Supplemental Materials for

Algorithmic personalization of information can cause inaccurate generalization and overconfidence

Giwon Bahg, Vladimir M. Sloutsky, Brandon M. Turner

Corresponding Author: Giwon Bahg (giwon.bahg.1@gmail.com)

This PDF file includes:

Supplemental Text
Supplemental Figures S1 to S10
Supplemental Tables S1 to S2

Supplemental Text

Personalization Algorithm

User Profile

Personalization methods require numerical expressions of users and their behavioral characteristics. In this study, we defined a user profile as a vector consisting of dimension-wise sampling probabilities, denoted $\mathbf{u} = (u_1, \cdots, u_D)^T$ where u_d represents the sampling probability of the d-th dimension and a superscript T is a transpose operator.

The item-level personalization algorithm (see Section "Item-level Personalization") updates the user profile every trial using Bayes' theorem assuming a beta prior and a binomial likelihood (Kruschke, 2014). The algorithm treats the dimension-wise sampling probability as a beta-distributed random variable U_d and uses its summary statistic to define the user profile:

$$U_d \sim \text{Beta}(\beta_{1d}, \beta_{2d}) \tag{1}$$

with a probability density function

$$\pi(U_d = u_d | \beta_{1d}, \beta_{2d}) = \frac{\Gamma(\beta_{1d} + \beta_{2d})}{\Gamma(\beta_{1d})\Gamma(\beta_{2d})} u_d^{\beta_{1d}-1}(1-u_d)^{\beta_{2d}-1}$$

and the expected value

$$\mathbb{E}[U_d] = \frac{\beta_{1d}}{\beta_{1d} + \beta_{2d}}. \tag{2}$$

The user profile is defined as a vector consisting of the expected dimension-wise sampling probabilities:

$$\mathbf{u} = \big(\mathbb{E}[U_1], \cdots, \mathbb{E}[U_D]\big)^T. \tag{3}$$

Also, a sampling decision for the d-th dimension is modeled as a Bernoulli random variable with the success probability u_d:

$$R_d \sim \text{Bernoulli}(u_d) \tag{4}$$

with a probability density function

$$\pi(R_d = r_d | u_d) = u_d^{r_d}(1-u_d)^{1-r_d}$$

where $r_d = 1$ if a participant encoded the d-th dimension and $r_d = 0$ otherwise.

At the beginning of each session, the algorithm initializes U_d using a Beta$(1, 1)$ prior for each dimension (i.e., $\beta_{1d} = \beta_{2d} = 1$). This prior is equivalent to a continuous uniform distribution defined at $[0, 1]$ and makes $\mathbb{E}[U_d] = 0.5$. Once a participant finishes feature sampling, the algorithm interprets the sampling decision as a Bernoulli response: $r_d = 1$ if the feature was sampled on Dimension d, and $r_d = 0$ otherwise. The algorithm then updates U_d using the following beta-binomial updating rule: For $d = 1, \cdots, D$,

$$\beta_{1d} \leftarrow \begin{cases} \beta_{1d} + 1 & \text{if } r_d = 1, \\ \beta_{1d} & \text{if } r_d = 0; \end{cases}$$

$$\beta_{2d} \leftarrow \begin{cases} \beta_{2d} & \text{if } r_d = 1, \\ \beta_{2d} + 1 & \text{if } r_d = 0. \end{cases}$$

The user profile $\mathbf{u}$ is updated according to the changes in β_{1d} and β_{2d} using Equation 2.

Item-level Personalization

We implemented an item-level personalization algorithm by adapting the method proposed by Covington et al. (2016) to our task setting. The original algorithm is a collaborative filtering algorithm using a deep neural network that aims to generate suggestions in an online video-sharing platform. The neural network learns to compress a user profile vector so that the dot-product similarity between the compressed user vector and vectors representing video clips explains the watch probability.

Our version of the recommendation algorithm aimed to recommend items with higher expected sampling rates. It means that our algorithm preferred items that lead learners to sample as many features as possible. This goal is analogous to that of Covington et al. (2016), which aimed to search for video clips with higher watch probabilities.

Our algorithm uses each user's sampling profile as a user vector (denoted $\mathbf{u}$) and each item's feature vector as an item vector (denoted $\mathbf{x}$). The basic mechanism of the recommendation architecture remains the same as collaborative filtering based on a transformation of user profiles. We further simplified the algorithm as follows: First, the algorithm simply does a linear transformation of the user profile, instead of nonlinear dimensionality reduction implemented by a deep neural network. There is no need for dimensionality reduction because $\mathbf{u}$ and $\mathbf{x}$ share the same dimensionality. Also, nonlinear transformations might be unnecessarily costly, considering the simplicity of our settings. The second modification is the loss function. The method by Covington et al. (2016) approached personalization as classification to predict the watch probability. We approached the same problem of personalization as binomial regression to predict the number of sampled dimensions.

Our user embedding model can be described as follows: Let $Y_i \in \{0, 1, 2, \cdots, D\}$ and $\mathbf{x}_i$ represent the number of sampled features and the feature vector of the i-th item, respectively. Also, let $\mathbf{B} \in \mathbb{R}^{D \times D}$ be a linear mapping that transforms a user profile $\mathbf{u}$. Then, a model for predicting the number of sampled features is a binomial regression model

$$
\begin{aligned}
P(Y_i = k | \mathbf{x}_i, \mathbf{B}, \mathbf{u}) &= \binom{D}{k} \mu_i^k (1 - \mu_i)^{D-k}, \\
\text{where } \mu_i &= \frac{1}{1 + \exp(-\mathbf{x}_i^T \mathbf{B} \mathbf{u})}.
\end{aligned}
$$

Assuming that we train the item-level personalization model using N_X items, the log-likelihood of B is derived as follows:

$$
\begin{aligned}
l(B | Y, X, \mathbf{u}) &= \log P(Y | X, \mathbf{B}, \mathbf{u}) \\
&= \log \left\{ \prod_{i=1}^{N_X} P(Y_i = y_i | \mathbf{x}_i, \mathbf{B}, \mathbf{u}) \right\} \\
&= \sum_{i=1}^{N_X} \left\{ \log \binom{D}{y_i} + y_i \log \mu_i + (D - y_i) \log(1 - \mu_i) \right\} \\
&= \sum_{i=1}^{N_X} \left[y_i \log \mu_i + (D - y_i) \log(1 - \mu_i) \right] + (\text{irrelevant constant})
\end{aligned}
$$

where $Y = \{Y_1, \cdots, Y_{N_X}\}$ and $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_{N_X}\}$ are collections of the number of sampled features and presented items, respectively. The best linear mapping is found by maximizing this log-likelihood, i.e.,

$$\hat{\mathbf{B}} = \arg\max_{\mathbf{B}} l(\mathbf{B}|Y, X, \mathbf{u})$$

In practice, we used the Nelder-Mead method with a relative convergence tolerance of 10^{-12} to find $\hat{\mathbf{B}}$.

In the original method for online video sharing (Covington et al., 2016), previously collected data (i.e., the profiles of individual users and their watch history) informed the neural network to compress user profiles and generate recommendations. Similarly, our experiment collected data from the active learning condition and used this dataset to train the mapping variable $\mathbf{B}$. Once we obtained the estimate $\hat{\mathbf{B}}$, we applied it to all three personalized learning conditions to generate learning sequences.

