Learning time-varying categories

Daniel J. Navarro
School of Psychology
University of Adelaide

Amy Perfors
School of Psychology
University of Adelaide

Wai Keen Vong
School of Psychology
University of Adelaide

Abstract

Many kinds of objects and events in our world have a strongly time dependent quality. However, most theories about concepts and categories are either insensitive to variation over time, or treat it as a nuisance factor that produces irrational order effects during learning. In this paper, we present two category learning experiments that explore peoples ability to learn categories whose structure is strongly time dependent. We suggest that order effects in categorization may in part reflect a sensitivity to changing environments, and that understanding dynamically changing concepts are an important part of developing a full account of human categorization.

At no two moments in time are we presented with the same world. Objects move, plants and animals are born and die, friends come and go, the sun rises and sets, and so on. More abstractly, while some of the rules that describe our world – like physical laws – are invariant over the course of our everyday experience, others – like legal rules – are not. Given some appropriate time scale, certain characteristics of an entity or class of entities can change; moreover, they may tend to change in *systematic* ways. For instance, the features that describe phones have changed considerably over recent decades: not only do modern phones perform many new functions, they are also physically smaller, sleeker, and smoother. Not surprisingly, people's expectations about category members change to

---

Correspondence concerning this article should be addressed to Daniel Navarro, School of Psychology, University of Adelaide SA 5005, Australia (daniel.navarro@adelaide.edu.au). Portions of this work were presented at the 2009 and 2012 annual conferences of the Cognitive Science society. Early work on this project, including salary support for DJN, was supported through ARC grant DP0773974. Later work was supported through ARC grant DP110104949, with salary support for DJN provided by ARC grant FT110100431 and for AP through ARC grant DE120102378. We thank Natalie May and Yiyun Shou for help running the experiments.

suit the environment as it stands: if asked to describe a phone in 2012, few people would refer to a rotary dial, but in 1970 nearly everyone would.

In one sense there is nothing terribly surprising about this observation. However, the changeable nature of many of the concepts and categories with which humans must interact is not greatly emphasized in the categorization literature (but see Elliott & Anderson, 1995). In most category learning experiments it is generally assumed that the underlying category is more or less static, and as such the order in which one encounters category members should not matter to a rational learner. In statistics, this is referred to as the assumption of *exchangeability*, and for reasons of simplicity it is generally the default assumption. Current probabilistic models of categorization make this assumption quite explicitly (e.g., Griffiths, Sanborn, Canini, & Navarro, 2008; Sanborn, Griffiths, & Navarro, 2010), and to the extent that standard exemplar and prototype models of categorization can be viewed as a kind of probabilistic model they can also be seen to abide by this assumption (Ashby & Alfonso-Reese, 1995; Griffiths et al., 2008).

Perhaps because exchangeability is assumed in most real world data analysis, it is generally taken to be a normative standard. However, human learners are also sensitive to the order in which category members are observed; this sensitivity appears to violate this normative standard. One way to account for order effects in cognitive models is by using learning rules that are sensitive to them. Some such rules can be viewed as modifications to standard probabilistic models. For instance, highlighting effects can be accounted for by assuming that people follow a "locally Bayesian" learning rule (Kruschke, 2006), whereas primacy effects can be captured by using a particle filtering learning rule (Sanborn et al., 2010). Another approach is to adopt connectionist error-driven learning rules, which implicitly assume that recent items are more salient and so are able to capture some kinds of recency effects, often better than a simple recency-weighting strategy would (e.g., Nosofsky, Kruschke, & McKinley, 1992; Sakamoto, Jones, & Love, 2008). A third approach is to alter the underlying stimulus representation: for instance, certain recency effects can be captured by the assumption that people track the *differences* between successive observations (Stewart, Brown, & Chater, 2002).

Although all of these approaches endeavor to account for order effects, it remains difficult to say whether it is rational to be sensitive to stimulus order. Many papers avoid any explicit discussion of whether this sensitivity should be called normative (e.g., Stewart et al., 2002), others argue that it reflects the cognitive limitations of the human learner (Sakamoto et al., 2008; Sanborn et al., 2010), and still others suggest that order sensitive learning rules are necessary if the learner is able to adapt to a changing world (Nosofsky et al., 1992; Elliott & Anderson, 1995). The latter perspective is mirrored rather explicitly in the literature on sequential effects (Yu & Cohen, 2009) and change detection (Brown & Steyvers, 2009).

Regardless of their views on the rationality of order effects, there is a great deal of uniformity in across all of these literatures: in particular, the models that best capture human performance weight the observations by their recency. In most models the weight assigned to a particular observation tends to decay approximately exponentially as a function of age (Nosofsky et al., 1992; Yu & Cohen, 2009; Brown & Steyvers, 2009), though in some cases the decay has been proposed to be a power function (Elliott & Anderson, 1995), which would be more in keeping with the literature on memory (Wixted & Ebbesen, 1997;

Rubin, Hinton, & Wenzel, 1999). A similar outcome exists in the judgmental forecasting literature, which examines how people perceive and extrapolate time series data: once again, an exponential weighting rule appears to account for human performance (see Goodwin & Wright, 1994; Lawrence, Goodwin, O'Connor, & Önkal, 2006, for overviews). An exponential weighting scheme also emerges from the literature on stimulus generalization (Shepard, 1987).

The combined weight of this work suggests that weighting more recent items according to an exponential function is both theoretically and empirically justifiable. In fact, if the world changes at unpredictable times and in arbitrary ways, an exponential weighting scheme is close to optimal (e.g., Yu & Cohen, 2009). However, the world does not always change in an unpredictable fashion. For instance, the changes to the category of "phone" have been at least partially predictable: newer phones tend to be faster, smaller, and more technologically capable. In this example at least, it is clear that people have a strong expectation that the category will change systematically and in a particular direction. This ability to extrapolate the direction of future change cannot be explained by assuming that sensitivity to order emerges simply from weighting more recent items more highly. A recency explanation would correctly predict that, for example, the iPad 4 would be more similar to the iPad 3 than to the iPad 2. What it does not predict, however, is that people expect iPad 4 to be systematically different from (e.g., faster than) both. In order to capture this effect, we need to move beyond simple recency towards an explanation based on the idea that people detect trends and extend them.

