Cognitive Models of Test-Item Effects in Human Category Learning

Xiaojin Zhu, Bryan R. Gibson, Kwang-Sung Jun {JERRYZHU,BGIBSON,DELTAKAM}@CS.WISC.EDU
Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 53706 USA

Timothy T. Rogers, Joseph Harrison {TTROGERS,JCHARRISON}@WISC.EDU
Department of Psychology, University of Wisconsin-Madison, Madison, WI 53706 USA

Chuck Kalish CWKALISH@WISC.EDU
Department of Educational Psychology, University of Wisconsin-Madison, Madison, WI 53706 USA

Abstract

Imagine two identical people receive exactly the same training on how to classify certain objects. Perhaps surprisingly, we show that one can then manipulate them into classifying some test items in opposite ways, simply depending on what other test items they are asked to classify (without label feedback). We call this the Test-Item Effect, which can be induced by the order or the distribution of test items. We formulate the Test-Item Effect as online semi-supervised learning, and extend three standard human category learning models to explain it.

1. Introduction

One common approach to studying human categorization in cognitive psychology is to build explicit statistical learning models that fit observed behaviors. This approach shares the same goal as statistical machine learning: both aim at uncovering the underlying mathematical principles of the learning process. In machine learning terms, human categorization is a classification task: given training data $\{(x_i, y_i)\}$ for $i = 1 \dots n$, learn a classifier $f : \mathcal{X} \mapsto \mathcal{Y}$. Here $x_i \in \mathcal{X}$ is a stimulus (or item), usually represented as a feature vector in $\mathbb{R}^d$, and $y_i \in \mathcal{Y}$ is the category label (e.g., 0 or 1) given to the learner. The classifier f comes from the implicit family of classification functions that the human mind can produce. Similar to machine learning, cognitive studies typically assume a training phase where the classifier f is trained, and a test phase where f is *held fixed* and applied to test items $x_{n+1}, x_{n+2}, \dots$ to predict $f(x_{n+1}), f(x_{n+2}), \dots$ The test phase is conducted without label feedback, i.e., without giving the human learner the true label $y_{n+1}, y_{n+2}, \dots$

While it is certainly possible to fix f during testing in machine learning, the validity of this assumption is less clear in human learning. Because there is no label feedback during test, one might be led to think that there is no information in the test items regarding the mapping from $\mathcal{X}$ to $\mathcal{Y}$ to merit a change in f. However, recent theoretical and empirical studies on *semi-supervised learning* cast doubt on this claim. These studies show that, under appropriate assumptions, both labeled and unlabeled data will influence the trained classifier (Chapelle et al., 2006; Zhu & Goldberg, 2009). The train/test split can be viewed as a special case of semi-supervised learning, where all the labeled data come first, and all the unlabeled data come next.

We propose the term *Test-Item Effect* to denote the possibility that unlabeled test items can induce changes to the classifier f in human category learning. *Specifically, the Test-Item Effect predicts that two otherwise identical people A, B receiving exactly the same training data can be made to disagree on certain test items x^*, i.e., $f_A(x^*) \neq f_B(x^*)$, simply by manipulating what other test data $x^A_{n+1} \dots$ and $x^B_{n+1} \dots$ they are asked to classify, respectively.* Note they do not receive label feedback on any test items. If the Test-Item Effect is real, it will have profound theoretical and practical implications. Existing category learning models will need to be revised to accommodate such effect. Education and training procedures will need to take the effect into account as well. As an example, consider airport security personnel that scan luggage, categorizing it as safe/suspicious. It would be important to understand whether encountering partic-

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

ular luggage items could unwittingly bias their category decision.

Some past research does show that classification behavior can be influenced by the construction of the test set. Zaki and Nosofsky showed that, in a one-class classification setting (where the training data consists of positively labeled items only, and the goal is to classify a test item as positive or not), the presence of a new, tight cluster in the test data shifts people's perception of the prototypical training item (Zaki & Nosofsky, 2007). More dramatically, Palmeri and Flanery showed that even with zero training data, people can perform one-class classification if there is a cluster in test data (Palmeri & Flanery, 1999). Fried and Holyoak investigated mixture models separately in labeled and unlabeled settings but did not mix them (Fried & Holyoak, 1984). Zhu *et al.* showed that humans use both labeled and unlabeled data to estimate the class-conditional distribution $p(x|y)$ in a mixture model setting (Zhu et al., 2007).

Despite such evidence, the Test-Item Effect is not yet well-understood. This paper makes two key contributions: (i) We describe a human experiment where the *order* of the same set of test items induces a strong Test-Item Effect. This experiment complements previous studies which change the test item distribution. (ii) We introduce novel "online semi-supervised" variants of three standard human category learning models: the exemplar, the prototype, and the rational model of categorization. Our new models perform online (incremental) semi-supervised learning, which is necessary to explain test-item order effect.