Generating Candidate Items

On each trial, we generated personalized learning sequences using the trained linear mapping $\hat{B}$. Let $V = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{N_V}\}$ $(N_V = 32)$ be a complete set or 'corpus' of items that can be presented during the learning phase. The expected sampling probability for all items in the corpus was computed as follows:

$$\hat{\mu}_i = \frac{1}{1 + \exp(-\mathbf{v}_i^T \hat{\mathbf{B}} \mathbf{u})}.$$

Among 32 items, the algorithm chose 12 items with the highest values of $\hat{\mu}_i$ as a candidate set. The item to be presented was determined by multinomial sampling using $\hat{\mu}_i$'s of the candidate items as sampling weights after normalization. That is, for the t-th trial,

$$\begin{aligned} \mathbf{x}_t &\sim \text{Multinomial}(\hat{\mu}_1^*, \hat{\mu}_2^*, \cdots, \hat{\mu}_{N_V}^*) \\ \text{where } \hat{\mu}_i^* &= \begin{cases} \hat{\mu}_i / \sum_{j \text{ for all } \mathbf{v}_j \in (\text{Candidate set})} \hat{\mu}_j & \text{if } \mathbf{v}_i \in (\text{Candidate set}); \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Feature-level Personalization

In Condition PP, the experiment uncovered the features of a presented item in descending order of the dimension-wise sampling probabilities in the user profile $\mathbf{u}$ (Equation 3). To break any possible ties within the user profile, a small number was randomly sampled from Uniform(-ϵ, ϵ) and added to each element of the user profile. The value of ϵ was small enough so that it would only break the tie but not disrupt other well-established orders. The randomly added numbers as a tiebreaker were not considered when updating user profiles.

Methods

Confusion Matrix Analysis: Other Performance Measures

In the main text, we presented the confusion matrix analysis to compare the accuracy of category representations between experimental conditions. The measure used in the main text was

the square-root sum of squared errors evaluated between the ideal confusion matrix (i.e., only its diagonal elements are non-zero) and the actual confusion matrix observed from participants. This measure is not commonly used in interpreting confusion matrices but served our purpose of comparing the details of inaccurate categorization represented in off-diagonal elements.

In the Supplemental Material, we discuss the same representational accuracy analysis but using other performance measures frequently used when analyzing confusion matrices. In particular, we focus on precision, recall, and overall accuracy.

Let's say that we have a classification task for C categories ($C > 2$), and a confusion matrix $M = [m_{ij}]_{C \times C}$ is defined by elements $\{m_{ij}\}$ representing the number of items from Category j classified to Category i ($i, j \in \{1, \cdots, C\}$). Category-wise precision (PR_c) and recall (RC_c) for Category c are defined as the number of correct categorization divided by the total number of predicted classification to Category c or the total number of Category c items, respectively. Overall accuracy (ACC) is defined as the number of correct categorization across all categories divided by the total number of presented items (Powers, 2011). Mathematical definitions of these terms are as follows:

$$PR_c := \frac{m_{cc}}{\sum_{j=1}^{C} m_{cj}}, \quad RC_c := \frac{m_{cc}}{\sum_{i=1}^{C} m_{ic}}, \quad ACC := \frac{\sum_{c=1}^{C} m_{cc}}{\sum_{i=1}^{C} \sum_{j=1}^{C} m_{ij}}.$$

There are two ways of summarizing category-wise precision and recall across multiple categories: macro-averaging and micro-averaging (Manning et al., 2008). Micro-averaging is obtained by summing the numerators and denominators across all categories and dividing the summed numerator by the summed denominator. Macro-averaging is obtained by averaging category-wise performance measures. Two types of measures (precision, recall) and two types of averaging schemes (micro-averaging, macro-averaging) allow us four possible summary measures.

In our experiment, category-wise precision may not be defined in some categories if participants never classified any item into these categories. Macro-average precision requires category-wise precision for all categories, and therefore, is not suitable for our analysis.

Meanwhile, micro-average precision (mPR), micro-average recall (mRC), and overall accuracy (ACC) are equivalent:

$$\begin{aligned} mPR &= \frac{\sum_{i=1}^{C} m_{ii}}{\sum_{i=1}^{C} (\sum_{j=1}^{C} m_{ij})} = \frac{\sum_{i=1}^{C} m_{ii}}{\sum_{i=1}^{C} \sum_{j=1}^{C} m_{ij}} = ACC, \\ mPR &= \frac{\sum_{i=1}^{C} m_{ii}}{\sum_{i=1}^{C} (\sum_{j=1}^{C} m_{ij})} = \frac{\sum_{j=1}^{C} m_{jj}}{\sum_{j=1}^{C} (\sum_{i=1}^{C} m_{ij})} = mRC. \end{aligned}$$

Also, if all categories were shown N/C times as in our case, macro-average recall (MRC) is equivalent to micro-average recall, micro-average precision, and overall accuracy:

$$\begin{aligned} MRC &= \frac{1}{C}\left(\sum_{i=1}^{C} RC_c\right) = \frac{1}{C}\sum_{c=1}^{C}\left(\frac{m_{cc}}{\sum_{i=1}^{C} m_{ic}}\right) \equiv \frac{1}{C}\sum_{c=1}^{C}\left(\frac{m_{cc}}{N/C}\right) \\ &= \frac{\sum_{c=1}^{C} m_{cc}}{C(N/C)} = \frac{\sum_{c=1}^{C} m_{cc}}{N} \\ &\equiv \frac{\sum_{c=1}^{C} m_{cc}}{\sum_{i=1}^{C} \sum_{j=1}^{C} m_{ij}} = mRC = mPR = ACC. \end{aligned}$$

Therefore, the summary performance metrics we can derive based on precision and recall are equivalent to overall accuracy, which we will use in the analysis.

The F_1 score also summarizes precision and recall by taking their harmonic mean. However, the macro-average F_1 cannot be defined due to the absence of macro-average precision in some datasets. The micro-average F_1 is equivalent to accuracy in our case (Manning et al., 2008).

When evaluating accuracy, we hypothesized that the control condition would have higher performance measure values than other conditions. The procedure for testing this hypothesis was identical to that in the main text except for the direction of the hypothesis (i.e., the control condition has a shorter representational distance and higher accuracy than other conditions).

Same-different Task: Quantifying the Representational Accuracy

The post-learning phase of the experiment included the same-different task in which participants evaluated the degree to which they believed two stimuli presented on the screen came from the same category. We hypothesized that the reported degree of belief reflects the latent category representation developed throughout the learning phase. The deviation of this latent representation from the ground truth (i.e., the similarity between categories computed from the feature values) would explain the influence of personalization or active learning on how participants learned about categories.

We quantified the latent category representation of participants using the same-different task data. The responses (i.e., the position on the slider) were translated into integer values in $[0, 100]$ where smaller numbers mean a stronger belief that the stimuli came from the same category. An (8×8) distance matrix was then constructed whose (i, j) element was the distance value evaluated between categories i and j, excluding all the diagonal elements that were left as zero. We computed the two-dimensional MDS solution from this distance matrix using the R function `cmdscale`. The MDS solution of category centroids served as a reference with which to compare.