The goal of this paper is to address these open questions by investigating how people learn categories that change in a systematic fashion over time. We begin by presenting an experiment that involves a simple "linear change" pattern, using a standard supervised classification task. We then follow this up with an experimental design that requires participants to generate new category members. In both experiments we find evidence that participants are sensitive to the systematic pattern of change in the observations they are shown.

Experiment 1

In this section we present a category learning experiment in which people were presented with fairly obvious and systematic temporal changes, and investigate how well people learned to anticipate those changes.

Method

Participants. 59 participants were recruited through a mailing list whose members consist primarily of current and former undergraduate psychology students. They were paid $10/hour for their time. The median age was 23, and the participants were predominantly (63%) female.

Materials & Procedure. The learning task was a standard supervised classification experiment, performed on a computer. Stimuli were little cartoon objects ("floaters"), which were displayed floating above a horizontal line ("the ground"). The height of the floater was the only respect in which the stimuli varied from each other. An example of what a floater looked like is shown in Figure 1.

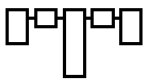

Figure 1. An example of the "floater" stimuli used in Experiment 1.

On each of 100 trials participants were shown a single floater and asked to predict whether it would flash red or blue. After making their prediction, they would receive feedback for two seconds while the floater flashed the appropriate color. As the left panel of Figure 2 illustrates, as the experiment progressed, all the stimuli shown to people tended to rise, regardless of which category they belonged to. In the figure, black circles correspond to items that belong to the "high" category and white circles correspond to times that belong to the "low" category. Assignment of flash color (red or blue) to category (high or low) was randomized across participants.

For the purposes of our analysis, we refer to the average rise (approximately 2mm) as 1 unit. The reason for doing this is that it allows us to write the true classification rule in a very simple form. Specifically, the classification rule was such that, if x_t denotes the height of the stimulus on trial t, then the optimal response is to select the response option corresponding to the high category if $x_t > t$. Such a rule, shown as the solid line in Figure 2, achieves 100% accuracy on the task. However, because most stimuli tend to lie quite close to the classification boundary, the task is relatively difficult even though the general trend is clear. Consistent with this, participants during informal discussions indicated that they detected the upward trend early in the experiment, but still found the task to be quite challenging.

A model for the task

Our data analysis relies on a simple categorization model. The model is inspired by decision bound models (e.g. Ashby & Gott, 1988; Ashby & Lee, 1991; Ashby & Maddox, 1993), though unlike most decision bound models it is not explicitly derived from general recognition theory. In this section we describe the structure of the model, since it is central to our analysis. It is worth noting, however, that we tried a variety of other categorization models, and the qualitative pattern of results was not affected.[1]

Recall from Figure 2 that the stimuli are designed to be approximately normally distributed with mean μ_t, where the mean μ_t increases linearly over time. Moreover, if a particular stimulus lies above the mean (i.e., $x_t > \mu_t$) then it belongs to the high category, and otherwise it belongs to the low category. In other words, tracking the category boundary

[1]For instance, we also tried prototype models, exemplar models, and a range of possibilities in whcih the learner estimates a linear or nonlinear regression function to describe how the category changes over time. Most of these models are outlined in Navarro and Perfors (2012). The key point is that the substantive story is robust: as long as we use a model that can fit the empirical data (not all of them can) we find a systematic but weak effect in the hypothesized direction.

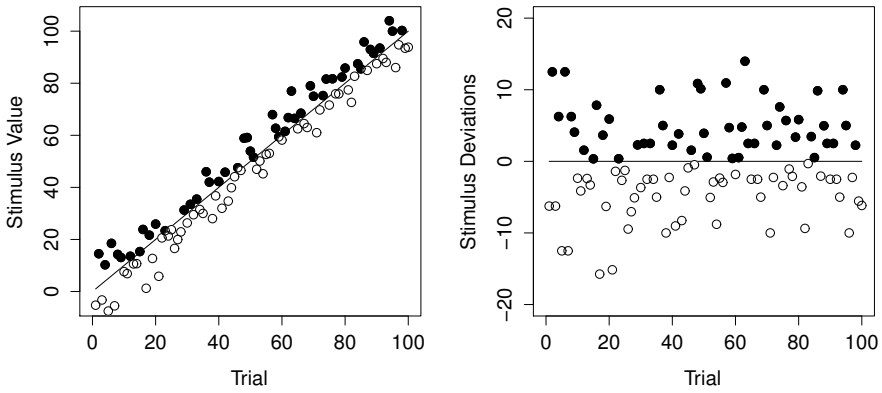

Figure 2. The experimental design. Black circles denote stimuli belonging to the "high" category and white circles denote stimuli belonging to the "low" category. Although the classification rule changes over time – that is, the classification boundary is constantly rising, as shown on the left panel – it does so in a regular fashion. The scale on the vertical axis is normalized so that the average "rise" from one trial to the next is 1 unit; on screen, this corresponded to an average rise of approximately 2mm per trial. Note that although it is logically possible to correctly classify all items, in practice the task is quite difficult, since most stimuli lie close to the boundary. The deviations of each stimulus from the true classification boundary are shown on the right.

over time is equivalent to tracking the value of μ_t over trials. A simple model for estimating the value of μ_t is as follows. Suppose that the learner has some estimate $\hat{\mu}_{t-1}$ of the location of the category boundary before the start of trial $t-1$. When the stimulus x_{t-1} is observed, the boundary is shifted by some proportion w in the direction of that observation, yielding the following estimate for the location of the category boundary before the start of trial t

$$\hat{\mu}_t = (1-w)\hat{\mu}_{t-1} + wx_{t-1} \tag{1}$$

where w is a "twitchiness" parameter that indicates the extent to which the learner relies on the very last observation that he or she has seen.[2] Expanding the recursion in Equation 1, we observe that this model produces an estimate $\hat{\mu}_t$ that is an exponentially-weighted

