: Methodology.* New York: Wiley.

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: 10.1093/cercor/bhr036

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290. doi: 10.1177/1745691611406920

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: 10.1037/xlm0000083

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures. *The Quarterly Journal of Experimental Psychology*, *68*(2), 1203–1222. doi: 10.1080/17470218.2014.978875

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: 10.1111/cogs.12607

Edmunds, C. E. R., & Wills, A. J. (2016). Modeling category learning using a dual-system approach: A simulation of Shepard, Hovland and Jenkins (1961) by COVIS. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 69–74.

Edmunds, C. E. R., Wills, A. J., & Milton, F. (2018). Initial training with difficult items does not facilitate category learning. *The Quarterly Journal of Experimental Psychology*. doi: 10.1080/17470218.2017.1370477

Gelman, A., & Rubin, D. B. (1992, January). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

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

Gureckis, T. M., & Love, B. C. (2002). Who says models can only do what you tell them? Unsupervised category learning data, fits, and predictions. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th annual conference of the cognitive science society* (pp. 399 – 404). Hillsdale, NJ: Erlbaum.

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: 10.1080/09528130210166097

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS and Stan* (2nd ed.). Academic Press.

Love, B. C., & Gureckis, T. M. (2007). Models in search of a brain. *Cognitive, Affective & Behavioral Neuroscience*, *7*(2), 90–108. doi: 10.3758/CABN.7.2.90

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: 10.1037/0033-295X.111.2.309

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide*. New Jersey, US: Lawrence Erlbaum Associates.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 650–662. doi: 10.1037/0278-7393.29.4.650

Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004). Evidence for a procedural-learning-based system in perceptual category learning. *Psychonomic*

Bulletin & Review, *11*(5), 945–952. doi: 10.3758/BF03196726

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. Retrieved from `http://www.jstatsoft.org/v40/i06/`

Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(3), 548–568. doi: 10.1037/0278-7393.21.3.548

Pothos, E. M., & Wills, A. J. (Eds.). (2011). *Formal approaches in categorization.* Cambridge, UK: Cambridge University Press.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018, February). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*(1), 102–114.

Smith, J. D., Boomer, J., Zakrzewski, A. C., Roeder, J. L., Church, B. a., & Ashby, F. G. (2014). Deferred feedback sharply dissociates implicit and explicit category learning. *Psychological Science*, *25*(2), 447–57. doi: 10.1177/0956797613509112

Smith, J. D., & Church, B. A. (2018). Dissociable learning processes in comparative psychology. *Psychonomic Bulletin and Review*, *25*(5), 1565–1584. doi: 10.3758/s13423-017-1353-1

Smith, J. D., Zakrzewski, A. C., Herberger, E. R., Boomer, J., Roeder, J. L., Ashby, F. G., & Church, B. A. (2015). The time course of explicit and implicit categorization. *Attention, Perception, & Psychophysics*, *77*(7), 2476–2490. doi: 10.3758/s13414-015-0933-2

Spiering, B. J., & Ashby, F. G. (2008). Initial training with difficult items facilitates information-integration but not rule-based category learning. *Psychological Science*, *19*(11), 1169–1177.

Wills, A. J., Dome, L., Edmunds, C., Honke, G., Inkster, A., Schlegelmilch, R., & Spicer, S. (2018). catlearn: Formal psychological models of categorization and learning [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=catlearn` (R package version 0.6)

Wills, A. J., Edmunds, C. E. R., Le Pelley, M. E., Milton, F., Newell, B. R., Dwyer, D. M., & Shanks, D. R. (2019). Dissociable learning processes, associative theory, and testimonial reviews: A comment on Smith and Church (2018). *Psychonomic Bulletin and Review*.

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.

Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin*, *138*, 102-125. doi: 10.1037/a0025715