We evaluated the representational accuracy of each participant by computing the RV coefficient (Robert & Escoufier, 1976) between the MDS representation from the same-different task responses and that of category centroids. The RV coefficient was proposed as a method to evaluate the dependency between two random variables coming from the same observation but possibly with different dimensionalities. Let $\mathbf{X}_1 \in \mathbb{R}^{n \times p}$ and $\mathbf{X}_2 \in \mathbb{R}^{n \times q}$ be two matrices consisting of p- or q-dimensional random vectors collected from the same n observations. The RV coefficient is computed as follows:

$$RV(\mathbf{X}_1, \mathbf{X}_2) = \frac{\text{tr}(\mathbf{X}_1 \mathbf{X}_1^T \mathbf{X}_2 \mathbf{X}_2^T)}{\sqrt{\text{tr}(\mathbf{X}_1 \mathbf{X}_1^T)^2 \text{tr}(\mathbf{X}_2 \mathbf{X}_2^T)^2}}$$

where $\text{tr}(\mathbf{X}) = \sum_{i=1}^{n} x_{ii}$ is the operator that calculates the trace of a matrix. This quantity evaluates the linear relationship between two random variables like typical correlation coefficients due to the following properties (Josse & Holmes, 2016; Josse et al., 2008). First, $RV(\mathbf{X}_1, \mathbf{X}_2) = 0$ if and only if $\mathbf{X}_1^T \mathbf{X}_2 = \mathbf{0}_{p \times q}$, i.e., a $(p \times q)$ matrix filled with zeros. Second, $RV(\mathbf{X}_1, \mathbf{X}_2) = 1$ if $\mathbf{X}_2 = a\mathbf{R}\mathbf{X}_1 + \mathbf{c}$ for $a \in \mathbb{R}$, $\mathbf{R}^T\mathbf{R} = \mathbf{R}\mathbf{R}^T = \mathbf{I}$ (i.e., an identity matrix), and $\mathbf{c} \in \mathbb{R}^n$. The second property means that the correlation between the two random variables is one if one can be seen as a linear transformation of the other.

As suggested above, the RV coefficient is always within the range of $[0, 1]$ and is invariant with overall scaling, orthogonal rotation, dimension-wise sign flip, and shift. These invariance

properties were ideal for evaluating the dependency between the category centroids and the latent category representation of participants. Although the original RV coefficients have been reported to overestimate the dependency between variables (Smilde et al., 2009), this was not a critical issue to us because our interest was in the comparison between the control and non-control conditions. We used the R package `MatrixCorrelation` (Indahl et al., 2018) to compute the RV coefficient.

Same-different Task: Representational Accuracy Analysis

To compare the degree of representational distortion between the control and non-control conditions, we used a Bayesian adaptation of the Wilcoxon rank sum test (van Doorn et al., 2020). The basic procedure is the same as the confusion matrix analysis (see the "Representational Accuracy Analysis" section in the Method section of the main text).

We hypothesized that the non-control conditions would develop inaccurate representations compared to the control condition. In the same-different task, our one-sided alternative hypothesis was that the RV coefficient would be greater in the control condition than in the non-control conditions. The null hypothesis was that there would be no between-condition difference in the RV coefficient.

Two different prior distributions of the between-group difference were used for this analysis to check the sensitivity of the Bayes factor to the prior. The first prior, which was reported in the main text, was Cauchy$(0, 1/\sqrt{2})$ and set following the default procedure of (van Doorn et al., 2020). A more diffuse distribution, Cauchy$(0, \sqrt{2})$, was chosen as the second prior.

Results

Category Structures

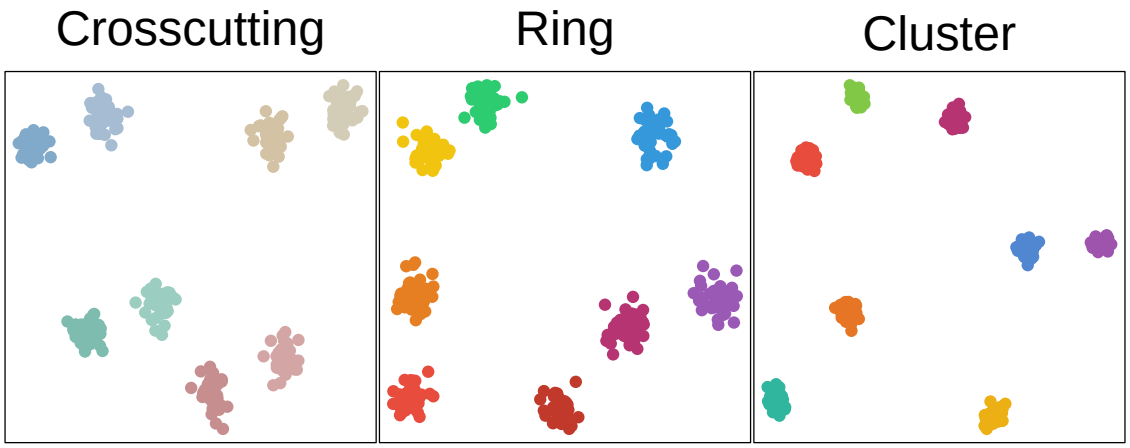

Figure S1: **Two-dimensional embedding of the category structures.** Two-dimensional embedding of the category structures used in the experiment is plotted. In the crosscutting structure, a tree structure defines four clusters of categories (colored in red, yellow, green, and blue), which are again crosscut within each cluster (thick and light colors). In the ring structure, categories have a circular dependency in a feature space. In the cluster structure, categories are arbitrarily distributed without any underlying regularity.

Figure S1 shows the two-dimensional multidimensional scaling solutions of category structures used in our study, including the crosscutting structure that was not discussed in the main text.

Stimulus Diversity

We compared the Shannon entropy to check whether the item-level personalization algorithm reduced the number of categories exposed to participants (Fig. S2).

Let $\mathbf{q} = (q_1, \cdots, q_C)^T$ $(C = 8)$ represent a vector consisting of the category presentation frequency. The Shannon entropy for category diversity is defined as follows:

$$\text{entropy}_{\text{category}} = \sum_{c=1}^{8} q_c \log \frac{1}{q_c} = -\sum_{c=1}^{8} q_c \log q_c$$

The maximum possible value of entropy for category diversity is achieved when all categories are equally presented, i.e., $q_c = 1/8$ for all c:

$$\max(\text{entropy}_{\text{category}}) = -\sum_{c=1}^{8} \frac{1}{8} \log \frac{1}{8} \approx 2.07944$$

Lower entropy values mean that a smaller number of categories were presented during the 128-trial learning phase and that the personalization algorithm successfully constrained the diversity of stimuli. On the contrary, large entropy values mean that categories were presented with similar

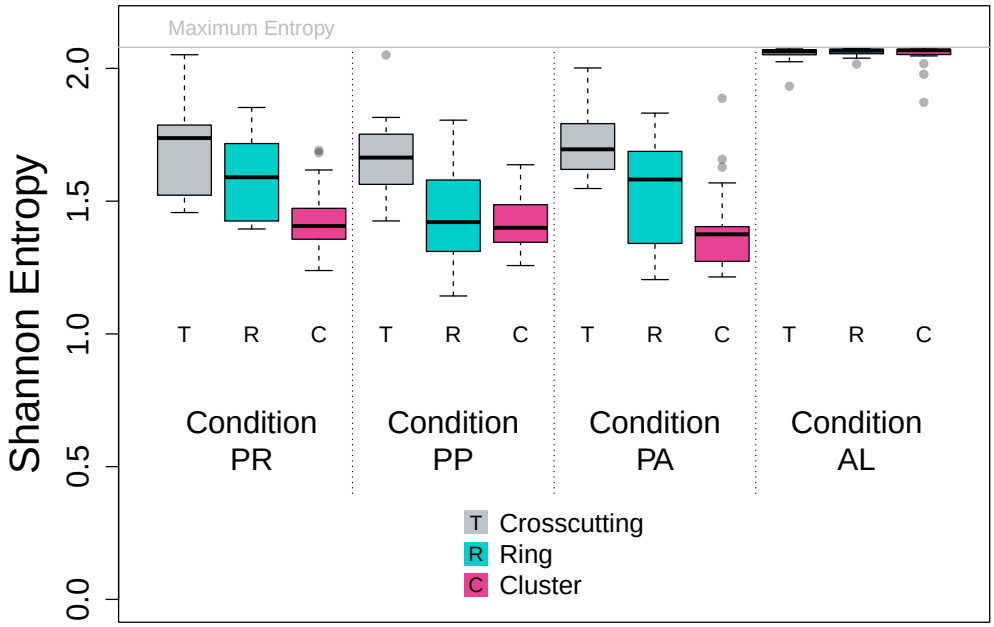

Figure S2: **Shannon entropy of category presentation frequencies.** The Shannon entropy of category presentation frequencies during the learning phase is plotted. Lower entropy values mean that a smaller number of categories were presented selectively. Category structures are differently color-coded (see the legend).

frequencies. Note that the entropy for the control condition is always the maximum value (≈ 2.07944).