[2]It may not be obvious on the surface how the model incorporates feedback, since Equation 1 does not explicitly specify a training signal. To illustrate how feedback is used, consider an error-driven learning model that adjusts its estimate of the category boundary every time the feedback indicates that an error was made. To learn from this feedback, the model needs a training signal. The true error here is $\mu_{t-1} - \hat{\mu}_{t-1}$, the deviation between the true and estimated category boundary, but this is not observable to the learner. However, the fact that x_{t-1} was misclassified implies that the magnitude of the error is *at least* $x_{t-1} - \hat{\mu}_{t-1}$. This lower bound on the error is observable and can be used as a training signal. If the learner adjusts the estimated category boundary by some proportion w of this error, we obtain $\hat{\mu}_t = \hat{\mu}_{t-1} + w(x_{t-1} - \hat{\mu}_{t-1})$, which is identical to Equation 1. In other words, if Equation 1 is applied only on error trials, it is equivalent to a standard error-driven learning model. We did consider using this restricted model, but felt that the full model (which learns from errors and from correct decisions) would be simpler.

average of the previous items:

$$\hat{\mu}_t = w \sum_{k=1}^{t} (1-w)^{k-1} x_{t-k} \tag{2}$$

In this equation the fictitious 'zero-th stimulus' x_0 corresponds to an initial value for the category boundary. The key thing to note in this equation is that recent trials will contribute more heavily to the estimate of μ_t; large values of w imply that only a few observations are used, small values of w allow multiple observations to be used.

Having formed an estimate of where the category boundary lies, the learner is assumed to generate responses probabilistically, as a function of the deviation Δ_t between the current stimulus and the estimated category boundary. This deviation is given by

$$\Delta_t = x_t - \hat{\mu}_t \tag{3}$$

and if $\Delta_t > 0$ then the item is more likely to be classified as a member of the high category. Specifically, we assume that the function relating distance to categorization probability is logistic (e.g., Navarro, 2007). If p_t denotes the probability of selecting the "high" category on trial t, then

$$p_t = \frac{1}{1 + \exp(-\lambda \Delta_t)} \tag{4}$$

where λ governs the rate at which the classification probability changes as a function of distance from the category boundary.

This model is closely related to the tracking model used by Brown and Steyvers (2009), but it has links to other models too. For instance, in the extreme case where $w = 1$, this heuristic corresponds to a relative judgment strategy in which each stimulus is compared only to the last stimulus in the experiment, and is a slight simplification of the model used by Stewart et al. (2002). When $w = 1$ and λ is large, the model produces a very simple heuristic: select the high category if and only if the current stimulus is larger than the previous one. Additionally, the model has links to prototype models of classification. In an equal variance prototype model, the learner represents separate means (the prototypes) for each category, and the category boundary lies equidistant from the two prototypes. If the estimate for a category prototype takes recency into account by taking an exponentially weighted average (see, e.g. Navarro & Perfors, 2012), then the classification probabilities will end up almost identical to those produced by our model.

The problem with using a heuristic such as this one is that it is very sensitive to random fluctuations in the data (when w is large), or else the estimate of the category boundary lags a long way behind the true one (when w is small). To see this, note that when w is large, then the learner is strongly influenced by the most recent observation, and to the extent that this observation is noisy or otherwise misleading as to the location of the category boundary, the learner will be unduly influenced by it. Decreasing the value of w allows the learner to avoid this mistake by aggregating information from multiple observations, but this comes at a price: because the category is moving, setting w to too small a value means that the estimate $\hat{\mu}_t$ will always be 'lagging' a long way behind the data.

This issue can be avoided to some extent if the learner is able to detect the pattern of change and anticipate the fact that the category boundary μ_t moves on each trial. For the sake of simplicity, we assume that this corresponds to the introduction of a simple correction factor b. This correction factor yields a slight modification to the model, in which the estimate of the category boundary on trial t is given by

$$\hat{\mu}_t = b + w \sum_{k=1}^{t} (1-w)^{k-1} x_{t-k} \tag{5}$$

We now have a simple classification model with three parameters: the twitchiness w that governs how reliant the learner is on the last stimulus, the slope parameter λ that describes the relationship between distance and generalization, and the bias parameter b that describes a correction factor, shifting the classification boundary to accommodate the fact that the task involves a clear trend over time. Although this model has not (to our knowledge) been used in the categorization literature previously, it has been used in the judgmental forecasting literature (see Goodwin & Wright, 1994) as a heuristic that is appropriate for modeling human extrapolation judgments for a trended time series.

Results

Human performance was significantly above chance for both categories: 76% of the "high" category items and 61% of the "low" category items were classified correctly, as shown in the left panel of Figure 3. Note that, while performance in both categories is significantly above chance, people perform better in the high category. This is not surprising: because both categories are rising throughout the experiment, novel items from the low category tend to be close to previous high category items. By contrast, new items from the high category are much less confusable, since they are not close to low category items.

When we plot the performance of all 59 participants separately, as in the right panel of Figure 3, it is clear that the improved performance on the high category items holds at the individual subject level as well. Nine of the 59 participants were not significantly above chance (the $p < 0.05$ significance threshold is plotted as a dashed line). Of the remaining 50 participants, 47 classified the high category items more accurately than the low category items, while performing above 50% correct for both categories. This is illustrated visually by the fact that the vast majority of dots in Figure 3 fall within the solid black triangle.

Model fits were performed at the individual subject level, obtaining separate values of b, w and λ for all 59 subjects, using maximum likelihood estimation to estimate parameter values. The first 5 trials were excluded for the purposes for the model fitting.[3] As shown in Figure 4 the model produces very similar classification behavior to human participants. Across participants, the correlation between human data and model predictions for the probability of correctly classifying low category items was $r = 0.94$ ($p < .001$), and $r = 0.92$ ($p < .001$) for the high category items. At a within-participant level, the model fit is significantly better than chance (as assessed via a likelihood ratio test at the $p < .05$ level)

[3]This decision was motivated by the intuition that the first few trials are "special" and not easily modeled. The intuition was also backed up by additional analyses using generalized linear mixed models: including a random effect term for trial number revealed that this is true for trial 2 in particular. Given this, it seemed prudent to exclude the first few trials.

