Title

A Model-Based Approach to the Wisdom of the Crowd in Category Learning

Journal

Cognitive Science, 42(S3)

ISSN

0364-0213

Authors

Danileiko, Irina
Lee, Michael D

Publication Date

2018-06-01

DOI

10.1111/cogs.12561

Peer reviewed

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2017) 1–23

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12561

A Model-Based Approach to the Wisdom of the Crowd in Category Learning

Irina Danileiko, Michael D. Lee

Department of Cognitive Sciences, University of California, Irvine

Received 16 March 2017; received in revised form 26 July 2017; accepted 13 September 2017

Abstract

We apply the "wisdom of the crowd" idea to human category learning, using a simple approach that combines people's categorization decisions by taking the majority decision. We first show that the aggregated crowd category learning behavior found by this method performs well, learning categories more quickly than most or all individuals for 28 previously collected datasets. We then extend the approach so that it does not require people to categorize every stimulus. We do this using a model-based method that predicts the categorization behavior people would produce for new stimuli, based on their behavior with observed stimuli, and uses the majority of these predicted decisions. We demonstrate and evaluate the model-based approach in two case studies. In the first, we use the general recognition theory decision-bound model of categorization (Ashby & Townsend, 1986) to infer each person's decision boundary for two categories of perceptual stimuli, and we use these inferences to make aggregated predictions about new stimuli. In the second, we use the generalized context model exemplar model of categorization (Nosofsky, 1986) to infer each person's selective attention for face stimuli, and we use these inferences to make aggregated predictions about withheld stimuli. In both case studies, we show that our method successfully predicts the category of unobserved stimuli, and we emphasize that the aggregated crowd decisions arise from psychologically interpretable processes and parameters. We conclude by discussing extensions and potential real-world applications of the approach.

Keywords: Categorization; Category learning; Wisdom of the crowd; General recognition theory; Generalized context model

1. Introduction

Imagine that a team of trainee doctors views a set of skin patches and must categorize them as being malignant or benign. These doctors receive feedback about their responses

Correspondence should be sent to Irina Danileiko, Department of Cognitive Sciences, University of California, Irvine, Irvine, CA 92697-5100. E-mail: idanilei@uci.edu

and, over time, learn to classify skin patches accurately. Presumably they learn which skin patch dimensions, such as color, size, or shape, are important. In addition, they may learn which levels of these dimensions indicate malignancy. A large skin patch that has a light color and a smooth outline might be benign, whereas a small skin patch that has a dark color and a jagged outline might be malignant. It is likely there will be differences in exactly which patches each doctor sees, or at least the sequence in which they see them. It is also likely that there will be individual differences in how well and how quickly the doctors learn to categorize.

The "wisdom of the crowd" is the phenomenon in which an aggregated group answer to a problem is more accurate than the answer of individuals in the group (Surowiecki, 2004). There are at least two ways an aggregate answer can improve upon an individual answer. One way is *signal amplification*, in which combining answers amplifies the common signal and reduces noise. For example, if a skin patch is malignant, that ground truth provides a common signal that competent doctors will reliably detect, while newer doctors may be less consistent in their categorizations. The net result is that the group overall will favor the ground truth of malignancy, even if some individuals believe it to be benign. A second way is *jigsaw completion*, in which different individuals solve different parts of the problem. For example, if there are various types of malignant patches, different doctors may specialize in different types. Relying on the categorizations of the doctors who specialize in each individual patch will maximize the accuracy of the group classification across all patches.

Surowiecki (2004) identifies four requirements for a wise crowd. The first is *diversity*: The individuals need to have a range of different opinions and backgrounds. As the doctor example makes clear, this will often be true of categorization problems, because of individual differences in learning. In general, some people may learn more quickly than others and some people may achieve eventual levels of categorization accuracy that are higher than other people's. It is also possible that not just the rate and final level of learning will differ, but the nature of the learning itself will differ, with some people learning incrementally and gradually improving their accuracy, and others switching between strategies, leading to sudden changes in accuracy. The second is *decentralization*: The individuals need to draw on different information sources. The doctor example again makes clear that categorization often satisfies this requirement. In general, the doctors will learn from different sets of skin patches or experience them in a different order. The third is *independence*: The individuals cannot know too much about what others think, so that they provide additional or different information to the group. If doctors are trained in an individual setting, or are otherwise unaware of the categorizations of the other trainees, this requirement will also be met.

Given that categorization satisfies these three requirements, applying the wisdom of the crowd idea hinges on satisfying Surowiecki's fourth requirement. This is *aggregation*: There must be a method for aggregating individual decisions into a group decision. Since categorization decisions are discrete (usually binary), the simplest method of aggregation is to take the majority decision. There is evidence, despite its simplicity, that the majority can lead to accurate and robust decisions, for both low-level

perceptual and higher order cognitive stimuli (Hastie & Kameda, 2005; Sorkin, Hays, & West, 2001).

In this paper, we study the wisdom of the crowd for category learning using majority decisions. We tackle this challenge in two ways: first, empirically, and then, using cognitive models. In the first part of the paper, we test empirically the accuracy of group learning curves produced by aggregating individual categorization decisions for a number of existing category learning datasets. We find that, in general, these aggregate learning curves perform as well or better than the learning curves of most individuals. This is perhaps not surprising, given the empirical success of aggregation in other behavioral tasks, including estimation (Herzog & Hertwig, 2009; Vul & Pashler, 2008), problem solving (Yi, Steyvers, Lee, & Dry, 2012), ranking and voting (Lee, Steyvers, & Miller, 2014; Selker, Lee, & Iyer, 2017), and competitions (Lee, Zhang, & Shi, 2011). Most of these tasks, however, involve sets of largely independent decisions, whereas category learning involves sequences of repeated decisions, sometimes with structure at the individual level based on the progress of learning. Thus, our finding of a wisdom of the crowd effect for category learning extends the generality of the empirical effect.

In the second half of the paper, we build on the empirical finding using cognitive models. The modeling approach allows the wisdom of the crowd to be extended to the categorization of new stimuli, for which behavioral data do not exist. The key idea is to use models to make predictions of the categorizations an individual would have produced for the new stimuli. The group majority can then be formed across these predictions. We demonstrate this approach in two case studies, involving two different models of categorization and two different stimulus sets. We conclude by discussing potential extensions and applications of the approach.

2. Behavior-based wisdom of the crowd

To test the accuracy of majority group decisions, we examine 28 existing experimental datasets from a set of previous category learning studies. These datasets were collected with ethical approval from the relevant academic institutions. Table 1 details the studies, including information about the total number of participants, the number of stimuli, the number of blocks (a set of trials typically presenting each stimulus once), the nature of the stimuli, and the number of experimental conditions. It is the total number of experimental conditions that totals the 28 datasets. These studies were chosen because they were the only ones for which we could find behavioral data at the level of individual participants and individual trials, and the true category membership of each stimulus is known.

As Table 1 shows, the datasets vary widely in all of these properties, especially in the nature of the stimuli. The stimuli include rectangular shapes of different sizes (Kruschke, 1993a), shapes varying in size and color (Lewandowsky, 2011), adult faces categorized in terms of gender, hair color, and trust (Navarro, Lee, & Nikkerud, 2005), Gabor patches varying in frequency and orientation (Zeithamova & Maddox, 2006), shapes varying in color and form (Lee & Navarro, 2002), Shepard circles of varying size and radial line

angle (Bartlema, 2013; Bartlema, Lee, Wetzels, & Vanpaemel, 2014), and nonsense words (Smith & Minda, 1998).

Fig. 1 shows the results for one dataset coming from the Kruschke (1993a) study. The *x*-axis shows the eight blocks of learning trials, and the *y*-axis shows categorization accuracy. Because there are two categories, an accuracy of 0.5 corresponds to chance performance. The thin gray lines show the performance of each individual participant, plotting their proportion of correct categorization decisions in each block of the experiment. The average of these individual participant accuracies is shown by the dashed blue line.