Figure S2 shows that the entropy of all personalized learning conditions (i.e., Conditions PR, PP, and PA) is smaller than that of the control condition. This means that the personalization algorithm used in the experiment successfully confined the diversity of learning content. By contrast, the entropy of Condition AL was close to that of the control condition, which means that active learners sampled eight categories impartially.

Overall, the entropy was greater in the crosscutting structure than in the other two category structures, implying that participants were exposed to all categories more evenly. It is likely because the spatial adjacency imposed by the category generation procedure exposed nearby categories more easily.

Results: Sampling Diversity

Figure S3 describes how participants sampled stimulus dimensions during the learning phase. Panel A illustrates the sampling profile of participants after completing the learning phase. Panel B shows the Shannon entropy of the sampling profile vectors after the learning phase (scatterplots) and the entropy changing over time (line plots). Lower entropy values mean that participants sampled a smaller number of dimensions selectively. Larger entropy values suggest that participants sampled six dimensions impartially. As participants in the control conditions were instructed to sample all features, the Shannon entropy of the control condition is fixed to a maximum value (≈ 1.79176).

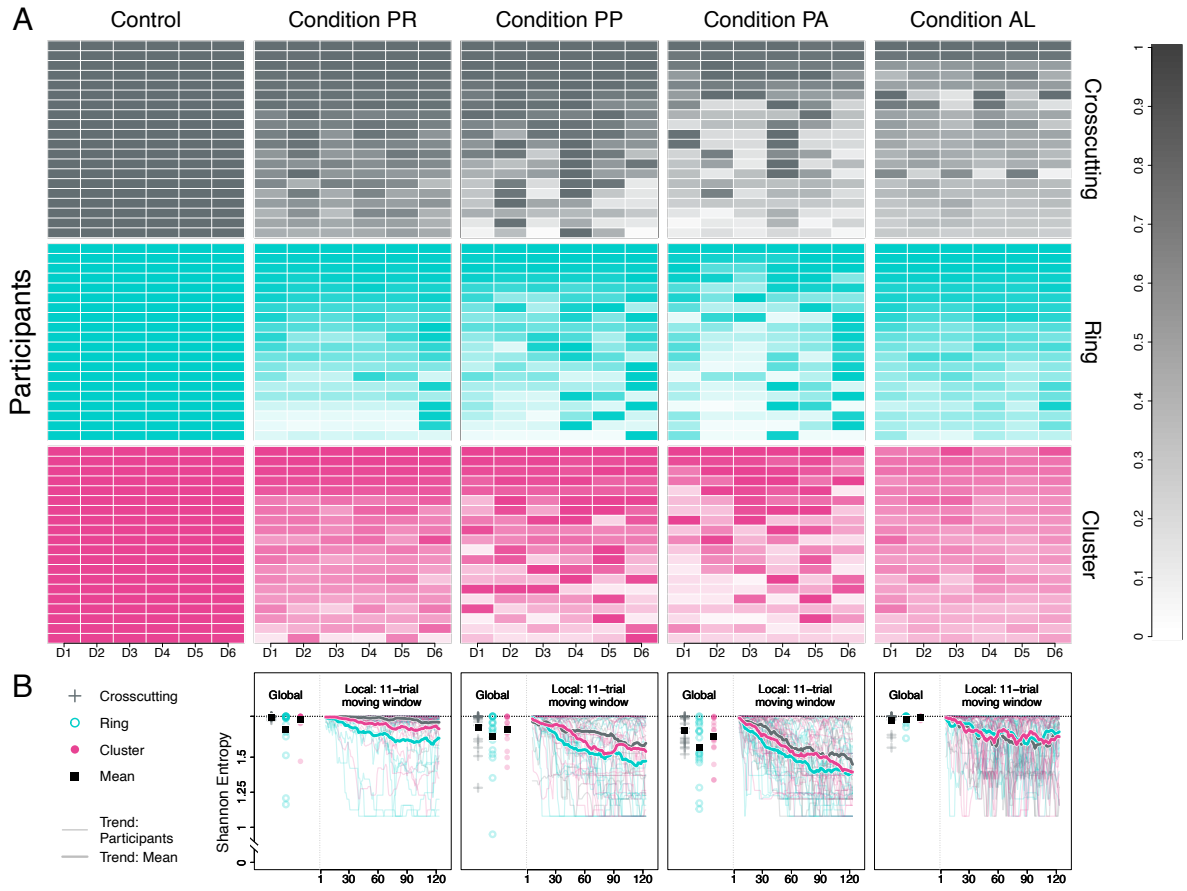

Figure S3: **Selectivity of the sampling profiles.** (A) Heatmaps illustrate the dimension-wise sampling profiles of participants after completing the learning phase. The rows and columns of each panel represent participants and stimulus dimensions, respectively. (B) The Shannon entropy of participants' sampling profile vectors is plotted. Scatterplots (left) show the entropy after completing the learning phase. Line plots (right) describe the local entropy changing over trials using a sliding window with a width of 11 trials. Category structures are differently color-coded.

The results suggest that personalized learning conditions, especially Conditions PP and PA, tend to develop more selective sampling profiles. Participants in Condition AL have relatively impartial sampling patterns.

The sampling entropy of the Condition PR participants tends to be greater than that of the other participants in the personalization conditions (PP, PA) and is close to its maximum value in the crosscutting structure. The category presentations that are relatively unbiased under the crosscutting structure (Figure S2) may have made the deviation from sampling biases more significant.

Figure S4 shows the number of sampled trials collapsed across participants and trials. Participants tended to sample all six dimensions or only one dimension, but moderate-range numbers were also observed. The variability in the number of sampled dimensions may have helped our personalization algorithm learn a more graded, not all-or-none, recommendation scheme.