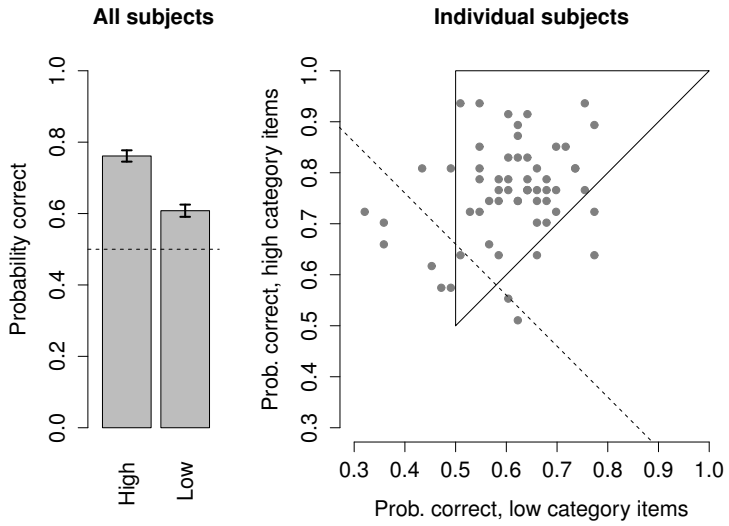

Figure 3. Overall performance (left) with 95% confidence intervals, as well as individual participant data (right). Both plots incorporate data from all participants and all trials. Note that 8 participants do not perform significantly above chance (dashed line corresponds to the $p < 0.05$ significance threshold), and for this reason are excluded from subsequent analyses.

Table 1: Means and standard deviations for model parameters, taken across all 43 subjects for which the model fit was deemed adequate.

	mean	std dev
bias, b	1.07	2.58
twitchiness, w	0.27	0.11
slope, λ	0.56	0.30

for 52 of the 59 subjects. Of the 7 subjects whose data were not well-fit by the model, 5 were among those who did not classify the stimuli above chance. It is no surprise that the model cannot account for the performance of those participants, as the empirical data from those participants is extremely noisy.

The key empirical test here relates to the parameter estimates, most notably of the bias parameter b. Given this, we restrict the analysis to those participants whose data are well-fit by the model. Our inclusion criterion here was that the model needed to explain at least 25% of the variance in the participant's choices. This criterion was met for 43 subjects. Among these participants, the model explained 43% of the variance on average (sd = 12%), corresponding to a 71% probability that the model would make the same response as the participant on any given trial (sd = 6%).

Descriptive statistics for the parameter estimates are provided in Table 1. From a theoretical perspective, the first analysis to consider is a test of whether the b parameter is in fact necessary within the model. To that end, for all 43 participants, we fit a restricted model with the bias parameter fixed at $b = 0$; a likelihood ratio test rejected this null model

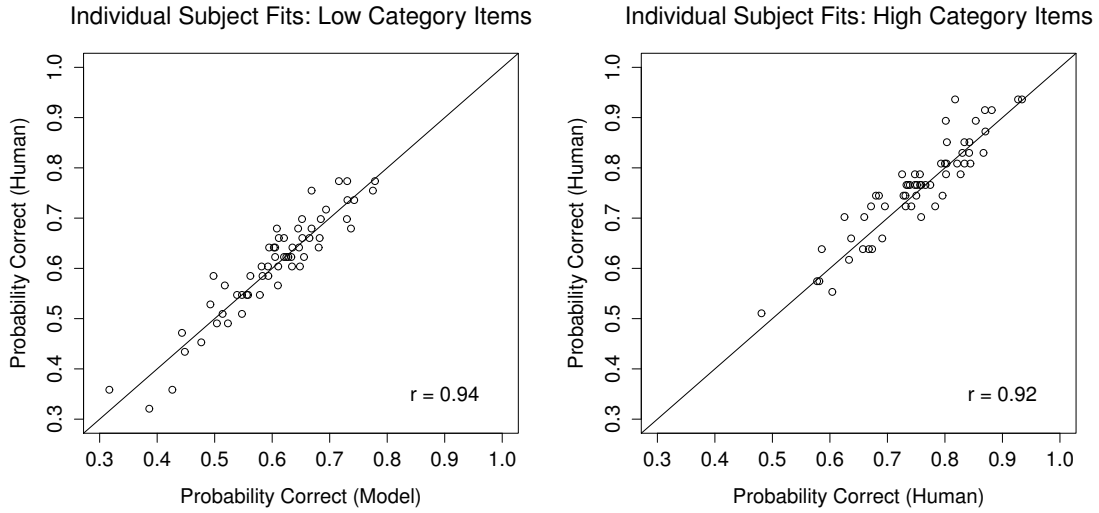

Figure 4. Model performance at an individual subject level. As with Figure 3 each participant is characterized in terms of their probability of correctly classifying low category items (left panel) and high category items (right panel). Each panel plots each participant as a circle, showing the model probability correct against the human probability correct.

(at $p < .05$) in all 43 cases, favoring the model that includes the bias. A second analysis to consider is to look at the magnitude of the bias parameter. The mean value of b was 1.07, with a 95% confidence interval of [0.28, 1.87], implying that participants did shift their classification boundaries upwards.

Closer examination reveals a somewhat more complicated story. There is a moderately strong negative correlation of $r = -.77$ ($p < .001$) between the bias parameter b and the twitchiness parameter w (this is the only significant correlation). This is to be expected: as noted earlier, smaller values of w mean that participants are aggregating information across more trials, which in turn implies that the uncorrected estimate of the category boundary will lag further behind the true boundary. In other words, the optimal value of b should be higher for smaller values of w. This is illustrated in Figure 5, which plots b against w for all participants (black dots), along with a regression line (solid line) that depicts the relationship between the two. The fact that confidence bands for the regression line (grey region) sit above zero (dotted line) for most values of w indicates that, in general, participants were extrapolating.

The fact that people extrapolate does not imply that the extent of the extrapolation is optimal. Indeed, Figure 5 also shows the optimal value of b for all values of w (dashed line). It is clear that the participant responses lie below the optimal value, indicating that they did not extrapolate far enough. This is consistent with the raw data in Figure 3: the simple fact that people are more accurate for high category items than for low category items provides strong evidence that people do not entirely anticipate the extent of the changes across trials. Moreover, when we plot the mean category boundary (taken across participants) extracted from the model, as per Figure 6, it is clear that the model reproduces this behavior.