Fig. 1 also shows the performance of the wisdom of the crowd aggregate. The aggregated decisions categorize a stimulus on each trial, just as individual participants did. The

Table 1
Details of the experimental category learning datasets

Dataset	n_p	n_s	n_b	n_c	Stimuli
Kruschke (1993a)	160	8	8	4	Rectangles
Lewandowsky (2011)	113	8	12	6	Shapes
Navarro et al. (2005)	40	25	8	4	Faces
Zeithamova and Maddox (2006)	170	80	5	4	Gabor patches
Lee and Navarro (2002)	22	9	Varied	4	SHAPES
Bartlema (2013)	34	8	40	1	Shepard circles
Bartlema et al. (2014)	31	8	40	1	Shepard circles
Smith and Minda (1998) Exp. 1	32	14	7	2	Nonsense words
Smith and Minda (1998) Exp. 2	32	14	10	2	Nonsense words

Note. n_p: number of participants across all conditions of the experiment; n_s: number of stimuli; n_b: number of blocks; n_c: number of conditions.

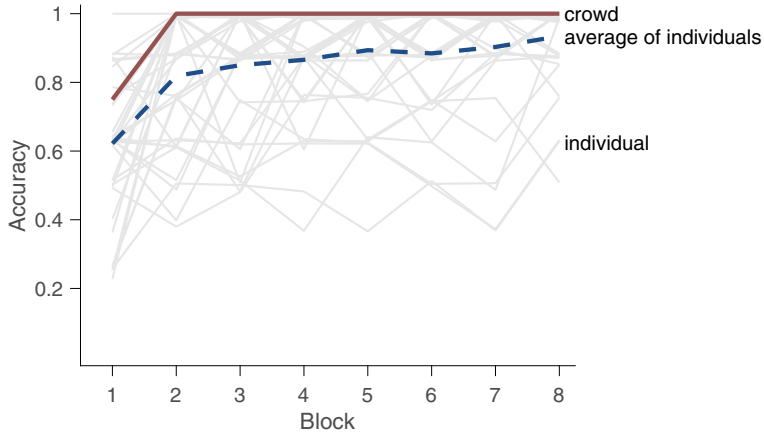


Fig. 1. Learning curves for one experimental condition from the Kruschke (1993a) dataset. The thin gray lines show each individual's proportion of correct answers for each block of the experiment. The dashed blue line shows the average of the individual participant accuracies. The single thick red line shows the categorization accuracy of the aggregated crowd majority decision.

difference is that the aggregate decision is based on the majority of the observed participant behaviors for that stimulus. The single thick red line shows the categorization accuracy of these aggregated majority decisions over the course of the experiment. The learning curve for the crowd achieves perfect accuracy as early as the second block of the experiment. This contrasts favorably with individual performance since only a few participants do slightly better in the first block, and clearly it is superior to the average performance of people.

Fig. 2 shows the same analysis for all of the conditions in all of the datasets from Table 1. Some experimental conditions are easier to learn, while others are harder. For example, the Bartlema et al. (2014) conditions are difficult because of the perceptual confusability of the stimuli, whereas the fourth Navarro et al. (2005) condition is difficult

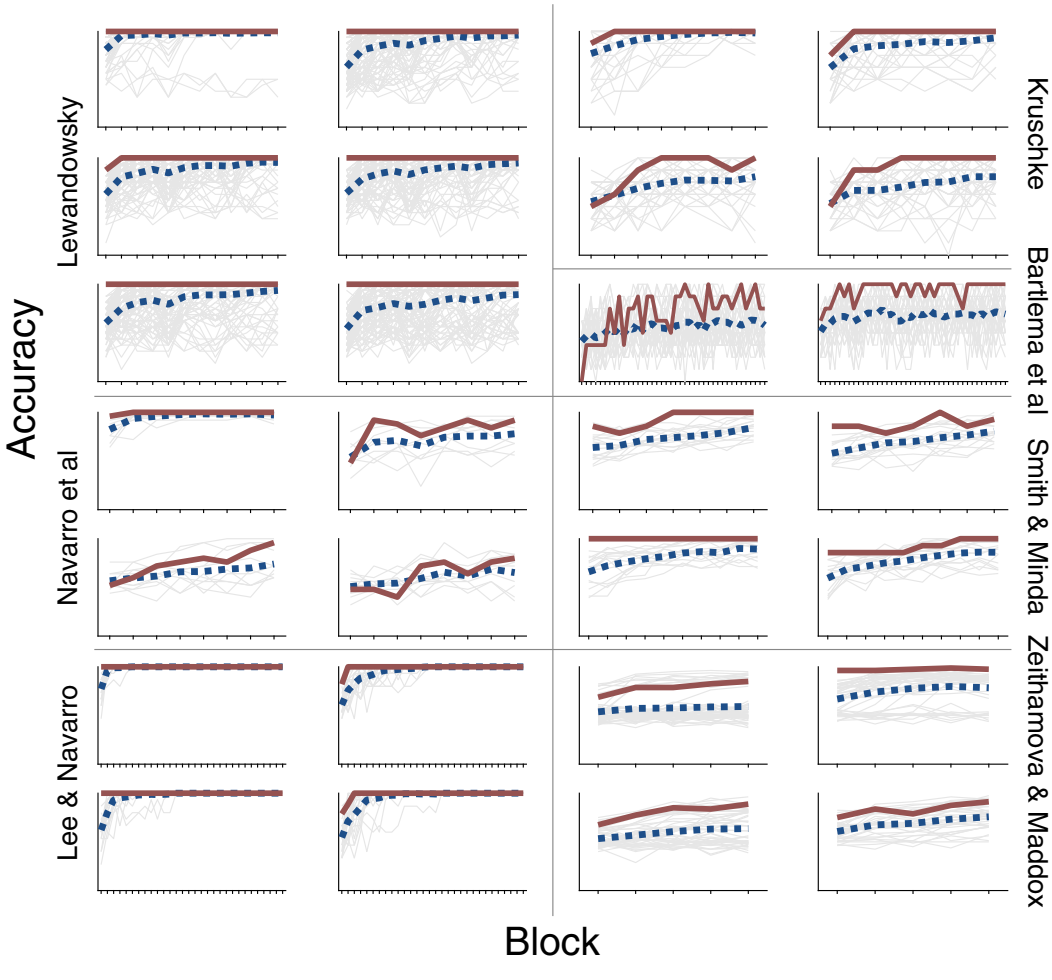


Fig. 2. Learning curves for 28 category learning experiments from eight datasets. As in Fig. 1, the gray lines show individual participant accuracy, the dashed blue lines show the average of the individual participant accuracies, and the red line shows the accuracy of the aggregated crowd majority decisions.

because the stimuli were randomly assigned to categories. In addition, some experimental conditions show clear evidence of individual differences. For example, in several of the Zeithamova and Maddox (2006) conditions, there appears to be two groups of participants, one learning the category structures and reaching high accuracy, and another failing to learn and remaining at poor accuracy throughout the experiment. Despite this variability, the red lines in Fig. 2 show that the crowd performs well. For nearly all of the experimental conditions, the crowd outperforms most or all of the individuals, and almost always outperforms the individual average. The basic result is that taking the majority decision is an effective aggregation method for category learning tasks.

3. Model-based wisdom of the crowd

Imagine now that there is a new skin patch that has not been categorized by any of the trainee doctors. In this case, it is not possible to aggregate observed categorization decisions, and so the behavior-based wisdom of the crowd approach does not apply. If it is possible to predict what each doctor would decide, however, these predictions can be aggregated as if they were behavioral decisions.

In this section, we develop a model-based approach for extending wisdom of the crowd categorization to new stimuli. The idea is to infer a cognitive model of each individual's categorization process based on their decisions for stimuli they have seen and to use that model to predict their decisions for new stimuli. We present two examples of this approach, using two different prominent models of categorization, and involving two different types of stimuli. The first uses general recognition theory (GRT: Ashby & Townsend, 1986) and simple perceptual stimuli, while the second uses the generalized context model (GCM: Nosofsky, 1984, 1986) and face stimuli.

3.1. An application using GRT

3.1.1. Stimuli and data

This application is based on two of the experimental conditions reported by Zeithamova and Maddox (2006). These are the two conditions without memory load: the unidimensional condition, which involves 41 participants, and the information-integration condition, which involves 34 participants. In Fig. 2, these two conditions are the top-right and bottom-right panels in the "Zeithamova & Maddox" section. In both of these conditions, each participant completed five blocks of 80 trials, categorizing Gabor patch stimuli that varied on two dimensions of spatial frequency and spatial orientation. Both conditions gave corrective feedback after every trial, so that the participants could learn to make more accurate categorizations. This application uses data from only the fifth and final block, when participants were the most informed about the category structures.