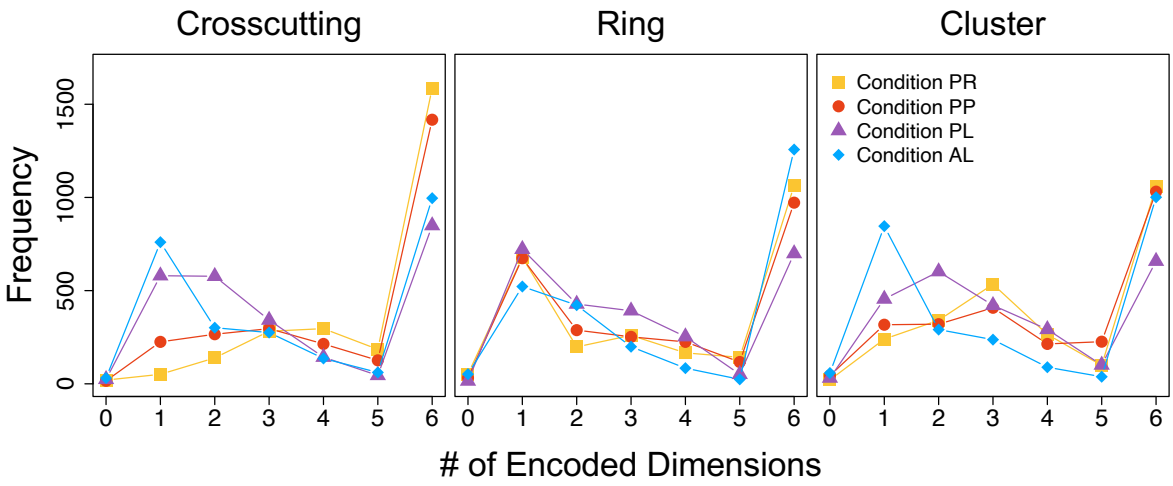

Figure S4: **The number of sampled dimensions.** The distributions of the number of sampled dimensions are presented, after collapsing across participants and learning-phase trials. Each panel represents the result from a category structure condition. The control condition was excluded because participants sampled all six dimensions in that condition.

Representational Accuracy: Same-different Task

Figure S5 illustrates the RV coefficients estimated from all conditions, along with the associated Bayes factors. $\log_{10} BF_{10}$ values are the base-10 log-transformed Savage-Dickey density ratio. Positive values mean that the RV coefficient of a non-control condition is lower than that of the control condition, indicating representational distortion. Negative values support the null hypothesis that there is no difference in the RV coefficients between the control and non-control conditions.

In the planned comparisons, all but two conditions failed to support either hypothesis when a prior Cauchy(0,$1/\sqrt{2}$) was used. Condition AL in the crosscutting structure and Condition PP in the cluster structure showed 'substantial' evidence supporting the null hypothesis. With a more diffuse prior (i.e., Cauchy(0, $\sqrt{2}$)), Conditions PR, PA, and AL in the cluster structure also showed 'substantial' evidence supporting the null hypothesis. In other cases, the evidence strength was weak to support either hypothesis. The Bayes factors largely influenced by the change of the prior also indicate that the evidence provided by the same-difference task data is not robust enough.

Representational Accuracy: Confusion Matrix

Figure S6 shows the condition-wise confusion matrices. If a confusion matrix looks less like a diagonal matrix, the category representation deviates from the ground truth. In both the ring and cluster structures (middle and bottom rows), the confusion matrices of the control and AL conditions are closer to the diagonal matrix. This observation suggests that participants in these conditions developed relatively correct category representations. Participants in the personalization conditions (PR, PP, and PA) show clear deviations from the shape of a diagonal matrix, suggesting that their category representations do not match that of the ideal learner who did not misclassify any test items.

Conditions in the crosscutting structure (top row) show that even the control condition suffers

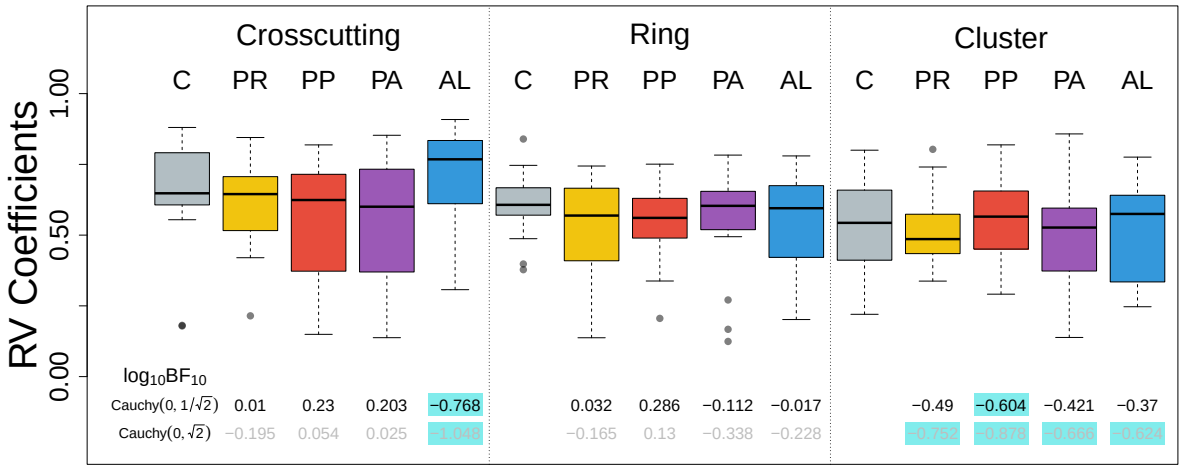

Figure S5: **Same-different task: RV coefficients between the latent category representation and the test items.** Boxplots show the association between the latent representation measured by the same-different task and the category centroids, quantified as RV coefficients (Robert & Escoufier, 1976). Higher values mean that the latent representations and the category centroids are strongly associated. $\log_{10} BF_{10}$ the base-10 log-transformed Savage-Dickey density ratio (rounded to three decimal points). Positive values support the alternative hypothesis that the control condition has higher RV coefficients than a non-control condition. Negative values support the null hypothesis that there is no difference between the two compared conditions. Any $\log_{10} BF_{10}$ values smaller than -0.5 (cyan) were highlighted, which respectively correspond to Jeffrey's suggestion for "substantial" evidence supporting the alternative and null hypotheses (Jeffreys, 1961). Numbers colored in black and gray are the Bayes factors obtained using the priors Cauchy(0, $1/\sqrt{2}$) and Cauchy(0, 2), respectively. C: Control.

from incorrect categorization. Two terminal categories stemming from the immediate parent (i.e., Categories 1 & 2, 3 & 4, 5 & 6, 7 & 8) are likely to be confused with each other even when all category members were equally exposed during the learning phase.

Figure S7 summarizes the representational metrics compared between conditions. The left column shows the representational distance (RSSE) between the confusion matrix of individual participants and that of the ideal responses. The right column shows overall accuracy. We discussed in the main text that all personalization conditions (PR, PP, PA) showed at least 'substantial' evidence of representational distortion compared to the control condition in the ring and cluster structures. By contrast, the crosscutting structure shows that we failed to control the task difficulty. The heatmaps in Figure S6 show that participants in this structure confused Category 1 with Category 2, Category 5 with Category 6, and Category 7 with Category 8 consistently, regardless of the sequence manipulation. The RSSE and overall accuracy measured from the crosscutting structure (Figure S7, top row) shows that the distortion in the control condition was similar to other non-control conditions.