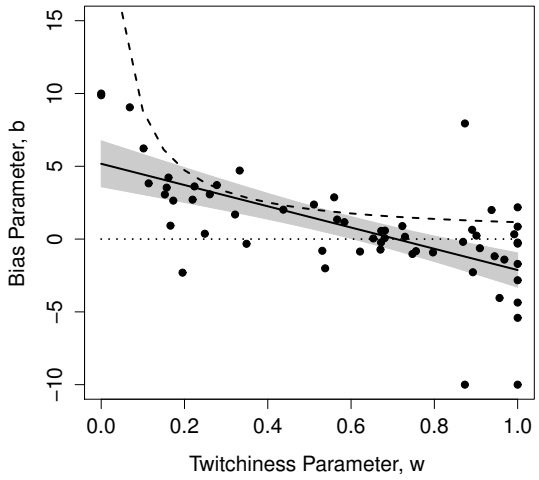

Figure 5. A scatterplot (black dots) showing the relation twitchiness parameter w and the bias parameter b, along with the linear regression (solid line) and corresponding 95% confidence bands. The fact that the regression line sits above zero (dotted line) for most values of w indicates that, in general, participants were extrapolating. However, the regression line also lies below the optimal value of b (dashed line) for most values of w, indicating that participants did not extrapolate far enough.

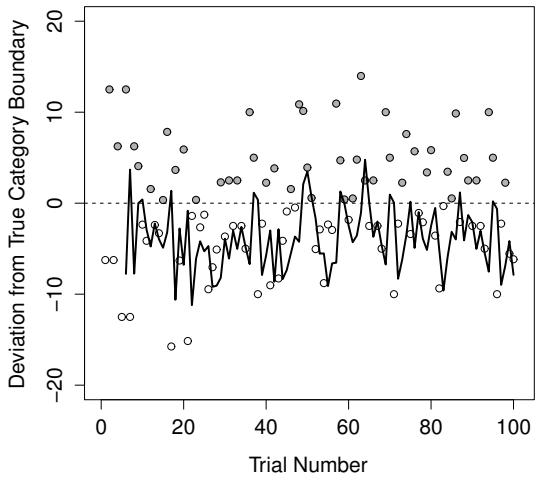

Figure 6. Estimated category boundary on each trial (solid line), expressed in terms of the deviation from the true category boundary (dashed line), and averaged across all 43 participants whose data is well-fit by the model. The fact that the solid line lies below the dashed line most of the time indicates that participants' implied category boundaries still lagged behind the true rule.

Discussion

The results from Experiment 1 provide some indication that people are capable of detecting changes to a category over time, and are able extrapolate a trend that they have observed when making decisions about new items. However, the effect is subtle, and relies on the assumption that, if the categories had not been changing over time, participants would have used an recency-weighted average to estimate the category representation (i.e., the model without b). This assumption seems sensible insofar as the task was a standard supervised categorization task, and models of that kind have been highly successful in explaining human behavior in such tasks. Nevertheless, it is desirable to show that the effect can be detected without requiring detailed modeling, if we allow the task to be modified to capture the effect more directly.

Experiment 2

One of the problems with the supervised classification design in Experiment 1 is the fact that participants provide only a very limited amount of data (a binary choice) on any particular trial. This makes it a little difficult to measure very subtle effects in how the category representation changes at a trial-by-trial level. To redress this, Experiment 2 employed a very different dependent measure: in this experiment, participants were asked to generate new category members on each trial. After training participants on a category that changes over time, we asked them to generate a sequence of future category members over multiple time points. If participants are genuinely able to extrapolate their category knowledge, then this sequence of responses should extend the trend that appears in the training data.[4]

Method

Participants. 110 participants (34 female, 76 male) were recruited via Amazon Mechanical Turk, an online service which co-ordinates people to complete tasks requiring human intelligence. Participants were located in 17 different countries, but the vast majority were in India (71) or the United States (17). They were required to be English speakers (assessed by a series of simple test questions) and were paid US$0.25 for their participation in the study. The structure of the experiment allowed a fairly straightforward way to check if participants had understood the task and were making a genuine attempt to complete it, based on an examination of their responses. Exclusions are thus discussed together with the data analysis.

Materials & Procedure

Participants completed the study online through the Amazon Mechanical Turk website. After accepting the task, they were asked demographic questions and then presented with the experiment instructions. They were then quizzed on the experimental procedure, to check that they understand the experiment instructions (this implicitly served as a check that they understood English). If they did not answer all questions correctly, they were redirected back to the instructions and required to repeat the instruction checks until all

[4]This experiment formed part of a larger study that made up the third authors Honours thesis.

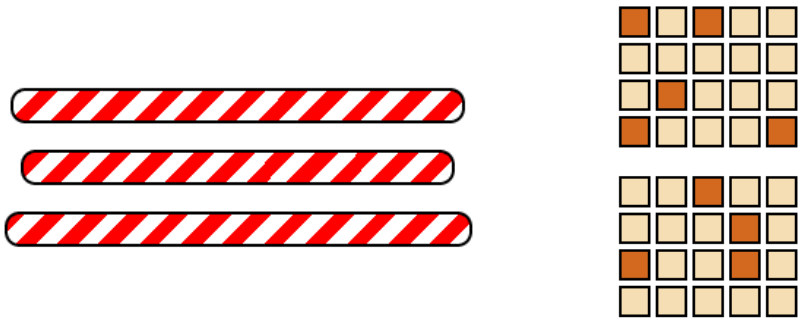

Figure 7. Examples of the stimuli used in Experiment 2. The image on the left shows a display containing 3 candy canes, and the image on the right shows a display containing 2 chocolate boxes. In both cases, the actual displays contained 10 stimuli.

questions were answered correctly. In order to verify that any effect observed does not depend on particular details of the stimulus or task, four versions of the task were employed. Participants were randomly allocated to different groups for the *stimulus type* and *prediction type* factors, yielding a 2×2 design. In the stimulus type manipulation, we altered the surface representation of the stimuli. The prediction type manipulation focused on the kind of responses people were asked to provide.