The two conditions varied in the way the category structures were defined. In the unidimensional condition, stimuli could be accurately categorized solely in terms of their spatial frequencies. In the information-integration condition, both spatial frequency and

spatial orientation were important for determining the correct categorization. Formally, the stimuli belonging to each category were defined using a multivariate normal distribution over the stimulus space. These distributions allow for the generation of new stimuli from each of the categories for both conditions. Thus, Zeithamova and Maddox (2006) provide behavioral decisions for the 80 stimuli in their conditions, but it is possible to generate any number of new stimuli, which participants did not see, but for which their true category membership is known.

3.1.2. GRT model

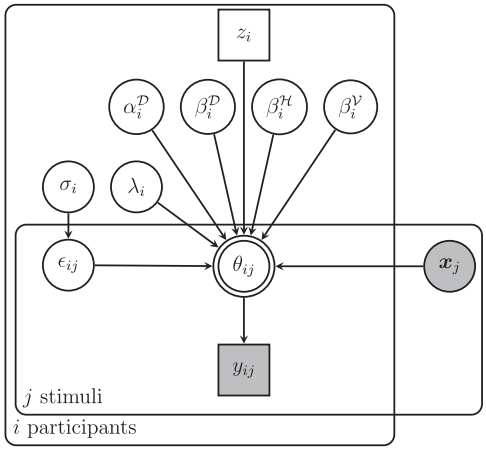
General recognition theory is a decision-bound model of categorization. It assumes that decisions are based on a decision boundary that divides the stimulus space into two categories. Each stimulus is represented as a point that defines its location in this space. In this application, the point $x_j = (x_{1j}, x_{2j})$ represents the spatial frequency and spatial orientation of the jth Gabor patch. GRT assumes that there is variability in the perceptual information associated with each stimulus point on each trial. To account for this, the representation is adjusted to include perceptual noise, so that $x_{pj} = x_j + \varepsilon_p$. Categorization decisions are based on which side of a decision bound this point lies.

The decision bound is a discriminant function of the two dimensions that satisfies the implicit line equation $h(x_1, x_2) = b_1x_1 + b_2x_2 + c$, with the three parameters, b_1, b_2, and c. GRT assumes that there is criterial noise ε_c added to the discriminant function to account for variations in how the participants remember the bound. It also allows for category bias δ, which can be conceived as shifting the decision bound to favor one category over the other. Putting these assumptions together, the probability a participant will choose category A is given by $\Pr(h(x_{pj}) + \varepsilon_c < \delta)$.

3.1.3. Implementation

The GRT has been implemented in a Bayesian framework as a graphical model (Danileiko, Lee, & Kalish, 2015). Graphical models are a formalism that makes it straightforward to implement individual differences using hierarchical structures (Lee, 2011, in press; Lee & Wagenmakers, 2013). In graphical models, parameters and data are represented by nodes, and the structure of the graph indicated the processes by which parameters are assumed to generate data. Unshaded nodes typically indicate latent parameters, while shaded nodes typically indicate observed data or other known values. Circular nodes indicate continuous values while square nodes indicate discrete values. Nodes with a double border represent deterministic variables that are defined as a function of other variables. Finally, rectangular plates indicate replications within the model.

Fig. 3 shows the implementation of the GRT that we applied to model individual categorization behavior for the Zeithamova and Maddox (2006) data. The node y_{ij} is a count of the number of times the ith participant categorized the jth stimulus into category A. The node $\boldsymbol{x}_j$ is the point that represents the jth stimulus in the stimulus space. The probability θ_{ij} is the probability that the ith participant categorizes the jth stimulus into category A, and it is calculated using the cumulative normal distribution $\Phi(\cdot)$. Following GRT, this categorization probability is determined by the decision bound the participant



$$
\begin{aligned}
\alpha_i^{\mathcal{D}} &\sim \text{Uniform}\left(-\tfrac{\pi}{2}, \tfrac{\pi}{2}\right) \\
\beta_i^{\mathcal{H}}, \beta_i^{\mathcal{V}}, \beta_i^{\mathcal{D}} &\sim \text{Uniform}\left(-\tfrac{1}{2}, \tfrac{1}{2}\right) \\
\sigma_i &\sim \text{Uniform}(0, 1) \\
\lambda_i &\sim \text{Uniform}(0, 10) \\
\epsilon_{ij} &\sim \text{Gaussian}\left(0, \tfrac{1}{\sigma_i^2}\right) \\
z_i &\sim \text{Categorical}\left(\tfrac{1}{6}, \ldots, \tfrac{1}{6}\right) \\
\theta_{ij} &= \begin{cases}
\frac{1}{2} & \text{if } z_i = 1 \\
0.01 & \text{if } z_i = 2 \\
0.99 & \text{if } z_i = 3 \\
\Phi\left(\left[x_{j1} - \beta_i^{\mathcal{V}} + \epsilon_{ij}\right]/\lambda_i\right) & \text{if } z_i = 4 \\
\Phi\left(\left[x_{j2} - \beta_i^{\mathcal{H}} + \epsilon_{ij}\right]/\lambda_i\right) & \text{if } z_i = 5 \\
\Phi\left(\left[\frac{x_{j2} - \tan\alpha_i^{\mathcal{D}} x_{j1} - \beta_i^{\mathcal{D}}}{\sqrt{\tan^2\alpha_i^{\mathcal{D}} + 1}} + \epsilon_{ij}\right]/\lambda_i\right) & \text{if } z_i = 6
\end{cases} \\
y_{ij} &\sim \text{Bernoulli}\left(\theta_{ij}\right)
\end{aligned}
$$

Fig. 3. Graphical model representation of the general recognition theory, as applied to the model individual categorization behavior for the Zeithamova and Maddox (2006) data.

uses and the criterial and perceptual noise for the trial on which the *j*th stimulus was presented. Our model assumes that the criterial and perceptual noise are combined into the value ε_{ij} which is drawn from a Gaussian distribution with mean 0 and a participant-specific standard deviation σ_i. Our model also assumes that the category bias δ is equal to 0, since the number of stimuli in each category are equal, and we expect people to be unbiased under these circumstances.

We assume there are, however, individual differences in the decision bounds that people use. In particular, we allow for simple unidimensional categorization strategies corresponding to strictly horizontal or vertical decision bounds, as well as more general diagonal bounds that involve both stimulus dimensions. This is implemented using a latent-mixture approach in which horizontal, vertical, and diagonal bounds are the mixture components. The parameter z_i functions as an indicator variable controlling which type of decision bound the *i*th participant uses. Because we use a *latent*-mixture approach, the z_i parameter is inferred for each participant. Depending on the type of decision bound, parameters that position that boundary in the stimulus space also need to be inferred. If the *i*th participant is inferred to use a horizontal bound, it is positioned at a spatial frequency value of $\beta_i^{\mathcal{H}}$. If they use a vertical bound, it is positioned at a spatial orientation value of $\beta_i^{\mathcal{V}}$. If they use a more general diagonal bound, it has a slope of $\alpha_i^{\mathcal{D}}$ and an intercept of $\beta_i^{\mathcal{D}}$.

For all of these possibilities, we assume that stimuli are probabilistically categorized according to which side of the decision boundary they lie. Stimuli closer to the boundary are categorized more probabilistically, with some probability they are categorized on the other side of the bound. Stimuli further from the boundary are categorized near deterministically. How quickly probabilistic categorization becomes deterministic is controlled by a participant-specific scale parameter λ_i, as part of a probit-link model of probabilistic responding.

The model in Fig. 3 also allows for three types of contaminant behavior, motivated by the clear presence of a group of participants exhibiting little learning and responding near chance, as discussed earlier. The first corresponds to the case in which a participant guesses, choosing category A and category B equally often regardless of the stimuli. The second corresponds to the case in which a participant almost always chooses category B regardless of the stimulus. The third corresponds to the case in which a participant almost always chooses category A regardless of the stimulus. Contaminant behaviors can be thought of as alternative response strategies to those coming from GRT, and so are naturally implemented by extending the latent-mixture approach (Zeigenfuse & Lee, 2010). Thus, overall, the parameter z_i indexes six possibilities for each participant: three possible GRT strategies based on different types of decision bounds, and three possible contamination strategies.

To complete the Bayesian implementation, we set equal prior probabilities on each participant using each of the six possible categorization strategies. We also set uniform prior distributions for the possible range of decision bound locations and for the noise variability and determinism parameters.