The goals of this paper are (i) to report two kinds of Test-Item Effects observed in a human category-learning task, and (ii) to assess whether the different models vary in how well they match the reported human behavior. If some models fit the human data better than others, this suggests that the Test-Item Effects might importantly constrain computational accounts of human category learning. The main contribution of this work is thus to cognitive psychology, although machine learning researchers may find some of the semi-supervised models interesting as well.

2. The Test-Item Effects in Human Category Learning

We first present a new human experiment on test item *order*, then review an experiment in (Zhu et al., 2007) on test item *distribution*. Taken together, these experiments demonstrate that the Test-Item Effect does occur, and can have multiple causes. A human category learning model needs to be able to explain all such Test-Item Effects.

2.1. The Test-Item Effect Due to Order

This experiment consists of two identical conditions except for one aspect: They share the same set of test items, but differ in the order the test items are presented to the subjects. As we show below, subjects in these two conditions consistently disagree on the label of certain test items.

Participants and Materials: 40 undergraduate students participated for partial course credit. The stimuli are novel shapes (Figure 1), varying according to a single continuous parameter $x \in [-2, 2]$. There are two classes, denoted as $y = 0$ or $y = 1$.

Procedure: In trial n, a stimulus x_n appears on a computer screen, and stays on until the subject presses one of two keys to label it. All subjects initially have the same 10 labeled trials, where two items occur alternatively: $(x_n, y_n) = (-2, 0), (2, 1), (-2, 0), (2, 1) \ldots$ For these 10 trials, after the subject presses her key, a label feedback appears on screen indicating whether her classification was correct (same as y_n). The computer screen is then cleared, and the stimulus for the next trial appears. After these labeled trials, subjects are presented with a series of 81 evenly spaced unlabeled test items: $x_n = -2, -1.95, -1.9, \ldots, 2$. The test items appear one at a time on the screen, and the subjects have to classify them using the same procedure. However, there is no longer labeled feedback after each classification. Importantly, the subjects are randomly divided into two conditions of equal size. In the "L to R" condition, the order of the test items is as above. In the "R to L" condition, the order is the opposite (i.e., $2, 1.95, 1.9, \ldots, -2$).

Result: In Figure 2(Left) we plot $P(y = 1|x)$, estimated by the fraction of subjects in each condition who classified x with label 1. The difference is striking[1]. Subjects in the "L to R" condition tend to classify more test items as $y = 0$, while those in the "R to L" condition tend to classify more as $y = 1$. For instance, for the same test item $x = -0.5$, only 4 out of 20 subjects in the "L to R" condition classified it as $y = 1$,

[1] We point out that the two curves in Figure 2(Left) are not symmetric about $x = 0$, as one would expect. We speculate that this is due to the stimulus space in Figure 1 not being perceptually uniform. Our feature x is a parameter used to generate the geometry of the shapes, and does not necessarily match the human perceived similarity between stimuli. Nonetheless, this does not affect the validity of the observed Test-Item Effect, which only depends on the two curves separating from each other.

Figure 1. Example stimuli

while 15 out of 20 subjects in the "R to L" condition did so. This is significantly different using log odds ratio at $p < 0.0004$. It is clear evidence of the Test-Item Effect, where the effect is produced by the order of test items. In fact, for test items $x \in [-1.2, 0.1]$ a majority-vote among subjects will classify them in opposite ways in those two conditions.

We postulate that the subjects might perform self-reinforcement: that once a person classifies a test item x as in class y, the predicted label y (perhaps weighted by its uncertainty) becomes a training label for the person. For example, for a subject in the "L to R" condition, the first few test items are all near $x = -2$. The subject can easily classify them as $y = 0$ from the training she just received. If self-reinforcement is in effect, these test items will act as additional training data for the $y = 0$ class. This will tend to favor classifying more test items as $y = 0$. The opposite can be said for the "R to L" condition. Such self-reinforcement corresponds to the self-training algorithm in semi-supervised learning. Under certain probabilistic models, it can also be interpreted as an Expectation-Maximization (EM) procedure. This insight will underlie our models in Section 3.

2.2. The Test-Item Effect Due to Distribution

We briefly review the experiment in (Zhu et al., 2007), which represents a different kind of Test-Item Effect, and whose data will be used in model fitting later. The task is again binary categorization in the same stimulus space. First, all subjects received 20 labeled items: $(x, y) = (-1, 0)$ and $(1, 1)$, repeated 10 times each in random order. In the subsequent 795 test trials, subjects were divided into two conditions. The two conditions differ only in how the majority of the test items were generated (the other test items were grid points shared across conditions): For 12 subjects in the "L shifted" condition, 690 of the test items were randomly sampled from a mixture of two Gaussian distributions with means at -1.43 and 0.57. These means are shifted to the left of the labeled training items. For 10 subjects in the "R shifted" condition, the means are -0.57 and 1.43. Again, there was no label feedback on any test items.