The cover story for the task asked participants to imagine they were helping a confectionery company to direct the future direction of some of their products. The experiment involved two phases for all conditions. Phase one consisted of 10 trials. Each trial corresponded to a different year: on each trial participants were shown a set of 10 stimuli, intended to represent different examples of confectionary on sale that year. All 10 items were displayed on screen together. The visual representation of the confectionary depended on the stimulus type: in the *candy cane* condition, stimuli were displayed as striped horizontal bars of different lengths. The length of the bar was the relevant stimulus dimension. In the *chocolate* condition, stimuli were boxes of chocolates, where each chocolate could be either white or dark. Each box showed a 4×5 grid of 20 chocolates, and the number of dark chocolates was the relevant stimulus dimension. Example stimuli are shown in Figure 7.

As with Experiment 1, the stimuli changed systematically across trials, as illustrated by the dashed lines in Figure 8. In the candy cane condition, the mean length of the stimuli was 300 pixels on trial 1, which rose linearly to 525 pixels by trial 10, corresponding to an average increase in length of 25 pixels per trial. In the chocolates condition, the mean number of dark chocolates in the boxes started at 3 on trial 1, rising by 1 chocolate per trial until the average number of chocolates on trial 10 was 12. On any given trial, however, participants were shown 10 stimuli that varied around this mean. In the candy cane condition, the length of any one stimulus was uniformly sampled from a range of ± 25 pixels around the true mean. In the chocolates condition, the number of dark chocolates in any one box was equal to the true mean plus -1, 0 or 1, with each outcome being equally likely. In our data analysis, however, all stimulus magnitudes are rescaled so that the mean magnitude on trial 1 was 0, and the mean magnitude on trial 10 was 1.

After being presented with a set of stimuli, people were asked to generate three new

examples of confectionary. The task differed depending on the prediction type: in the *current* condition, participants were asked to generate three more examples that would be representative of the confectionary for the current year; in the *future* condition the examples were supposed to represent predictions for the next years confectionary. Participants who saw candy canes made their predictions by dragging a slider bar to change the length of the candy canes, while participants who saw chocolate boxes were able to click on each individual chocolate piece to change its color from white chocolate to dark chocolate or vice versa.

The critical part of the experiment was phase 2, which was identical regardless of what prediction type participants were assigned to. During this phase, participants were asked to generate three examples of candy canes or chocolate boxes for each of the next three years (i.e., generating nine examples in total). During this phase no information was given as to what kinds of candies or chocolate blocks were popular for those years. Note that, because these unsupervised trials followed immediately from the supervised ones in phase 1, we did not ask participants in the future-prediction condition to make any responses for the last trial of phase 1. The reason for this is that the last trial of phase 1 is essentially equivalent to the first trial of phase 2 in the future-prediction condition, and we felt it more important to preserve the comparability of judgments in phase 2.

Results

Because the stimuli increased across trials in a linear fashion, one would expect that the predictions generated by participants would also rise across trials. At a minimum, one would expect the magnitude of the responses generated by participants in phase 1 to correlate with the stimulus magnitude. This suggests a natural exclusion criterion: if the Pearson correlation between the stimuli and responses is less than or equal to zero, we have evidence that the participant either did not understand the task, or failed to make a genuine attempt to do so. In total 16 participants were excluded on this basis, including two participants who produced identical responses on every trial. After these exclusions were made, we ended up with 24 participants in the candy-stimulus current-prediction condition, 29 in the chocolate-stimulus current-prediction, 34 in the candy-stimulus future-prediction condition, and 23 in the chocolate-stimulus future-prediction condition. The mean response on each supervised trial is plotted in Figure 8 broken down by condition (error bars denote 95% confidence intervals). There is some evidence in these plots that participant responses differed between conditions during phase 1, but the important thing for our purposes is to note that in all conditions, the participants understood the task, correctly realizing that the stimulus magnitude was increasing over trials. Given that the primary interest lies in how participants responded in phase 2, it is to this topic that we now turn.

Mean responses and confidence intervals[5] for the phase 2 trials (i.e., extrapolations over a three year range) are plotted in Figure 9: on the left the data are shown aggregated across condition, and on the right each condition is shown separately. Given that the conditions involved different stimuli, response methods, and task instructions, it is not

[5]Confidence intervals in the right panel of Figure 9 are the intervals associated with each group mean, constructed in the usual way. Because the left panel collapses across groups with different means, the error bars in this case show the confidence interval associated with the grand mean in the corresponding 2x2 ANOVA (i.e., confidence interval for the intercept term when Helmert contrasts are used).

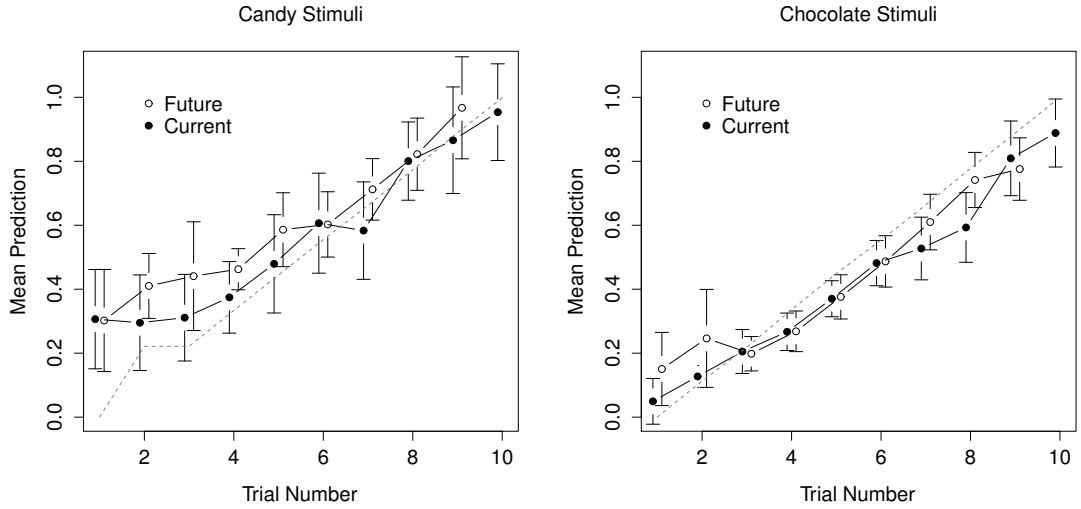

Figure 8. Mean responses on each trial during the supervised phase. In the current-prediction condition, participants at trial t generated predictions for additional examples at trial t. In the future-prediction condition, participants at trial t generated predictions for examples they would expect to see at trial $t + 1$. One would therefore expect that in most cases, the future predictions would be higher than the current predictions. This effect is evident for the candy-cane stimuli but not the chocolate stimuli.