3.2. Categorization modeling results

We implemented the graphical model in Fig. 3 in JAGS (Plummer, 2003). Our results are based on six independent chains with 10,000 samples each after discarding the first 50,000 burn-in samples from each chain and thinning by collecting only every third sample. The chains were assessed for convergence using the standard $\hat{R}$ statistic (Brooks & Gelman, 1997).

The posterior distribution of the z_i parameter provides the probability that the ith participant is using each of the six possible strategies. We make the simplifying practical assumption that they use the most likely strategy, corresponding to the mode of the posterior distribution. Similarly, for the GRT-based strategies, we assume they use the decision bound given by the posterior mean of the relevant parameters for the horizontal, vertical, and diagonal cases.

These results are summarized in Fig. 4, which shows the 80 stimuli as points, colored by their true category. The shading of each point corresponds to the proportion of correct categorizations, with darker shades correspond to more accurate decisions. The decision bounds for the participants—36 in the unidimensional condition and 33 in the information-integration condition—inferred to be using GRT-based strategies are shown as gray lines. In the unidimensional condition, most participants use a vertical decision bound. However, in the information-integration condition, there is a group of participants who use a vertical decision bound and another group who use a diagonal decision bound. In addition to these individual differences in the type of decision bound, there are also individual differences in the location of the bounds themselves. For example, different participants use vertical decision bounds that correspond to different thresholds of spatial frequency.

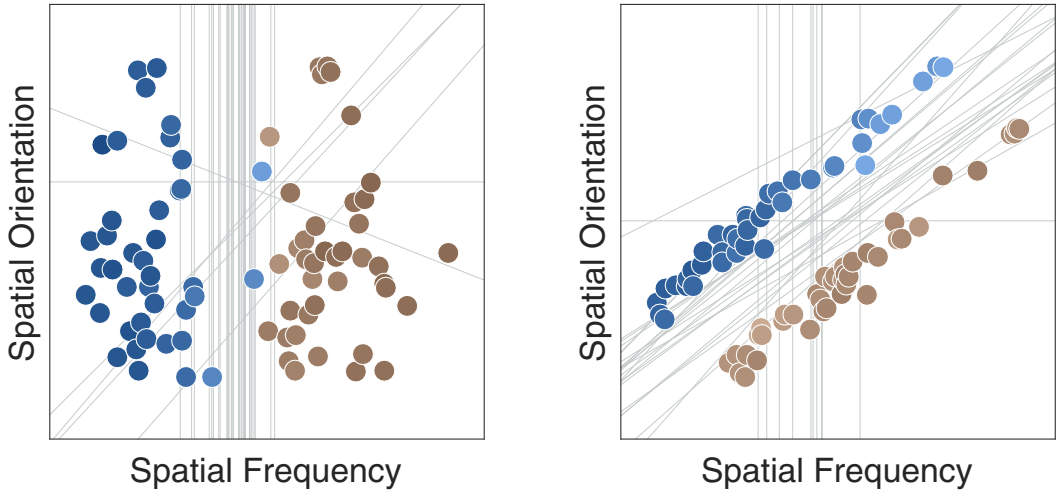


Fig. 4. Categorization behavior and inferred decision boundaries for the Zeithamova and Maddox (2006) data. The left panel corresponds to the "unidimensional without memory load" condition, and the right panel corresponds to the "information-integration without memory load" condition. In each panel, the true category structure is shown by the marker color, and the proportion of correct categorization decisions is shown by shading. The gray lines show the inferred decision bound for each participant.

Fig. 5 highlights the behavior of three selected participants from both the unidimensional and the information-integration conditions. Each of these participants corresponds to a pair of panels. The "observed" panels show the presented stimuli, with color corresponding to the categorization decision and the marker shape corresponding to the true category. The inferred boundary for the participant is shown by the gray line. This single boundary represents each participant's most likely boundary strategy, either vertical, horizontal, or diagonal, as well as the inferred location of that boundary in the stimulus space. The GRT is able to describe observed behavior to the extent that the decision bound separates the stimuli (i.e., that different colors lie on different sides of the bound). The selected participants vary as to whether they use a vertical bound, a diagonal bound, or one of the contaminant strategies.

The "new" panels in Fig. 5 show how the inferred strategies are applied to make predictions about how each participant would categorize the *newly generated* set of stimuli. For GRT-based strategies, new stimuli are simply categorized according to which side of the bound they lie. Otherwise, the stimuli are categorized according to the inferred contaminant strategy. It is these predictions that allow us to apply the wisdom of the crowd method to new stimuli.

3.2.1. Wisdom of the crowd results

Fig. 6 summarizes our wisdom of the crowd analysis for the unidimensional and information-integration conditions. For the observed stimuli, the gray bars show the distribution of the categorization accuracy across participants. This is based on their behavioral

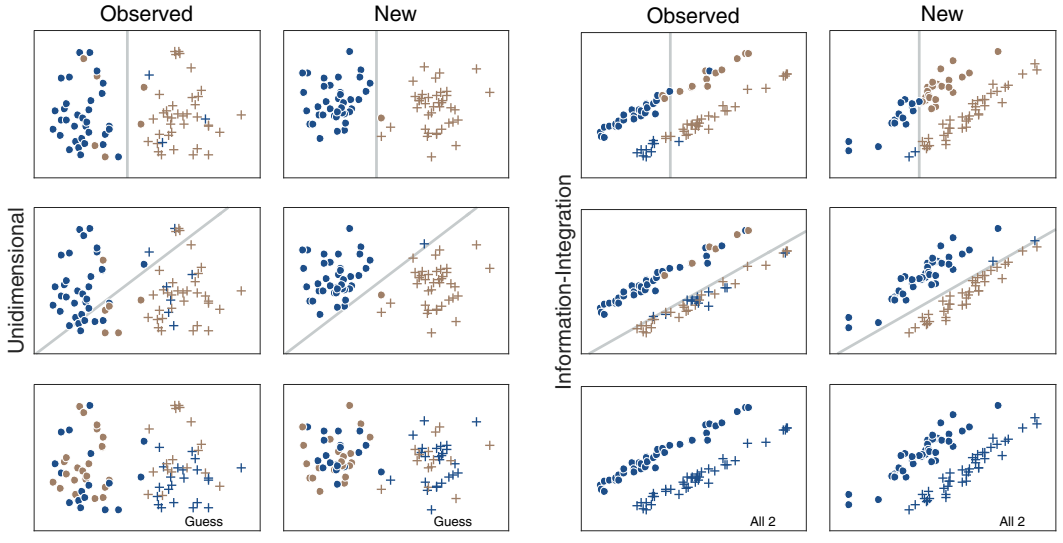


Fig. 5. Observed and new categorization decisions for a subset of the participants from the Zeithamova and Maddox (2006) dataset. Panels on the left correspond to the "unidimensional without memory load" condition, and panels on the right correspond to the "information-integration without memory load" condition. Rows correspond to individual participants, and columns correspond to observed and predicted behavior. In all of the panels, the marker shape represents the true category. In observed panels, the marker color represents participant behavior in the experiment. In the new panels, the marker color represents predicted behavior of the newly generated stimuli, based on the inferred categorization strategy. For both conditions, the first two participants are inferred to use the decision bounds shown by the gray line, while the third participant uses either a guessing or repetitive contaminant strategy.

responses in the experiment. The broken line shows the average of individual accuracy. The black dot shows the accuracy of the majority of these observed categorization decisions.

The "new" panels of Fig. 6 involve 1,000 sets of newly generated stimuli. For these new sets, the gray bars show the distribution of accuracy for the *predicted* categorization decisions across participants. The broken line is again the average accuracy. The black dot is the average accuracy of the majority predicted categorization across all of the new sets, and the error bar shows its 95-percentile range. There is more variability in the error bar of the information-integration condition because the inferred boundaries used are themselves more variable and prone to suboptimal behavior of using vertical boundaries, as seen in the right panel of Fig. 4.

The observed results mirror those presented in Fig. 2, showing that the majority decision is generally very accurate compared to individual performance. The similarly good performance for the new stimuli shows the effectiveness of the model-based approach. The majority of the predicted decisions, where the predictions are generated by models of individual categorization behavior, is able to categorize accurately stimuli that have never been observed.