We show their results in Figure 2(Right) by pooling all subjects in a condition together, and fit two logistic regression curves: human categorization on the first 50 test items (marked "early") and the last 50 test items ("late"). In early test, the two conditions are essentially the same, and the curves overlap. In late test however, the curves are dramatically different. The final decision threshold, i.e., the x with $P(y = 1|x) = 1/2$, shifted in opposite directions in these two conditions. During late test, items $x \in [-0.07, 0.50]$ received the opposite majority classification in those two conditions. This difference represents the Test-Item Effect due to test-item distributions.

3. Online Semi-Supervised Learning Models for the Test-Item Effect

The human behavioral data in Section 2 suggests strong Test-Item Effects due to the order or distribution of test data. We follow these desiderata in proposing new quantitative cognitive models to explain such data: *i*) These models should explain the effect of unlabeled data for classification, as opposed to unsupervised clustering, which is well studied in psychology. This suggests semi-supervised learning models. *ii*) They need to learn incrementally to explain order effects. Besides which, incremental learning is a better fit to human learning experience than batch learning. *iii*) They should build upon existing human category learning models. *iv*) They should have as few parameters as possible to prevent overfitting human behavioral data.

With these in mind, we propose three online semi-supervised learning algorithms to model the Test-Item Effect in humans. They extend the exemplar, the prototype, and the rational categorization models in psychology, which correspond to kernel regression, Gaussian mixture models, and Dirichlet process mixture models in machine learning, respectively. Although the variety of semi-supervised learning models has flourished in recent years, we will only employ self-training and generative models, and leave more advanced semi-supervised assumptions such as manifolds or large margin separation to future work.

3.1. Semi-Supervised Exemplar Model

The generalized context model (Nosofsky, 1986) in cognitive psychology assumes that people store training items in memory, and make a category prediction for a new item by a weighted average of training item categories. The weight decreases as the distance between the new item and the training item increases. This exemplar model can be

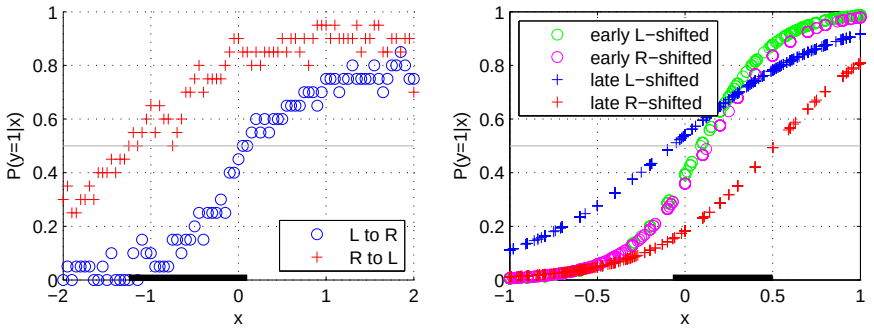

Figure 2. The Test-Item Effect due to order (Left) and distribution (Right). The thick black lines mark items on which the majority human classification differs in the two conditions.

related to the Nadaraya-Watson kernel estimator in machine learning (Nadaraya, 1964; Wasserman, 2006; Shi et al., 2008). In batch supervised learning, given training data $\{(x_i, y_i)\}_{i=1\ldots n}$, one can create a nonparametric regression function $r(x)$ as follows: $r(x) = \sum_{i=1}^{n} \frac{K(\frac{x-x_i}{h})}{\sum_{j=1}^{n} K(\frac{x-x_j}{h})} y_i$. The kernel K, which determines the weight, can be any non-negative smooth function that integrates to one, having zero mean and non-zero variance. A common choice is the Gaussian kernel $K(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$, although in theory the choice of kernel is not critical. More important is the bandwidth parameter h, which controls how narrow or broad the kernel is. For binary classification with $y \in \{0, 1\}$, $r(x)$ can be viewed as a direct estimate of $P(y = 1|x)$, and label prediction can be made by thresholding $r(x)$ at 0.5.

We propose an online semi-supervised extension of the Nadaraya-Watson estimator via self-training, as in Algorithm 1. If x_n is unlabeled, the algorithm predicts its label based on $r(n)$ as estimated from the current model. The algorithm then treats the real-valued $r(n)$ as if it is a soft label for x_n, and uses it in subsequent iterations. This is a nonparametric, semi-supervised learning model that learns incrementally, with a single bandwidth parameter h.