Personalization conditions (PR, PP, PA) under the ring and cluster structures tend to have lower accuracy than the control condition. Although this pattern is consistent with the results from RSSE, evidence strength is noticeably weaker and often makes the comparison indecisive, particularly in the ring structure. The nature of the accuracy metric may have failed to incorporate the details of

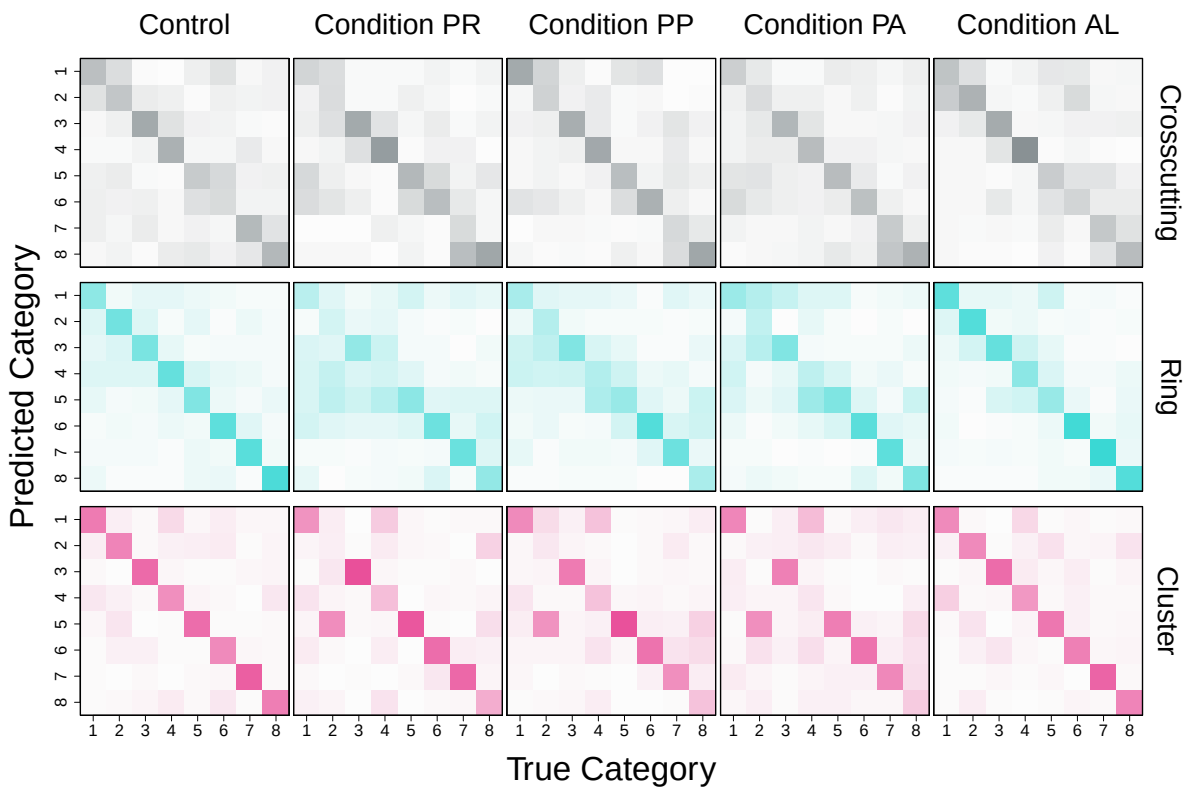

Figure S6: **Confusion matrices.** Heatmaps illustrate the condition-wise average confusion matrix. Densely colored cells are the ones with higher average response frequency.

incorrect categorization induced by the distorted category representation.

"Novel category" Responses

Figure S8 shows the histograms of the "novel category" responses reported during the post-learning categorization test. Given a test item from the categories that were not presented during the learning phase, if a participant selected previously unobserved category labels, we defined this response pattern as a "novel category" response regardless of the accuracy of such choices. Participants who observed all categories during the learning phase could not produce the "novel category" responses by definition, and therefore, were excluded from the plot.

Compared to the ring and cluster structures reported in the main text, the crosscutting structure is not likely to produce "novel category" responses. As shown in Figure S6, the categories defined in the crosscutting structure can be easily confused. The environmental structure may easily cause incorrect generalization and reduce the possibility of the "novel category" responses.

It is also noteworthy that a noticeably smaller number of participants were considered in the histogram in the crosscutting structure than in the ring and cluster conditions. Due to the adjacency of categories imposed by the crosscutting structure, the personalization algorithm may have exposed all eight categories to many participants.

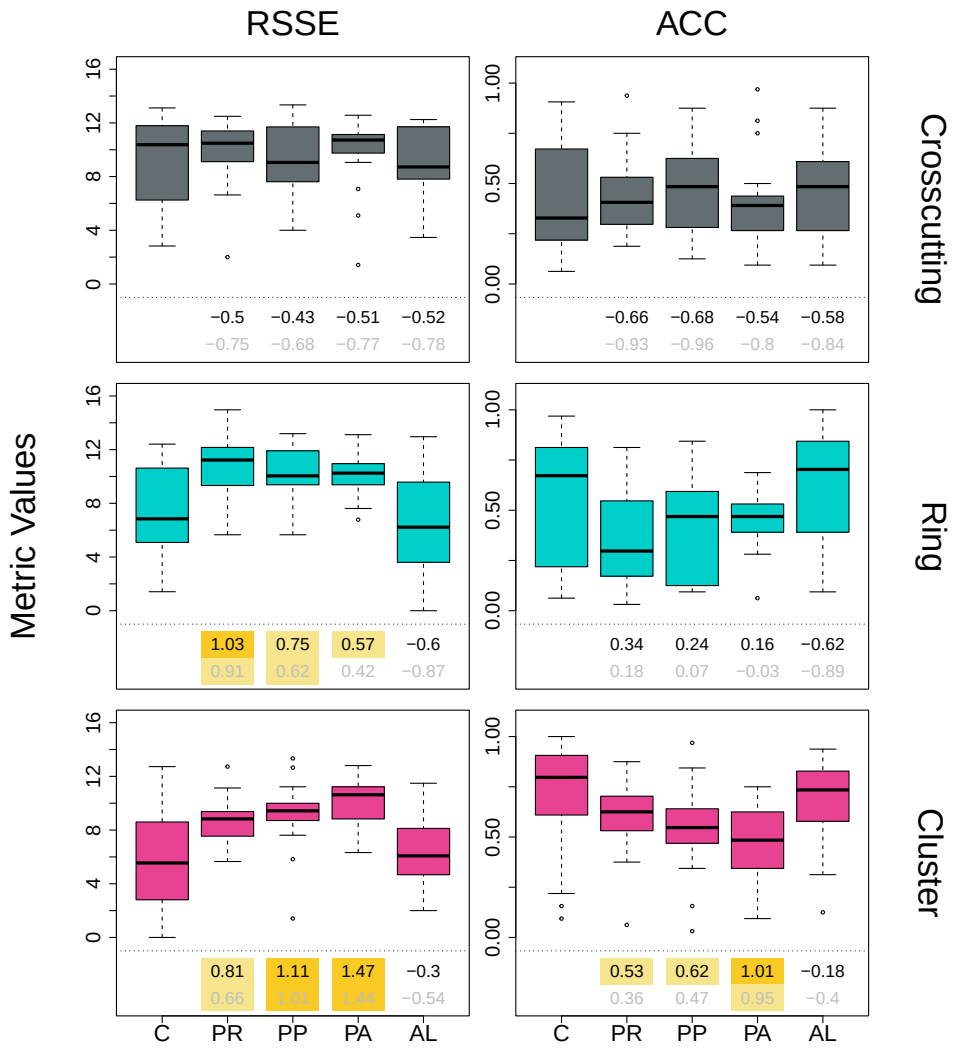

Figure S7: **Representational distances and performance measures.** Boxplots summarize representational distance (left column) and classification accuracy (right column). The planned comparisons between the control and non-control conditions were reported below using the Bayes factor. $\log_{10} BF_{10}$ means the base-10 log-transformed Savage-Dickey density ratio (rounded to two decimal points). Positive values support the alternative hypothesis that the representational distance (accuracy) of the non-control conditions is greater (lower) than that of the control condition. Negative values support the null hypothesis that there is no difference between the two compared conditions. Any $\log_{10} BF_{10}$ values greater than 0.5 (light yellow) and 1 (yellow) were highlighted, which respectively correspond to Jeffrey's suggestion for "substantial" and "strong" evidence supporting the alternative hypothesis (Jeffreys, 1961). Numbers colored in black and gray indicate the result from two different priors, Cauchy$(0, 1/\sqrt{2})$ and Cauchy$(0, \sqrt{2})$, respectively. C: Control, RSSE: Square-root sum of squared errors (see the main text), ACC: accuracy.