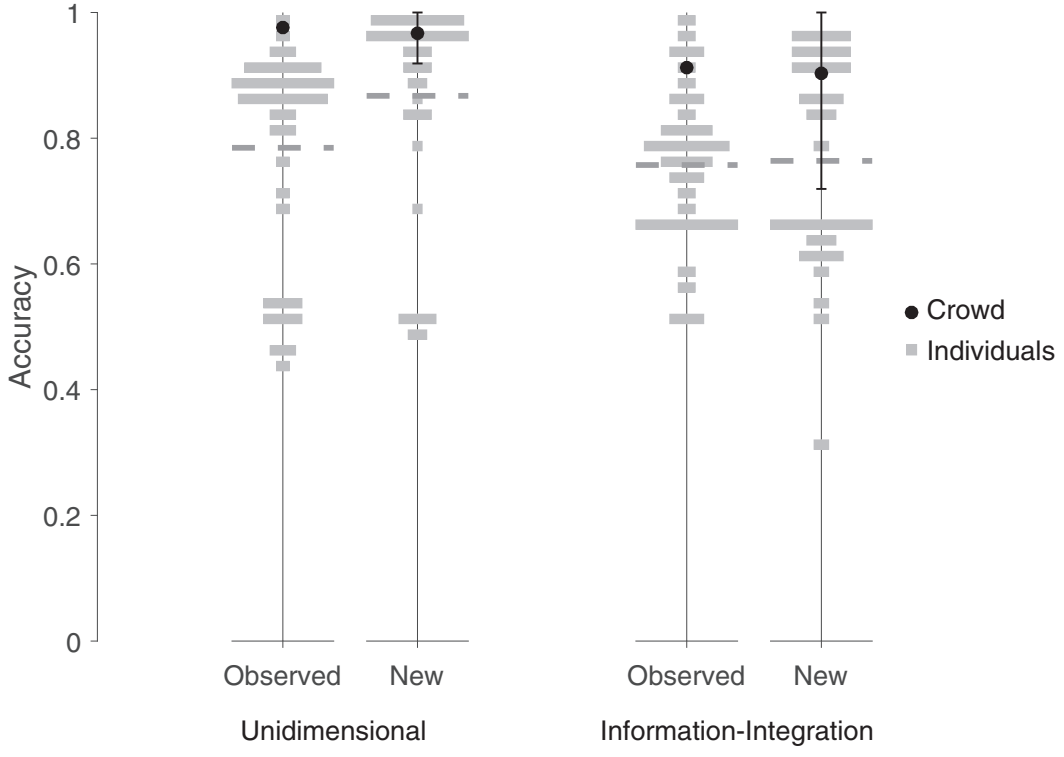


Fig. 6. The left panels show the individual and crowd accuracy for the Zeithamova and Maddox (2006) "unidimensional without memory load" condition. The right panels show the individual and crowd accuracy for the Zeithamova and Maddox (2006) "information-integration without memory load" condition. For both observed and new stimuli, the gray bars show the distribution of individual accuracy, the dashed line shows the average of these individual accuracies, and the black dot shows the accuracy of the majority crowd decision.

3.3. An application using the GCM

3.3.1. Stimuli and data

This application is based on one of the four conditions in the category learning experiment reported by Navarro et al. (2005). This experiment involved a set of 25 faces. The four conditions differed in the way these faces were assigned to two categories. We consider only the category structure that divided the faces in terms of hair color. In Fig. 2, this condition is the top-right panel in the "Navarro et al." section. In this condition, 10 participants completed eight testing blocks in which each stimulus was presented once with corrective feedback.

Fig. 7 shows each of the faces, labeled A–Y, in terms of their representation in a two-dimensional stimulus space. The space was derived using the individual-differences multidimensional scaling method presented by Okada and Lee (2016), based on previously collected similarity data involving 14 participants rating each pair of faces on

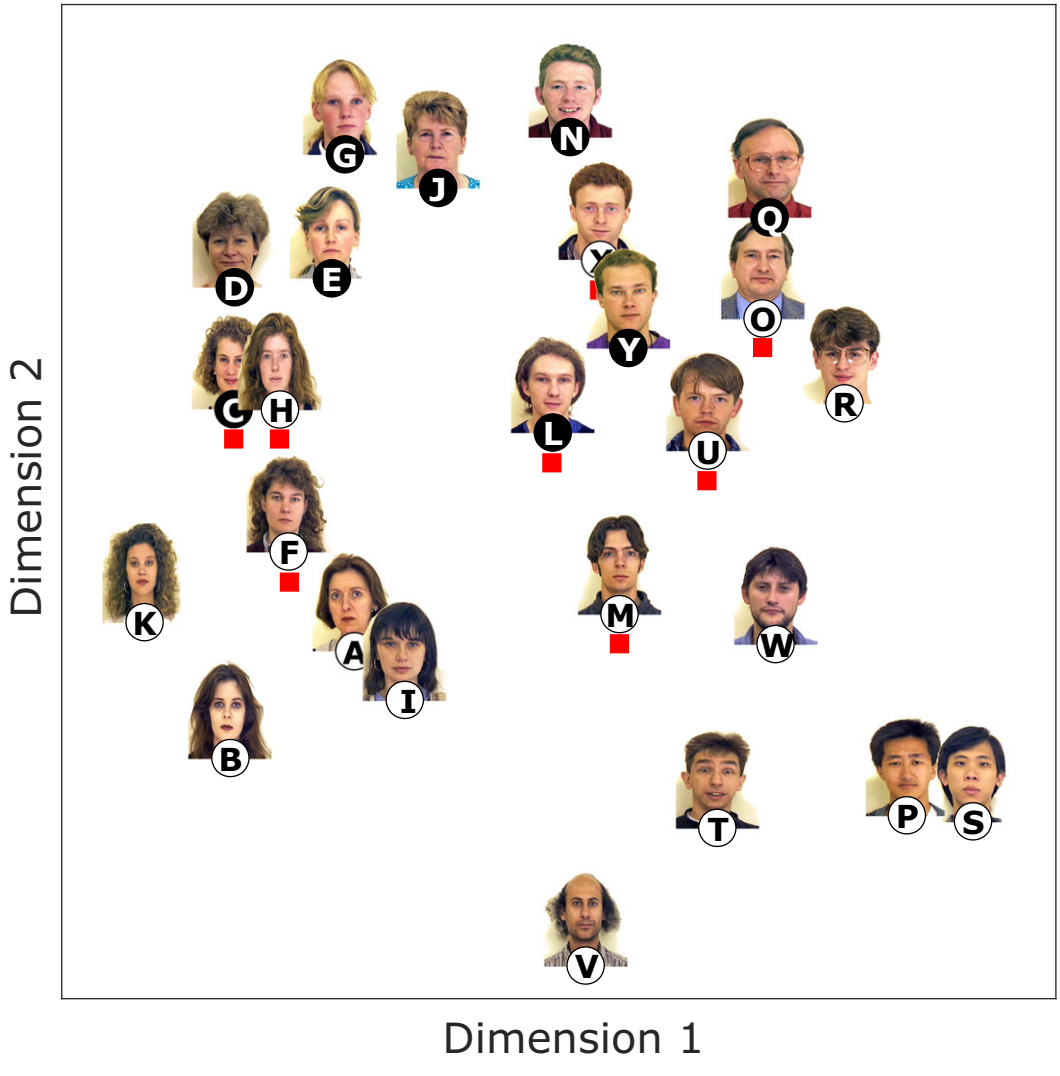


Fig. 7. Dimensional representation of the face stimuli from the Navarro et al. (2005) dataset. The true category structure is shown by filled and unfilled alphabetic labels. The red squares underneath the labels indicate the eight faces that were removed.

a 5-point scale. A key feature of this multidimensional scaling method is that it derives stimulus spaces with psychologically interpretable dimensions. The dimensions in Fig. 7 can be interpreted as corresponding to gender and hair color.

The category structure for the hair color condition is indicated in Fig. 7 by the black and white coloring of the stimulus labels. Unlike the Zeithamova and Maddox (2006) experiment, there is no rule for generating new stimuli with known category assignments. Accordingly, we removed eight faces from the Navarro et al. (2005) dataset. The

removed faces are highlighted in Fig. 7 by red squares beneath the stimulus labels. We treat these faces as if they were new stimuli, never seen by the participants.