3.2. Semi-Supervised Prototype Model

Prototype models in cognitive psychology assume that people abstract the central tendency or "prototype" of each category, and make a category prediction by comparing the new item to the prototypes (Posner & Keele, 1968; Reed, 1972; Rosch, 1973; Rosch et al., 1976). Generative mixture models are their parallels in machine learning. For concreteness, we discuss two-component Gaussian Mixture Model with parameters $\theta = (\alpha, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$. The model is defined by

Algorithm 1 Semi-Supervised Exemplar Model

Parameter: kernel bandwidth h
for $n = 1, 2, \ldots$ **do**
  Receive x_n, predict its label by thresholding $r(x_n) = \sum_{i=1}^{n-1} \frac{K(\frac{x_n - x_i}{h})}{\sum_{j=1}^{n-1} K(\frac{x_n - x_j}{h})} \hat{y}_i$ at 0.5
  Receive y_n (may be unlabeled), update model:
  if y_n is unlabeled **then**
    $\hat{y}_n = r(x_n)$
  else
    $\hat{y}_n = y_n$
  end if
end for

$P(x, y|\theta) = P(y|\theta)P(x|y, \theta)$ where

$$P(y|\theta) = \alpha^y (1-\alpha)^{1-y}$$
$$P(x|y, \theta) = \left(2\pi\sigma_y^2\right)^{-1/2} \exp\left(-(x-\mu_y)^2/(2\sigma_y^2)\right).$$

The label of a new item x_n is predicted according to the posterior $q(y)$:

$$q(y) \equiv P(y|x_n, \theta) = \frac{P(x_n, y|\theta)}{\sum_{y'=0,1} P(x_n, y'|\theta)}.$$

In batch supervised learning, the parameters θ can be trivially estimated from training data as the Maximum Likelihood Estimate. In batch semi-supervised learning where some items are unlabeled, the parameters can still be well-estimated (up to local optima) using the (batch) EM algorithm (Dempster et al., 1977). However, in online semi-supervised learning, parameter estimation is more difficult. The closest models are the incremental EM algorithm of (Neal & Hinton, 1998) and the category density model of (Fried & Holyoak, 1984). Neal and Hinton assumed that one can "loop back" and revisit the items repeatedly when updating parameters. From a cognitive modeling perspective, it is advantageous to assume a true

online learning setting, where items are encountered only once and not revisited indefinitely. Fried and Holyoak's model does this, but it assumes a "transfer" (i.e. test) phase where the model parameters are no longer updated, and thus will not model the Test-Item Effects.

With this constraint, we define a variant of the incremental EM algorithm in Algorithm 2. In a human learning analogy, one keeps incremental track of the number of items in each Gaussian component, and the component's mean and variance, through sufficient statistics. If an input item (x, y) is labeled, its sufficient statistics vector is

$$\tilde{\phi}(x, y) = (1 - y, (1 - y)x, (1 - y)x^2, y, yx, yx^2).$$

Importantly, if the item is unlabeled, one first estimates the label distribution $q(y)$ using the current parameters, and then assigns the contribution of the item to both Gaussian components, weighted by $q(y)$. The sufficient statistics vector for an unlabeled item x is

$$\mathbb{E}_{y\sim q}[\tilde{\phi}(x, y)] = \sum_{y=0,1} q(y)\tilde{\phi}(x, y).$$

With the accumulative sufficient statistics, one re-estimates the parameters as in the M-step.

It is easy to show that Algorithm 2 maximizes a lower bound of the log likelihood of all observed data so far. But unlike the incremental EM algorithm of (Neal & Hinton, 1998), it is not guaranteed to maximize the log likelihood itself. This is a price we pay for not revisiting old items. Nonetheless, it can still be a valid cognitive model for human category learning.

It is possible to initialize the accumulative sufficient statistics vector ϕ with certain values, which is equivalent to a prior encoded in pseudo counts. In particular, the vector $\phi = (n_0, 0, n_0, n_0, 0, n_0)$ represents n_0 pseudo items with mean zero and variance one for each Gaussian. In our model fitting experiments below, we will treat n_0 as the sole parameter of Algorithm 2. Then, Algorithm 2 is a parametric, semi-supervised learning model that learns incrementally, with a single prior parameter n_0.