Post-learning Categorization: Accuracy

Figure S9 shows the accuracy in the post-learning categorization task as a function of the representativeness scores of test items, including the crosscutting condition that was not reported in the

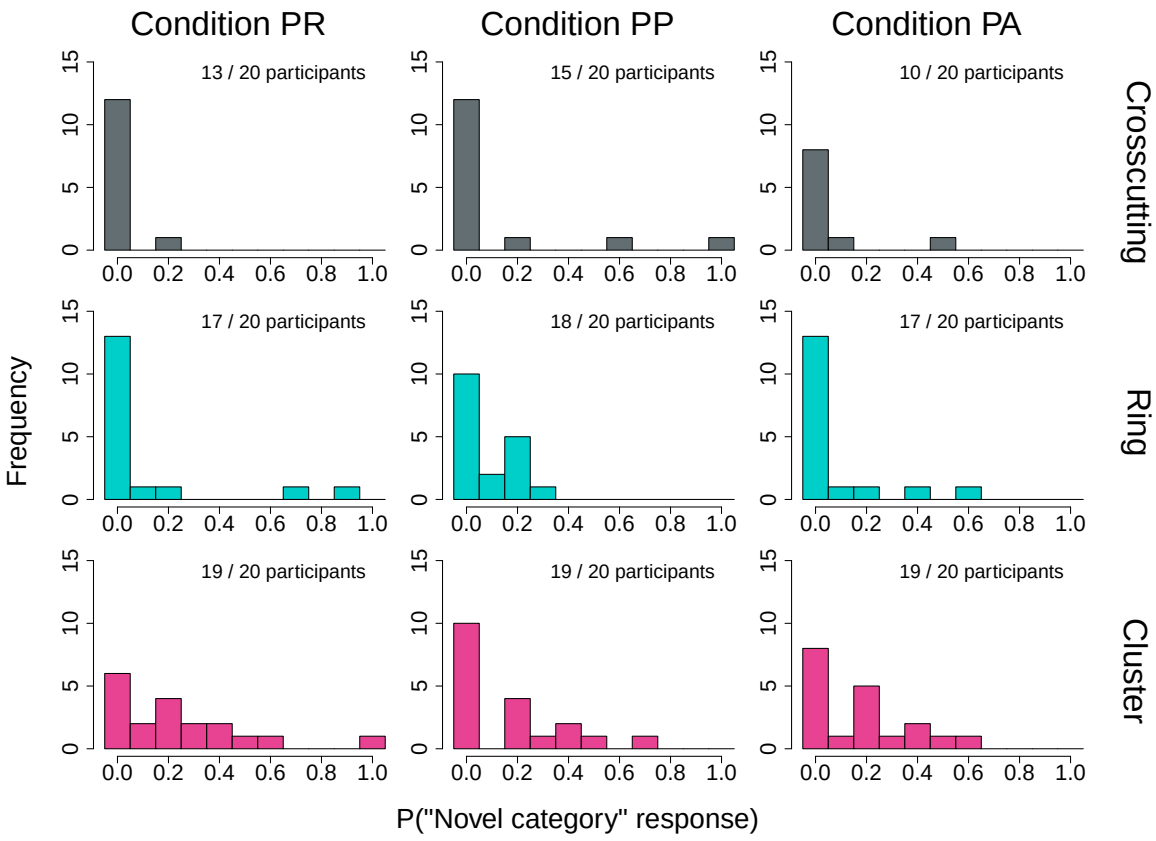

Figure S8: **"Novel category" responses.** Histograms of the probability that participants give the "novel category" responses during the post-learning categorization test. The number of participants validly included in the histogram is specified in the top right corner of each panel. The control and active learning conditions were excluded because participants observed all eight categories at least once.

main text. In all three category structures, higher representativeness scores are associated with greater accuracy in the categorization task.

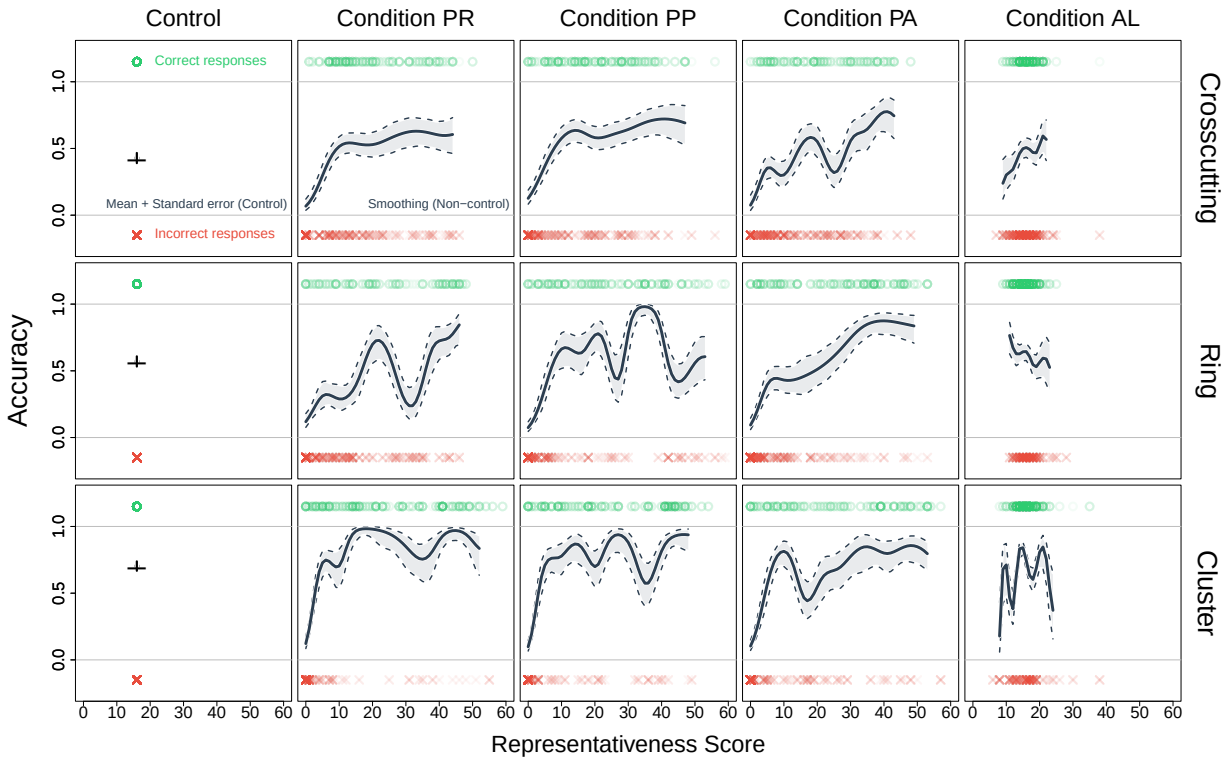

Figure S9: **Data description: Accuracy.** Post-learning categorization accuracy (y-axis) as a function of the representativeness score of the test items (x-axis). The accuracy of the individual test items (empty circles and "×" markers) is illustrated along with the mean trend for each condition (bold lines). Shaded areas represent the 95% confidence intervals of the estimated trend. The accuracy of trials was color-coded differently. The trend lines and associated confidence intervals for the non-control conditions are based on the predictions of generalized additive models using cubic spline bases fitted to each condition.