3.3.2. GCM model

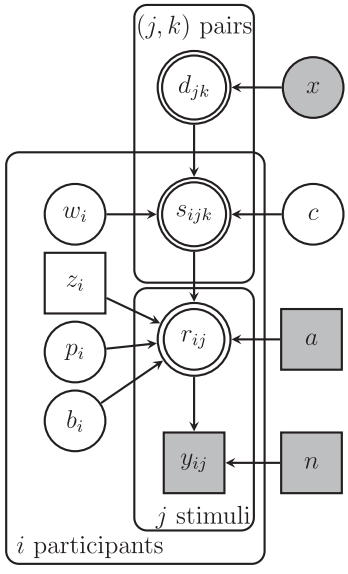
The GCM is an exemplar-based model of categorization that uses selective attention and similarity comparison processes to categorize stimuli. There are several variants of the GCM designed to accommodate specific categorization situations. These situations include some stimuli being presented more frequently than others, category assignment being inherently probabilistic, or the use of stimuli for which perceptual learning is possible (Nosofsky, 1992).

We use a variant of the GCM that we think is appropriate for the Navarro et al. (2005) categorization task. Formally, the ith stimulus is represented as a two-dimensional coordinate location $\boldsymbol{x}_i = (x_{i1}, x_{i2})$, as shown in Fig. 7. The attention-weighted distance between the ith and jth stimuli is then $d_{ij} = w \mid x_{i1} - x_{j1} \mid + (1 - w) \mid x_{i2} - x_{j2} \mid$, where w is a parameter controlling how much attention is given to the first dimension. This means that a dimension receiving more attention will be more influential in determining distances than the one receiving less attention. We assume that there may be individual differences in attention, and so there are individual-level w parameters. The similarity between these stimuli is $s_{ij} = \exp(-cd_{ij})$, where c is a parameter controlling the generalization gradient. Because we assume individual differences perceptual learning for the face stimuli are unlikely, the same generalization parameter is used for all participants. We do, however, allow c to vary over blocks, allowing for the possibility the degree of generalization is adapted to the learned category structures. The similarity of the ith stimulus to category A is then the sum of the similarities to all the stimuli in the category: $s_{iA} = \Sigma_{j \in A}\, s_{ij}$. Finally, the probability of a category response placing the ith stimulus in category A is $p_{iA} = bs_{iA}/(bs_{iA} + (1 - b)s_{iB})$, where b is a parameter controlling the response bias to category A. Because the categories are fixed, we do not include a response-determinism parameter in the category response model. We do, however, allow for individual bias, consistent with the assumption that there may be individual differences in the way participants learn the unequal category sizes.

3.3.3. Implementation

The GCM has also been implemented as a graphical model (Lee & Wagenmakers, 2013; Vanpaemel, 2009). Fig. 8 shows the implementation of the GCM that we applied to model individual categorization behavior for the Navarro et al. (2005) data. Unlike the GRT application, we consider every block in the category learning experiment. Since the GCM does not model learning, we did this by applying it cumulatively over the sequence of blocks.

In the graphical model, the y_{ij} node counts the number of times the ith participant categorizes the jth face into category A. The cumulative approach means that this count includes the current block as well as all previous blocks, and n counts how many times it has been presented over these blocks. Following the GCM, the category A response probability r_{ij} is determined from the similarities s_{ij} for the kth participant, which in turn are determined from the distances d_{jk} between the stimulus representations $\boldsymbol{x}$. The response



$$
\begin{aligned}
w_i &\sim \begin{cases} \text{Uniform}(0,1) & \text{if } z_i = 1 \\ 0 & \text{if } z_i = 2 \\ 1 & \text{if } z_i = 3 \end{cases} \\
c &\sim \text{Gaussian}(1,1)_{T(0,1)} \\
p_i &\sim \text{Uniform}(0,1) \\
b_i &\sim \text{Gaussian}(0.5,0.2)_{T(0,1)} \\
d^m_{jk} &= |x_{jm} - x_{km}| \\
s_{ijk} &= \exp\left\{-c(w_i d^1_{jk} + (1-w_i) d^2_{jk})\right\} \\
z_i &\sim \text{Categorical}\left(\tfrac{1}{6},\ldots,\tfrac{1}{6}\right) \\
r_{ij} &= \begin{cases} \frac{b_i \sum_j a_j s_{ijk}}{b_i \sum_j (a_j s_{ijk}) + (1-b_i)\sum_j (1-a_j) s_{ijk}} & \text{if } z_i = 1,2,3 \\ p_i & \text{if } z_i = 4 \\ 0.99 & \text{if } z_i = 5 \\ 0.01 & \text{if } z_i = 6 \end{cases} \\
y_{ij} &\sim \text{Binomial}(r_{ij}, n)
\end{aligned}
$$

Fig. 8. Graphical model representation of the generalized context model, as applied to the model individual categorization behavior for the Navarro et al. (2005) data.

probabilities depend upon individual bias b_i, and the similarities depend upon the generalization gradient c and individual attention weights w_i.

We assume there are individual differences in the attention weights that people use, and we give theoretical weight to the use of simple attention strategies that focus on just one stimulus dimension. We implement this using a latent-mixture approach in which the mixture components are the attention weight w_i values of 0, 1, or drawn from a uniform distribution. An attention weight of 0 corresponds to a person attending only to dimension 2 in Fig. 7. An attention weight of 1 corresponds to a person attending only to dimension 1 in Fig. 7. A person inferred to be using an attention weight drawn from a uniform distribution devotes attention to both dimensions, but possibly not equally. Similar to our GRT latent-mixture model, the z_i parameter functions as an indicator variable controlling which attention weight value the ith participant uses.

To complete the Bayesian implementation, we set a prior on the generalization gradient consistent with the distribution of distances in the MDS representation, and a prior for bias that corresponds to expecting any deviation from unbiased responding to be smaller rather than larger. We also set equal prior probabilities on each of the six possible categorization strategies.

3.3.4. Categorization modeling results

We again implemented the graphical model in JAGS. Our results are based on three independent chains with 1,000 samples each after discarding the first 5,000 burn-in samples from each chain. The chains were again assessed for convergence using the standard $\hat{R}$ statistic.

The top panel of Fig. 9 shows the most likely model for each participant in each block. It is clear that some participants change their attention weights over time, and that there are individual differences in these patterns of change. The most common attentional strategy is to attend just to the second dimension. The first two participants always attend to the second stimulus dimension, and the next two participants do the same from the second block onwards. Other participants show guessing contaminant behavior on many of the blocks. It is relatively rare for participants to attend to both dimensions. In general, the patterns of change are interpretable, such as participant 7 who initially attends to the first dimension, then distributes their attention for a few blocks, and finishes by guessing for the remainder of the experiment.

The bottom-left panel of Fig. 9 shows the inferred similarity gradients, over the distances in the MDS representation of the faces, based on the posterior of the c parameter. The histogram shows the distribution of distances between all pairs of faces. The eight