3.3. Semi-Supervised Rational Model of Categorization

Although our task is classification, it is instructive to think what "clusters" the previous two models conceptually produce. Exemplar models have effectively n singleton clusters, one for each input item. Prototype models have only two clusters (i.e., Gaussian components) for binary classification. Anderson's Rational Model of Categorization (RMC) (Anderson, 1990) is somewhere in between: it groups items into clusters, but the number of clusters can grow. This is a balance between memory load and classifier flexibility. In fact, the RMC was later discovered to be equivalent to Dirichlet process mixture model (DPMM) (Sanborn et al., 2006; Griffiths et al., 2008), a nonparametric Bayesian model (Neal, 2000; Rasmussen, 2000). We present a semi-supervised variant of DPMM with particle filtering. Our model is similar to the AClass model of (Mansinghka et al., 2007), which was used for supervised learning. But unlike AClass where each category has its own private DPMM, we stack (x, y) into an extended feature vector and use one global DPMM: $G \sim DP(G_0, \alpha_2)$, $\theta_1 \dots \theta_n \sim G$, $(x_i, y_i) \sim F(x, y|\theta_i)$, where G_0 is a base distribution which we take to be the product of Normal-Gamma and Beta, conjugate priors for Normal and binomial: $G_0 = \text{NG}(\mu_0, \kappa_0, \alpha_0, \beta_0)\text{Beta}(\alpha_1, \beta_1)$. $\theta = (\mu, \lambda, p)$ is a parameter vector with the mean and precision of a Gaussian for the x component, and the "head" probability for the y component. Due to the property of the Dirichlet process, many θ's will be identical, creating an implicit clustering of items. F is a product of Gaussian and Bernoulli: $F = \text{Norm}(x; \mu, \lambda)p^y(1-p)^{1-y}$. As is common with DPMM, we introduce cluster membership indices $z_1 \dots z_n$, and integrate out θ and G via particle filtering (Fearnhead, 2004). That is, at iteration $n - 1$ we assume the distribution $P(z_{1:n-1} \mid x_{1:n-1}, y_{1:n-2})$ is well-approximated by the empirical distribution on m particles $z^{(1)}_{1:n-1}, \dots, z^{(m)}_{1:n-1}$, each particle is a vector of indices:

$$P(z_{1:n-1} \mid x_{1:n-1}, y_{1:n-2}) \approx \frac{1}{m}\sum_{l=1}^{m} \delta(z_{1:n-1}, z^{(l)}_{1:n-1}),$$

Algorithm 2 Semi-Supervised Prototype Model

Parameter: Prior encoded in ϕ
Initialize $\theta^{(0)}$ from ϕ (see M-step below)
for $n = 1, 2, \dots$ **do**
  Receive x_n, classify by $q(y) = P(y|x_n, \theta^{(n-1)})$
  Receive y_n (may be unlabeled), update model
  E-step:
  if y_n is unlabeled **then**
    $\phi = \phi + \mathbb{E}_q[\tilde{\phi}(x_n, y)]$
  else
    $\phi = \phi + \tilde{\phi}(x_n, y_n)$
  end if
  M-step: Let $\phi = (n_0, s_0, ss_0, n_1, s_1, ss_1)$. Compute $\theta^{(n)}$ as follows: $\alpha = \frac{n_1}{n_0+n_1}$, $\mu_0 = \frac{s_0}{n_0}$, $\sigma_0^2 = \frac{ss_0}{n_0} - \left(\frac{s_0}{n_0}\right)^2$, $\mu_1 = \frac{s_1}{n_1}$, $\sigma_1^2 = \frac{ss_1}{n_1} - \left(\frac{s_1}{n_1}\right)^2$
end for

where $\delta(u, v) = 1$ if $u = v$, and 0 otherwise. Then, at iteration n, after we observe the input item x_n but before seeing its label y_n, the distribution $P(z_{1:n} \mid x_{1:n}, y_{1:n-1})$ can be shown to be proportional to

$$\sum_{l=1}^{m} \delta(z_{1:n-1}, z^{(l)}_{1:n-1}) P(y_{n-1} \mid z^{(l)}_{1:n-1}, y_{1:n-2}) \\ P(z_n \mid z^{(l)}_{1:n-1}) P(x_n \mid z_n, z^{(l)}_{1:n-1}, x_{1:n-1}). \quad (1)$$

One would further sample from (1) m new particles $z^{(1)}_{1:n}, \ldots z^{(m)}_{1:n}$. The empirical distribution on these new particles will approximate $P(z_{1:n} \mid x_{1:n}, y_{1:n-1})$. This update is the key to particle filtering, which uses a fixed number of particles to approximate an increasingly complex distribution.

In (1), one needs to compute three conditional probabilities. The conditional probability of z_n is computed from the Chinese Restaurant Process prior. Let there be K unique index values $1 \ldots K$ in $z_{1:n-1}$, then

$$P(z_n = k \mid z_{1:n-1}) = \begin{cases} n_k/(\alpha_2 + n - 1), & k \le K \\ \alpha_2/(\alpha_2 + n - 1), & k = K + 1 \end{cases}$$

where n_k is the number of indices with value k in $z_{1:n-1}$. The conditional probability of x_n is computed from a student-t distribution,

$$P(x_n \mid z_{1:n}, x_{1:n-1}) = t_{2\alpha}\left(x_n \mid \mu, \beta(\kappa + 1)/(\alpha\kappa)\right),$$

with $\mu = (\kappa_0\mu_0 + N\bar{x})/(\kappa_0 + N)$, $\kappa = \kappa_0 + N$, $\alpha = \alpha_0 + N/2$, $\beta = \beta_0 + \frac{1}{2}\sum_{i=1}^{N} \delta(z_i, z_n)(x_i - \bar{x})^2 + (\kappa_0 N(\bar{x} - \mu_0)^2)/(2(\kappa_0 + N))$, and $N = \sum_{i=1}^{n-1} \delta(z_i, z_n)$, $\bar{x} = \frac{1}{N}\sum_{i=1}^{n-1} \delta(z_i, z_n) x_i$.