entirely surprising to see that the baseline response in phase 2 (i.e., the response on the first trial of phase 2) differs across condition. The key question of interest with respect to phase 2, however, is not whether the four *conditions* differ, but whether the three *trials* do. That is, do the responses show a significant rise from trial to trial? Do people extend the linear trend that they observed in phase 1 into the future?

To address this question, we fit a linear mixed effect model to the phase 2 data using the using the `lme4` package in R (Bates, Maechler, & Bolker, 2011). The model included fixed effects of prediction type, stimulus type and trial number (i.e., year). Individual variation was captured by including a random intercept and random effect of trial number (i.e., random slope) for each subject. The Bayesian information criterion (BIC) was used for model checking, testing each effect by comparing the full model to a restricted model in which the terms corresponding to that effect removed. The BIC deteriorates very substantially if we remove the random intercepts (BIC difference = 100.0) or the random slopes (BIC difference = 14.3). Removing the fixed effect of prediction condition improves model performance slightly (BIC difference = -5.0), whereas removing the fixed effect of stimulus condition causes it to decline slightly (BIC difference = 6.1). Most importantly for our purposes, the effect of removing the fixed effect of trial causes a large deterioration (BIC difference = 22.4). If we convert the BIC difference of 22.4 to approximate posterior odds, we observe that this corresponds to odds of 73,000:1 in favor of the model that includes the effect of trial number. This suggests that participants are extrapolating sensibly over the next three years, capturing the linear trend in the data they are shown.

Having verified that participants' responses did show a significant rising trend in

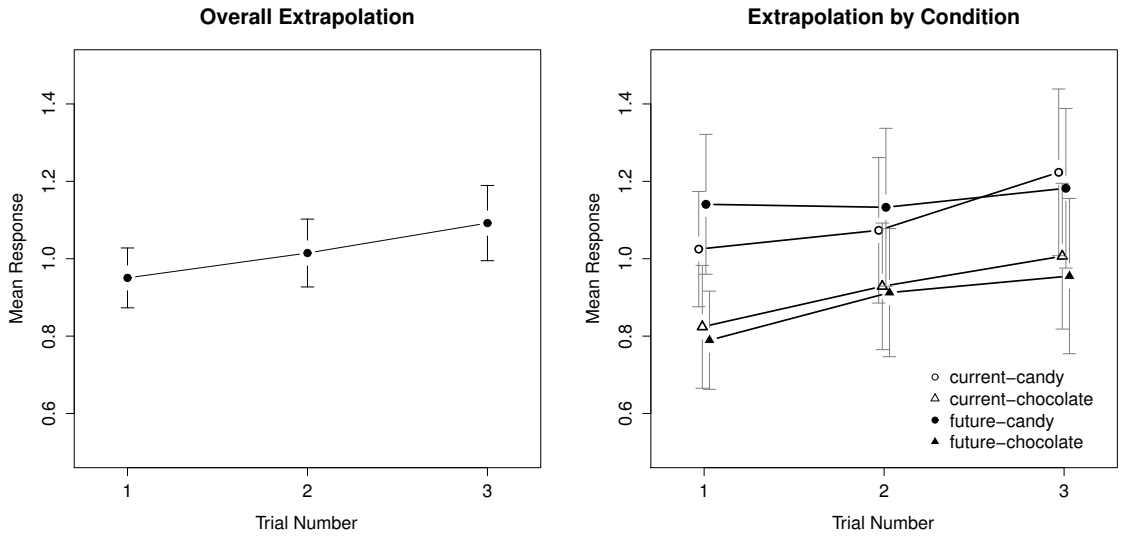

Figure 9. The data from phase 2 of experiment 2, showing future predictions made over three years without further examples shown. The plots display the mean responses and 95% confidence intervals for participant responses in phase 2 of the experiment. The left panel shows the mean response averaged across conditions, whereas the right panel shows the responses broken down across all four conditions. As this figure indicates, participants' generalizations in phase 2 reflected the linearly increasing trend that they experienced in phase 1 of the experiment.

phase 2 of the experiment, it is useful to consider how large that trend is. For simplicity, we consider the aggregated data plotted on the left hand side of Figure 9. The slope in this plot is 0.077. The linear trend in phase one involved the stimuli rising (in rescaled units) from an average value of 0 on trial 1 up to an average of 1 on trial 10. This corresponds to a rise of 1 unit in 9 steps; that is, a slope of 0.111. In other words, although participants did extrapolate a linear trend in phase 2, the slope of the trend was shallower than the slope of the original trend in the data, with a slope that is only 70% as large as the empirical slope in phase 1.

Discussion

The results of Experiment 2 are in agreement with the earlier study. In Experiment 1 we found evidence that people extrapolated their knowledge about categories in a manner that is consistent with the pattern of change in the task. However, the effect was subtle and relied on a detailed model fitting exercise. By asking people to generate new category members that they would expect to observe several years into the future, we are able to observe the effect more directly. As illustrated in Figure 9, across all four conditions participants' judgments showed a linearly rising trend when asked to make predictions for the next three years even in the absence of any explicit feedback. Similarly, just as with Experiment 1, we found that the extent of the extrapolation was suboptimal: people do the right thing qualitatively, but not to the extent required to perfectly match the trend in the raw data. Taken together, the two experiments tell a fairly consistent story about how

people make sense of categories that change over time.