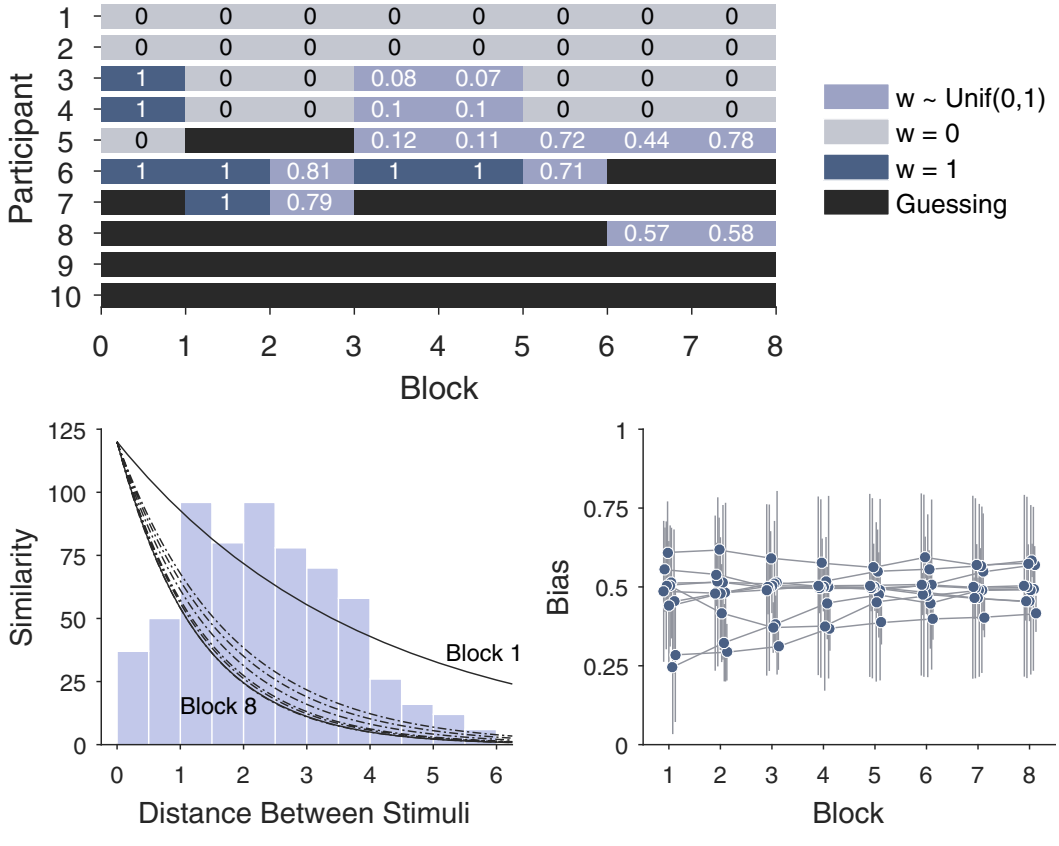


Fig. 9. Most likely model for all 10 participants for all eight blocks shown via color-coding (top panel). Inferred similarity gradients over all stimuli distances in the MDS representation (bottom-left panel). Inferred bias parameter for all 10 participants over the eight blocks (bottom-right panel).

gradients shown correspond to the eight blocks, and the first and last blocks are labeled. The gradient narrows over the course of the experiment, consistent with some form of adaptation. Comparing the gradients to the distribution of distances shows that in the first block, there is broad generalization from one face to all other faces, but in later blocks, what is known about one face generalizes only to relatively nearby faces. This is consistent with the principle of semi-distributed representation (Kruschke, 1993b).

The bottom-right panel shows the inferred pattern of change in bias for each of the 10 participants over the course of the experiment, with error bars representing 95% credible intervals. A few participants show some small initial bias, but the general result is that most participants on most blocks do not favor one category response over the other.

3.3.5. Wisdom of the crowd results

We used the GCM, with inferred individual differences in attention and bias, to make categorization predictions for each of the withheld faces from Fig. 7. As before, the prediction is the most likely category response and the crowd categorization decision is the majority of the individual-participant predictions. For each individual, we used the most likely strategy that the model inferred they were using to generate their categorization decision for the withheld stimuli. When that most likely strategy involved the GCM, we used attention and bias parameters corresponding to the inferred posterior mean for the individual. We then took the modal predicted categorization decision from the non-contaminant participants for each withheld face for each block of the experiment. This final step generated the crowd categorization decision.

Fig. 10 summarizes the wisdom of the crowd analysis. It shows the average categorization accuracy for the withheld faces for the individual participants (i.e., the categorization decisions actually made by the participants before we removed them for modeling

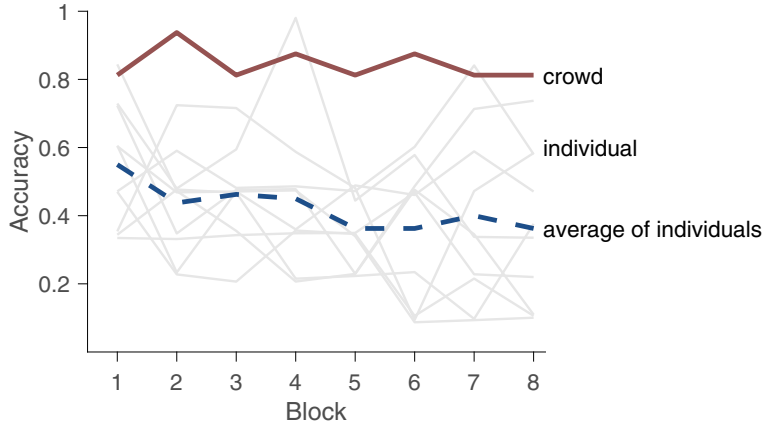


Fig. 10. Learning curves for the removed faces from the Navarro et al. (2005) dataset. The thin gray lines show each individual's proportion of correct answers for each block of the experiment. The dashed blue line shows the average of the individual participant accuracies. The single thick red line shows the categorization accuracy of the aggregated crowd majority decision.

purposes), the average of these individual accuracies, and the crowd category decisions. It is clear that from the first block, the crowd is more accurate than any individual and maintains this superiority over the subsequent blocks. The crowd is always more accurate than the average of the individuals. The performance of the aggregate decision is especially impressive, given the difficulty of the withheld faces in terms of the category structure, as evidenced by the average of individual accuracy decreasing even with feedback.

4. Discussion

In this paper, we have developed and demonstrated a wisdom of the crowd approach to categorization. The basic idea is to use the majority categorization decision over a set of individuals as the crowd decision. We showed that this approach leads to accurate crowd decisions for a number of existing category learning datasets, varying widely in the size of the crowd, the difficulty of the category structures, and the nature of the stimuli. We also developed a model-based extension of this idea, using categorization models that allow for individual differences in categorization behavior. We showed that individual-level models can be inferred from available categorization decisions and then used to predict how that individual would categorize an unseen stimulus. Our results show that the majority of these predictive decisions continues to produce relatively accurate crowd decisions.

The two case studies we presented highlight the potential generality of the approach. One involved GRT and decision-bound categorization, while the other involved the GCM and similarity-based exemplar categorization. One involved low-level perceptual Gabor patch stimuli, while the other involved more complicated and holistic face stimuli. One focused on individual differences in the form of different decision strategies, such as horizontal, vertical, and diagonal decision bounds, while the other focused on individual differences in the form of selective attention to different stimulus dimensions. The basic approach simply needs a predicted decision for each individual for a new stimulus, and any model of categorization decisions and individual differences is potentially applicable.

The particular versions of the GRT and GCM we used worked effectively, but we do not claim they are the best possible models. In both case studies, we made a number of modeling decisions, about the inclusion or exclusion of parameters in the GRT and GCM, about the existence of contaminant subgroups, and so on. These decisions usually had some basis in theory or the specific nature of the category learning task. For example, we allowed the generalization parameter in the GCM to vary across blocks but not across people, because that would imply some individual variation in perceptual learning over the course of the experiment, which we think is unlikely for the face stimuli. Similarly, our GRT model did not include a category bias parameter, because the number of Gabor stimuli in each category was equal, consistent with what we would expect participants to assume, but we did include such a parameter in the GCM model, because the number of face stimuli in each category was unequal, and we expect individual differences in participants learning this imbalance.

Despite these sorts of justifications, however, it would be possible to explore a large number of alternative GRT and GCM models by combinatorially varying the assumptions we made. This would be interesting theoretically to test which assumptions are key to good performance, and useful practically to optimize performance. We noted some interesting possibilities in constructing our case studies but did not attempt a systematic investigation. For example, in the GCM analysis, we observed that crowd performance was significantly worse before we included the contaminant behavior mixture components. Without allowing for these individual differences, the crowd performance did not go above an accuracy of 75%. Removing contaminants reduces the number of decisions contributing to the majority, but evidently this deficit is more than compensated by identifying those participants who are learning the category structure. In this case, the additional theoretical complicated of including contaminant behavior was worthwhile. It might also be that sometimes a simpler model is a better account of people's behavior, and improves performance. For example, even though the possibility of individual differences in category bias for the GCM case study was well motivated, the inferences in Fig. 9 suggest that assuming unbiased responding for all participants might describe the data well, and it could potentially lead to better crowd performance. Exploring these sort of possibilities is an interesting direction for future research. It will be challenging territory to navigate, because of possible tensions between modeling assumptions that follow from established theory, those that are required to describe the current behavioral data, and those that best achieve the applied goal of crowd accuracy. Ultimately, we need to understand potentially complicated relationships between the quality of a cognitive model of individual categorization behavior, the quality of a model of individual differences in that behavior, and the quality of the crowd performance it underpins.

Moving forward, one attraction of our approach is its generality. It is possible for both the current case studies, and for other case studies—involving other stimuli or category structures—that quite different categorization models will be appropriate. For example, some category structures will need nonlinear decision bounds in the GRT, and individual differences in generalization gradients in the GCM will be needed for stimuli that allow for perceptual learning. Beyond the GRT and GCM there are many other theories and models of categorization, including ALCOVE, COVIS, RULEX, SUSTAIN, and hybrid models (Ashby, Alfonso-Reese, Turken, & Waldon, 1998; Busemeyer, Dewey, & Medin, 1984; Kruschke, 1992; Love, Medin, & Gureckis, 2004; Nosofsky & Palmeri, 1998; Smith & Minda, 2000), that could be used as the underlying psychological models in our wisdom of the crowd framework.