Importantly, for our semi-supervised variant of DPMM, the conditional probability of y_{n-1} is computed from a beta-binomial distribution

$$P(y_{n-1} \mid z_{1:n-1}, y_{1:n-2}) = \frac{c_1 + \alpha_1}{c_0 + c_1 + \alpha_1 + \beta_1}. \quad (2)$$

Note some of the y's might be unlabeled. If y_{n-1} is unlabeled, the probability is simply 1 since it must take either one of the labels. If some y's in $y_{1:n-2}$ are unlabeled, one can show that those are marginalized over, resulting in the following counts: $c_1 = \sum_{i=1}^{n-2} \delta(z_i, z_{n-1})\delta(y_i, 1)$, and $c_0 = \sum_{i=1}^{n-2} \delta(z_i, z_{n-1})\delta(y_i, 0)$. Here, we define $\delta(y_i, 1) = \delta(y_i, 0) = 0$ if y_i is unlabeled. Once the particles are updated with (1), predicting y_n is straightforward:

$$p(y_n \mid x_{1:n}, y_{1:n-1}) \approx \frac{1}{m}\sum_{l=1}^{m} p(y_n \mid z^{(l)}_{1:n}, y_{1:n-1}) \quad (3)$$

Algorithm 3 Semi-Supervised Rational Model of Categorization

Parameters: $\alpha_2, \mu_0, \kappa_0, \alpha_0, \beta_0, \alpha_1, \beta_1$
Initialize m empty particles; y_0 =unlabeled
for $n = 1, 2, \ldots$ **do**
  Receive y_{n-1} (may be unlabeled) and x_n
  Re-sample m particles from (1)
  Predict y_n with new particles from (3)
end for

where $p(y_n \mid z^{(l)}_{1:n}, y_{1:n-1})$ is computed with (2). The complete algorithm is given in Algorithm 3. In our experiments, we use $m = 64$ particles, set the hyperparameters at $\mu_0 = 0, \kappa_0 = 1, \alpha_0 = \beta_0 = 1, \alpha_1 = \beta_1 = 1$, and leave the Dirichlet process concentration parameter α_2 as the sole parameter.

4. Model Comparison

Parameter tuning. Let $(x^{[s]}_n, y^{[s]}_n)$, $n = 1, 2, \ldots$ be the sequence of training and test data that the s-th subject saw during human experiments, where some y's may be unlabeled. Furthermore, let $h^{[s]}_n \in \{0, 1\}$ be the binary classification response the s-th subject made at trial n. Each of our models predicts the label probability $P(y_n | x_{1:n}, y_{1:n-1}, \theta)$ at trail n, given parameter $\theta = h, n_0$, or α_2. We define training set log likelihood as

$$\ell_{tr}(\theta) \equiv \sum_{s \in tr} \sum_{n} \log P(h^{[s]}_n \mid x^{[s]}_{1:n}, y^{[s]}_{1:n-1}, \theta).$$

Because the order and distribution tasks used the same stimuli, we merge their subjects and fit a single parameter for both tasks.[2] Specifically, we take 32 subjects, eight each from the "order task L to R", "order task R to L", "distribution task L shifted", and "distribution task R shifted" conditions to form the training set tr. The remaining 4, 2, 12, 12 subjects in those conditions form the test set te, and define test set log likelihood $\ell_{te}(\theta)$ accordingly. These sets are shared by the three models. For each model, we find the maximum likelihood estimate parameter $\hat{\theta} = \arg\max_\theta \ell_{tr}(\theta)$ on the training set using a coarse parameter grid as shown in Figure 3.

Observations. We report in Table 1 the log likelihood $\ell_{te}(\hat{\theta})$ on the *test set*, which was not involved in parameter tuning. In addition, Figure 4 shows the behavior of the three models over a wide range of param-

[2] This reduces data sparsity. We assume that because the stimulus space is the same, and the learners have no prior knowledge that the tasks are different, they will use the same parameter setting in both tasks.