General Discussion

The world is not a static place: the rules and regularities that characterize our environments do not remain constant over time, and as such a sensitivity to change is a necessary characteristic of any intelligent learner that operates in such an environment. As trivial as that sounds, it has important implications. Any model that is invariant to stimulus ordering, such as the original GCM (Nosofsky, 1984), standard decision bound models (Ashby & Lee, 1991), or the statistical model underpinning the "rational" model of categorization (Sanborn et al., 2010) is fundamentally the wrong model for some kinds of categorization problem that humans need to solve in real life, namely those that involve change over time. As such, models that are sensitive to order information such as the Generalized Context Model with recency weighting (Nosofsky et al., 1992) or models that rely on connectionist learning rules (e.g. Kruschke, 1992; Sakamoto et al., 2008) are better suited to this sort of category learning problem. Similarly, to the extent that learning about change is a key task facing the learner, directly encoding stimulus differences (as per Stewart et al., 2002) may be viewed as a rational thing to do. In some respects, these models are closer to the correct rational analysis of this learning problem than are most "rational" categorization models.[6]

This work shows that once we start to think of category learning as a problem that applies in a changing environment, there is much more to the problem than simply assigning more importance to recent observations. When the pattern of change is systematic rather than arbitrary, then it is not sufficient merely to *detect* changes (e.g., Brown & Steyvers, 2009): an ideal learner is able to anticipate them by forming sensible expectations about when the world changes and how that change occurs. Although our experiments are very simple examples of this, involving regular linear change in one dimension, they provide evidence that people do form these sorts of expectations during category learning.

The current research also opens up questions about the mechanisms by which people actually form these expectations. The model that we used in Experiment 1 does not provide an answer to this question. We do not provide any theory for how the bias parameter is learned, though it seems likely that since $b > 0$ produces superior classification performance in this task almost any sensible learning rule (e.g., error-driven learning, Bayesian updating) would infer a positive value. Indeed, it seems likely that such models would infer the *optimal* value for b, which would open up the question of why human extrapolations were suboptimal in both experiments. More generally, we are not convinced that a simple "bias" is the right answer to the problem in general: the model from Experiment 1 would be highly inappropriate for a category that changed in a nonlinear fashion, for instance. These questions are left open for future work.

Although the model from Experiment 1 was used primarily as an aid to data analysis, the bias parameter b within that model is perhaps the most explicit attempt to formalize the notion of extrapolation in this paper. As such, it is worth giving some consideration to what b actually does. Viewed purely as a mathematical device, the effect that b has

[6] In fact, although we do not discuss it in this paper, it is clear that models that lack a sensitivity to change will perform extremely poorly when applied to experiments such as the one in this paper. This issue is discussed in the conference paper that this article extends (Navarro & Perfors, 2012).

on the estimated category boundary is identical to the effect of a category base rate: it moves the boundary closer to one category or the other. Given this, one might wonder if the learned bias in Experiment 1 is just a form of base rate learning. However, this conflates the learning mechanism with the substance of what is learned: in the experiments presented in this paper, the base rates for both categories were always 50%. If the learned bias b is equivalent to base rate learning, then one would expect an optimal value of $b = 0$ in any experiment for which categories do not differ in base rate. Yet, as Figure 5 shows, the optimal value of b is decidedly non-zero. Even if the mechanism used to learn b is formally equivalent to a base rate learning rule, the thing that is learned is clearly not a base rate. Rather, it is a bias term whose substantive effect is to systematically shift the learners expectations about future events away from their experience of past events. "Extrapolation" seems as good a name as any for this inference.

To sum up, this paper illustrates that people are capable of tracking changes over time during category learning. This work also opens up a range of additional issues. Regular linear change is only one pattern of change that could characterize real world categories. In our initial work on this topic (Navarro & Perfors, 2009) we also considered the possibility that changes could be discrete jumps similar to Brown and Steyvers (2009) or sinusoidal patterns similar to Sakamoto et al. (2008). However, these possibilities are not exhaustive either. Understanding how people adapt to systematic changes in categories requires a broader investigation across a wider range of possible change patterns.

References

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216-233.

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*, 33-53.

Ashby, F. G., & Lee, W. W. (1991). Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General*, *120*, 150-172.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*, 372-400.

Bates, D., Maechler, M., & Bolker, B. (2011). lme4: Linear mixed-effects models using s4 classes [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=lme4` (R package version 0.999375-42)

Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, *58*, 49-67.

Elliott, S., & Anderson, J. R. (1995). Effect of memory decay on predictions from changing categories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*, 815-836.

Goodwin, P., & Wright, G. (1994). Heuristics, biases and improvement strategies in judgmental time series forecasting. *International Journal of Management Science*, *22*, 553-568.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as nonparametric Bayesian density estimation. In M. Oaksford & N. Chater (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (p. 303-328). Oxford: Oxford University Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22-44.

Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review*, *113*, 677-699.

Lawrence, M., Goodwin, M., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, *22*, 493-518.

Navarro, D. J. (2007). Similarity, distance and categorization: A discussion of Smith's (2006) warning about "colliding parameters". *Psychonomic Bulletin & Review*, *14*, 823–833.

Navarro, D. J., & Perfors, A. (2009). Learning time-varying categories. In *Proceedings of the 31st annual conference of the cognitive science society* (p. 412-424). Austin, TX: Cognitive Science Society.

Navarro, D. J., & Perfors, A. (2012). Anticipating changes: Adaptation and extrapolation in category learning. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 809–814). Austin, TX: Cognitive Science Society.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *10*, 104-114.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*(2), 211-233.

Rubin, D. C., Hinton, S., & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *25*, 1161-1176.

Sakamoto, Y., Jones, M., & Love, B. C. (2008). Putting the psychology back into psychological models: Mechanistic versus rational approaches. *Memory and Cognition*, *36*, 1057-1065.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*, 1144-1167.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *28*, 3-11.

Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition*, *25*, 731–739.

Yu, A. J., & Cohen, J. D. (2009). Sequential effects: Superstition or rational behavior? *Advances in Neural Information Processing Systems*, *21*.