As well as considering other models of categorization, our approach would benefit from extended models of categorization decisions and category learning. For example, it is possible that people change strategies during the course of learning a category structure. In the GCM case study, Fig. 9 shows that some participants change how they attend to the different stimulus dimension over the course of learning. These changes are not formally part of the GCM model that we used, nor is it a common capability in most established psychological models of category learning. It would also be possible to extend the modeling approach to allow for individual differences in terms of which

psychological model each person uses. It may be that some people use decision bounds while others use exemplar-based similarity, and it may even be that some people start with an exemplar strategy and switch to a decision bound strategy as the number of presented stimuli increases. Both of these extensions could be naturally accommodated by hierarchical and latent-mixture extensions within the graphical modeling approach we have used and could continue to be applied to data using Bayesian methods.

Turning to applied possibilities, a challenge for our approach is determining how to represent the stimuli. The simple perceptual nature of the Gabor patch stimuli is not true of all stimuli, and the representation of the faces that we used was based on independent similarity data collection and multidimensional scaling analysis. Even then, the representation of the face stimuli only applied to 25 faces, and we have no method for determining how a new face should be represented in this same space. What is needed for real-world application is a formal method for determining an appropriate representation of any possible stimulus. To return to our motivating example of doctors learning to diagnose skin patches, it seems possible, but far from trivial, that image processing methods could automatically map a visual skin patch stimulus into a dimensional psychological representation. In general, the applicability of our approach to real-world situations hinges on finding such a representational method. When such methods are available, our approach has the attractive property of requiring relatively limited effort on the part of people to categorize large numbers of stimuli. Once a categorization model has been inferred for each individual, it can be applied to any number of new stimuli. The accuracy of the crowd categorizations should increase as both more individuals are included in the group, and as individuals categorize more stimuli.

One way to interpret our wisdom of the crowd approach comes from machine learning where it would be called boosting (Hastie, Tibshirani, & Friedman, 2001). Under this interpretation, the model of each individual functions as a weak classifier and there is a simple majority rule for aggregation of the categorization decision. In fact, the GRT is closely related to decision-bound methods like support vector machines, and the GCM is closely related to radial basis classifiers, nearest-neighbor, and other clustering methods (Gomes, Welinder, Krause, & Perona, 2011; Welinder, Branson, Belongie, & Perona, 2010). From a machine learning perspective, the contribution of our approach is to help identify useful weak classifiers, by recognizing that the classification problem is a problem of human categorization, and so domain-specific cognitive models should be effective in ways that more domain-general statistical methods may not. Nevertheless, it is almost certainly possible to improve categorization accuracy in the case studies we have presented using established and successful machine learning techniques. In particular, it is likely that discriminative machine learning methods could outperform the generative approach to probabilistic modeling we have used. The strength of the psychological nature of our approach comes not from relative accuracy, but from significantly greater interpretability. A recognized challenge for machine learning methods relates to issues of interpretability and trust (Ribeiro, Singh, & Guestrin, 2016). While a deep neural net may only be able to give a sequence of connection weights as a justification for a decision, it is generally easy to give complete and meaningful accounts of how and why our

aggregated crowds decided to categorize a new stimulus a certain way. These explanations will reference interpretable decision strategies and individual differences in those strategies. This should not only increase the probability that people trust the crowd decision, but also make training and remediation of individuals possible, especially by comparing their categorization strategies to others.

In terms of psychological understanding, our approach is a good example of what Watts (2017) calls "solution-oriented" social science. The general goal is to seek to solve a practical problem, using existing theories and models where possible, and identifying gaps where they exist. In our case, the wisdom of the crowd problem demands that the modeling of individual differences be taken seriously, and both of our case studies incorporated different sorts of categorization strategies as well as allowing for different types of contaminant behavior. This is relatively new theoretical territory for the modeling of human category learning, and there certainly is not wide exploration or agreement on the number and type of these individual differences. In this way, our results provide new empirical evaluation and are relevant to the development of theory. More tellingly, the results for the faces case study, shown in Fig. 10, identify the need for models of how people change or adapt their categorization strategies over the course of learning. There are few such theories, and no established categorization models that include this capability. In these sorts of ways, our case studies not only demonstrate the applicability of current categorization models to have a useful real-world application, but highlight the role of applications in focusing attention of important theoretical and modeling problems that need to be solved to understand how categorization works.

Acknowledgments

There is a project page on the Open Science Framework associated with this paper at https://osf.io/j95q6/ providing data, code, and other supplementary information. We thank Mike Kalish for helpful conversations, and Stephan Lewandowsky for sharing data. This work was supported by NSF Award 1461365.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldon, E. M. (1998). A neuropsycholoigcal theory of multiple systems in category learning. *Psychological Review*, *105*, 442–481.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*, 154–179.

Bartlema, A. (2013). Selective attention in category learning. Unpublished raw data.

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, *59*, 132–150.

Brooks, S. P., & Gelman, A. (1997). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Busemeyer, J. R., Dewey, G. I., & Medin, D. L. (1984). Evaluation of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 638.

Danileiko, I., Lee, M. D., & Kalish, M. (2015). A Bayesian latent mixture approach to modeling individual differences in categorization using General Recognition Theory. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 501–506). Austin, TX: Cognitive Science Society.

Gomes, R. G., Welinder, P., Krause, A., & Perona, P. (2011). Crowdclustering. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (pp. 558–566). Red Hook, NY: Curran Associates, Inc.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, *112*, 494–508.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer Verlag.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*, 231–237.

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

Kruschke, J. K. (1993a). Human category learning: Implications for backpropagation models. *Connection Science*, *5*, 3–36.

Kruschke, J. K. (1993b). Three principles for models of category learning. *The Psychology of Learning and Motivation*, *29*, 57–90.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1–7.

Lee, M. D. (in press). Bayesian methods in cognitive modeling. In J. T. Wixted, (Ed.), *The Stevens Handbook of Experimental Psychology and Cognitive Neuroscience* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review*, *9*, 43–58.

Lee, M. D., Steyvers, M., & Miller, B. J. (2014). A cognitive model for aggregating people's rankings. *PLoS ONE*, *9*, 1–9.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge: Cambridge University Press.

Lee, M. D., Zhang, S., & Shi, J. (2011). The wisdom of the crowd playing the Price is Right. *Memory & Cognition*, *39*, 914–923.

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 720.

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

Navarro, D. J., Lee, M. D., & Nikkerud, H. (2005). Learned categorical perception for natural faces. In B. G. Bara, L. W. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 1600–1605). Mahwah, NJ: Erlbaum.

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

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, *43*, 25–53.

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimensionspaces. *Psychonomic Bulletin & Review*, *5*, 345–369.

Okada, K., & Lee, M. D. (2016). A Bayesian approach to modeling group and individual differences in multidimensional scaling. *Journal of Mathematical Psychology*, *70*, 35–44.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing*. Vienna, Austria.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). New York, NY: ACM.

Selker, R., Lee, M. D., & Iyer, R. (2017). Thurstonian cognitive models for aggregating top-n lists. Decision. *Decision*, *4*, 87–101.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *24*, 1411–1436.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 3.

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision-making. *Psychological Review*, *108*, 183–203.

Surowiecki, J. (2004). *The wisdom of crowds*. New York: Random House.

Vanpaemel, W. (2009). BayesGCM: Software for Bayesian inference with the generalized context model. *Behavior Research Methods*, *41*, 1111–1120.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*, 645–647.

Watts, D. J. (2017). Should social science be more solution-oriented? *Nature Human Behaviour*, *1*, 1–5.

Welinder, P., Branson, S., Belongie, S. J., & Perona, P. (2010). The multidimensional wisdom of crowds. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems* (vol. 23, pp. 2424–2432).

Yi, S. K. M., Steyvers, M., Lee, M. D., & Dry, M. J. (2012). The wisdom of the crowd in combinatorial problems. *Cognitive Science*, *36*, 452–470.

Zeigenfuse, M. D., & Lee, M. D. (2010). Finding the features that represent stimuli. *Acta Psychologica*, *133*, 283–295.

Zeithamova, D., & Maddox, W. T. (2006). Dual task interference in perceptual category learning. *Memory & Cognition*, *34*, 387–398.