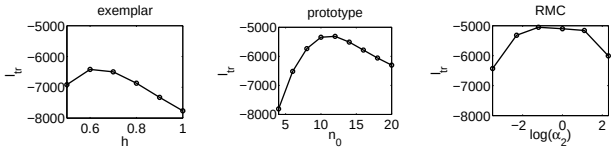

Figure 3. Training set log likelihood $\ell_{tr}(\theta)$ for some θ

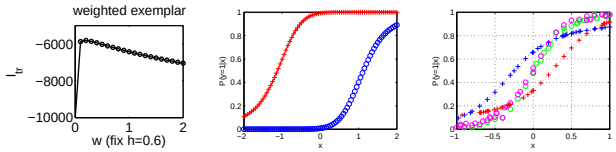

Figure 5. Down-weight unlabeled exemplars

Table 1. Test set log likelihood $\ell_{te}(\hat{\theta})$

	exemplar	prototype	RMC
$\hat{\theta}$	$h = 0.6$	$n_0 = 12$	$\alpha_2 = 0.3$
$\ell_{te}(\hat{\theta})$	-3727	-2460	**-2169**

eters (including $\hat{\theta}$). We make a few observations: (i) All three models predict test-items effects. All models show different classification behavior following the same supervised training depending upon the order and distribution of the test items. (ii) Some models are more consistent with the empirical data than others. Specifically, the semi-supervised RMC model showed a qualitatively similar pattern (and the best log-likelihoods) to both datasets under a range of parameter values. The prototype model fared well under some parameter choices but not others; and the exemplar model failed to qualitatively match the empirical data under any of the studied parametrization. The test item effect thus provides evidence useful for constraining theories of human categorization. In this case, it suggests that the RMC provides a better approximation of human category learning than either prototype or exemplar theories, though to more firmly assess this hypothesis it will be necessary to consider other parametrization of the later kinds of models.

Down-weight unlabeled exemplars. Our semi-supervised exemplar model has the lowest likelihood. On the "order" task, the two curves are too wide apart; on the "distribution" task, they overlap, cross, or even flip. A natural idea for improvement is to afford a weight parameter $w < 1$ to unlabeled exemplars: perhaps a self-assigned label is worth less than a true label. Specifically, one can adapt the Nadaraya-Watson kernel estimator into $r(x) = \sum_{i=1}^{n} \frac{w_i K(\frac{x-x_i}{h})}{\sum_{j=1}^{n} w_i K(\frac{x-x_j}{h})} y_i$, with $w_i = w$ if x_i is unlabeled, and $w_i = 1$ otherwise. Figure 5(left) shows $\ell_{tr}(w, h = 0.6)$ for the exemplar model with w ranging from 0 (supervised learning) to 2 (overweight). Clearly, semi-supervised learning ($w > 0$) is much better than supervised learning at explaining the human data. Training likelihood peaks at $w = 0.2$ and decreases thereafter. The *test set* log likelihood with $w = 0.2, h = 0.6$ is -2934, still worse than the other two models (which have only one parameter). The other two panels in Figure 5 show exemplar model predictions similar to the top row of Figure 4, but with $w = 0.2, h = 0.6$. Overall, down-weight unlabeled exemplar helps, but not overwhelmingly.

5. Conclusions

We have presented a novel Test-Item Effect in human categorization, induced by test item order. Together with previously known distribution-induced effect, it calls for new online semi-supervised learning models. We presented such extensions to the standard exemplar, prototype, and RMC models. Simulations show that all of our models exhibit the Test-Item Effect, with semi-supervised RMC giving the best fit.

Acknowledgments We thank Ruichen Qian for assisting with human experiments. This work is supported in part by AFOSR FA9550-09-1-0313, NSF IIS-0916038, IIS-0953219, DLS/DRM-0745423, and the Wisconsin Alumni Research Foundation. The voluntary, fully informed consent of the subjects used in this research was obtained as required by 32 CFR 291 and AFI 40-402.

References

Anderson, J. R. *The adaptive character of thought*. Erlbaum, Hillsdale, NJ, 1990.

Chapelle, Olivier, Zien, Alexander, and Schölkopf, Bernhard (eds.). *Semi-supervised learning*. MIT Press, 2006.

Dempster, A., Laird, N., and Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 1977.

Fearnhead, P. Particle filters for mixture models with an unknown number of components. *Statistics and Computing*, 14:11–21, 2004.

Fried, L. S. and Holyoak, K. J. Induction of category distributions: a framework for classification learning. *Journal of experimental Psychology*, 10(2):234–257, 1984.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., and Navarro, D. J. Categorization as nonparametric Bayesian density estimation. In Oaksford, M. and Chater, N. (eds.), *The probabilistic mind: Prospects for rational models of cognition*. Oxford University Press, Oxford, 2008.