Post-learning Categorization: Confidence

Figure S10 shows the confidence in the post-learning categorization task as a function of the representativeness scores of test items, including the crosscutting condition that was not reported in the main text. When a participant makes correct decisions, higher representativeness scores are associated with greater confidence ratings. If a participant makes incorrect decisions, (near-)zero representativeness seems to induce higher confidence ratings in the personalization conditions. However, this pattern is relatively not clear in the crosscutting structure (first row) because the overall confidence rating in this condition seems higher compared to the other two category structures.

As suggested by the exploratory trend analysis and discussed in the main text, the results show that personalization can cause inflation in confidence even when participants have learned almost nothing about an item (i.e., with (near-)zero representativeness).

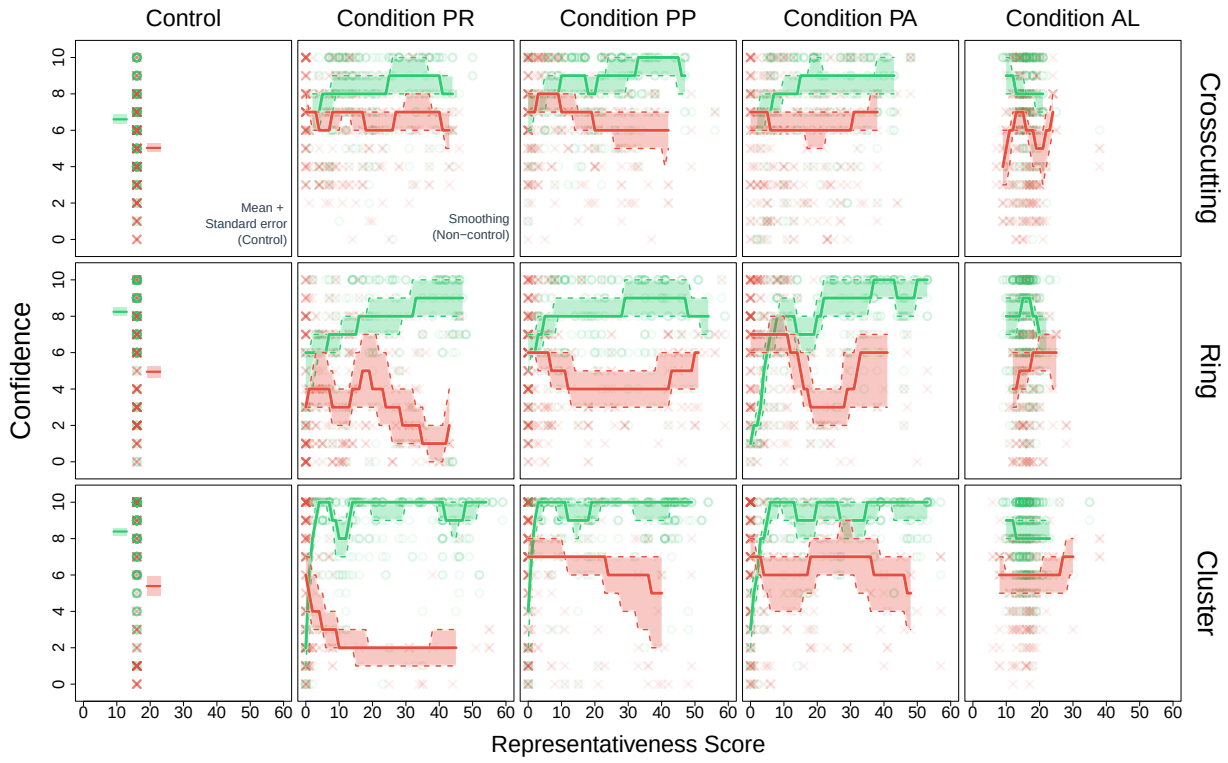

Figure S10: **Data description: Confidence.** Post-learning categorization confidence rating (y-axis) as a function of the representativeness score of the test items (x-axis). The reported confidence of the individual test items (green empty circles and red "×" markers) is illustrated along with its mean trend for each condition (bold lines). The accuracy of trials was color-coded differently. In both panels, shades indicate the 95% confidence interval of the estimated trend. The trend lines and associated confidence intervals for the non-control conditions are based on the predictions of generalized additive models using cubic spline bases fitted to each condition and accuracy level.

Mixed-effect modeling

Table S1 and S2 show the posterior estimates and their 95% equal-tailed credible intervals of the fixed-effect coefficients and random-effect standard deviation terms.

Table S1: **Posterior estimates and credible intervals of the fixed-effect coefficients.**

Continuous variables		Category structure		Learning sequence					Posterior mean	Credible interval
C	S_R	Cluster	Ring	Control	PR	PP	PA	AL		95%
(Intercept)		X		X					0.168	(-1.343, 1.739)
		X			X				-1.554	(-3.180, 0.035)
		X				X			-0.932	(-2.575, 0.684)
		X					X		-1.416	(-3.089, 0.186)
		X						X	-2.589	(-4.304, -0.895)
X		X		X					-0.326	(-0.630, -0.039)
X		X			X				0.340	(0.040, 0.651)
X		X				X			0.129	(-0.170, 0.438)
X		X					X		0.210	(-0.090, 0.520)
X		X						X	0.401	(0.081, 0.722)
	X	X		X					-0.202	(-0.276, -0.133)
X	X	X		X					0.058	(0.046, 0.070)
			X	X					-1.788	(-3.928, 0.262)
			X		X				1.614	(-0.473, 3.784)
			X			X			0.003	(-2.255, 2.296)
			X				X		2.836	(0.680, 5.059)
			X					X	2.822	(0.615, 5.114)
X			X	X					0.317	(-0.084, 0.722)
X			X		X				-0.355	(-0.789, 0.066)
X			X			X			0.000	(-0.437, 0.431)
X			X				X		-0.451	(-0.880, -0.039)
X			X					X	-0.144	(-0.583, 0.292)
	X		X	X					0.087	(0.002, 0.173)
X	X		X	X					-0.028	(-0.044, -0.013)

Note: C: Confidence ratings, S_R: Representativeness scores.

Table S2: **Posterior estimates and credible intervals of the random-effect standard deviation terms.**

Variables	Posterior mean	95% Credible interval
Intercept	1.283	(0.775, 1.855)
C	0.237	(0.154, 0.329)
S_R	0.139	(0.097, 0.185)
C:S_R	0.031	(0.024, 0.040)

Note: C: Confidence ratings, S_R: Representativeness scores, $A : B$: Two-way interaction between A and B.

References

Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for YouTube recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems*, 191–198.

Indahl, U. G., Næs, T., & Liland, K. H. (2018). A similarity index for comparing coupled matrices. *Journal of Chemometrics*, *e3049*.

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). OUP Oxford.

Josse, J., & Holmes, S. (2016). Measuring multivariate association and beyond. *Statistics Surveys*, *10*, 132.

Josse, J., Pagès, J., & Husson, F. (2008). Testing the significance of the rv coefficient. *Computational Statistics & Data Analysis*, *53*, 82–91.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (Second). Academic Press.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Powers, D. (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Robert, P., & Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: The RV-coefficient. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *25*(3), 257–265.

Smilde, A. K., Kiers, H. A., Bijlsma, S., Rubingh, C., & Van Erk, M. (2009). Matrix correlations for high-dimensional data: The modified RV-coefficient. *Bioinformatics*, *25*, 401–405.

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2020). Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman's ρ. *Journal of Applied Statistics*, *47*(16), 2984–3006.