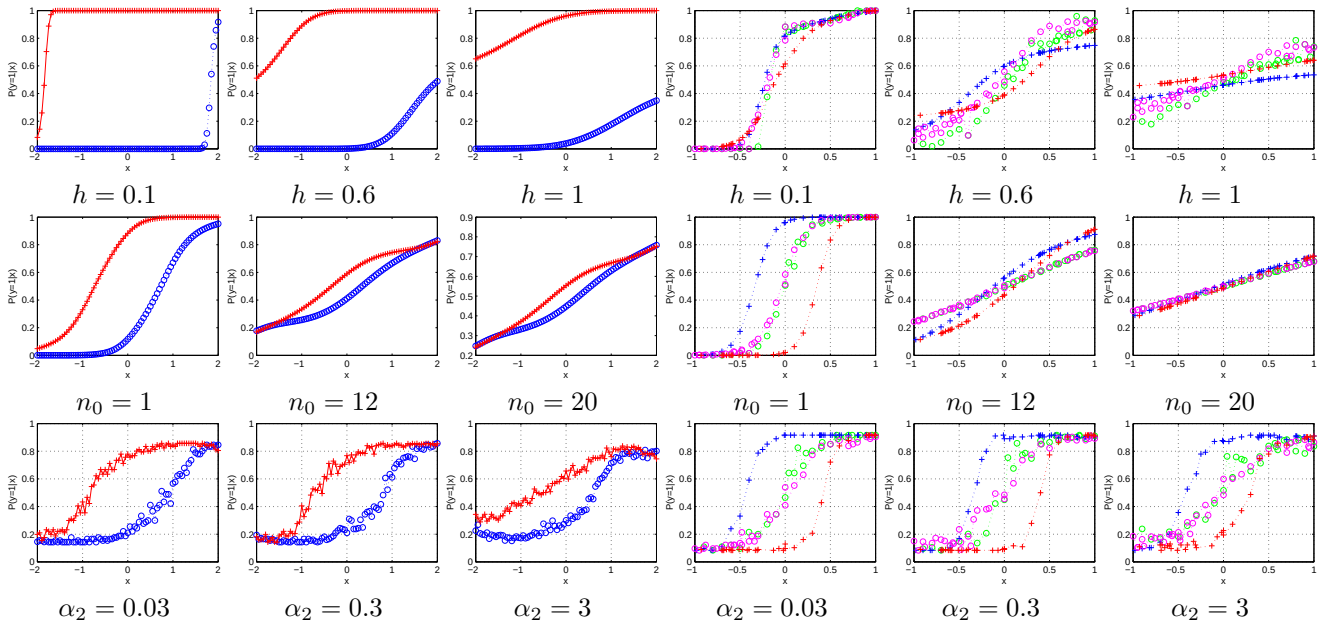

Figure 4. (Top) Semi-supervised exemplar model, (middle) Semi-supervised prototype model, (bottom) Semi-supervised rational model of categorization. Columns 1–3 show model predictions $P(y_n = 1|x_{1:n}, y_{1:n-1})$ on the "order" task (Section 2.1), and columns 4–6 the "distribution" task (Section 2.2). The legend is the same as in Figure 2.

Mansinghka, V. K., Roy, D. M., Rifkin, R., and Tenenbaum, J. B. Aclass: An online algorithm for generative classification. In *AISTATS*, 2007.

Nadaraya, E.A. On estimating regression. *Theory of Probability and Its Application*, 9, 1964.

Neal, R. M. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

Neal, R. M. and Hinton, G. E. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M. (ed.), *Learning in Graphical Models*. 1998.

Nosofsky, R. M. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39–57, 1986.

Palmeri, Thomas J. and Flanery, Marci A. Learning about categories in the absence of training: Profound amnesia and the relationship between perceptual categorization and recognition memory. *Psychological Science*, 10(6): 526–530, November 1999.

Posner, M. I. and Keele, S. W. On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77:353–363, 1968.

Rasmussen, C. E. The infinite Gaussian mixture model. In *NIPS 12*, 2000.

Reed, S. K. Pattern recognition and categorization. *Cognitive Psychology*, 3:382–407, 1972.

Rosch, E. On the internal structure of perceptual and semantic categories. In Moore, T. E. (ed.), *Cognitive development and the acquisition of language*. Academic Press, New York, 1973.

Rosch, E., Simpson, C., and Miller, R. S. Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, 2:491–502, 1976.

Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. A more rational model of categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 2006.

Shi, L., Feldman, N. H., and Griffiths, T. L. Performing Bayesian inference with exemplar models. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 2008.

Wasserman, Larry. *All of Nonparametric Statistics (Springer Texts in Statistics)*. Springer, 2006. p. 71.

Zaki, S.R. and Nosofsky, R.M. A high-distortion enhancement effect in the prototype-learning paradigm: Dramatic effects of category learning during test. *Memory & Cognition*, 35(8):2088–2096, 2007.

Zhu, Xiaojin and Goldberg, Andrew B. *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, San Rafael, CA, 2009.

Zhu, Xiaojin, Rogers, Timothy, Qian, Ruichen, and Kalish, Chuck. Humans perform semi-supervised classification too. In *AAAI*, 2007.