AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

© 2025 The Author(s)
ISSN: 0096-3445

2025, Vol. 154, No. 9, 2503–2522
https://doi.org/10.1037/xge0001763

Algorithmic Personalization of Information Can Cause Inaccurate Generalization and Overconfidence

Giwon Bahg[1], Vladimir M. Sloutsky[2], and Brandon M. Turner[2]
[1] Department of Psychology, Vanderbilt University
[2] Department of Psychology, The Ohio State University

Personalization algorithms are widely used online to deliver recommendations fine-tuned to individual users. This specificity comes at the cost of the diversity of information presented to users, limiting exposure to alternative perspectives and potentially reinforcing existing beliefs. We investigated the degree to which personalization can hinder the acquisition of new knowledge of categories. We asked participants to learn about alien categories under different levels of personalization and tested their knowledge using a post-learning categorization task. Our results show that learners in personalized environments sample feature information more selectively during the learning phase and develop inaccurate representations about the categories. Critically, they also report inflated confidence about their inaccurate decisions for categories for which they had little exposure. Our results suggest that personalization can distort learners' understanding of the environment, bias information sampling, and induce incorrect generalization of knowledge.

Public Significance Statement

Many online content-sharing services use "personalization" algorithms to recommend information based on a user's preferences and interests. However, personalization has been criticized because it can severely limit users' access to the world, leaving them constrained in what they learn and believe. We tested whether personalization can actually harm learning by systematically varying the degree of personalization and asking participants to learn alien categories. Our study shows that the personalized learning environment significantly biases information, leading to learners' limited information search, decreased learning accuracy, and overconfidence about inaccurate responses. Our results suggest that learners in personalized environments incorrectly expect their knowledge to apply in domains for which they have little experience. A broad application of one's limited and biased categorical knowledge can be problematic for our society because it can result in stereotypical thinking and conceptual biases.

Keywords: category learning, personalization, recommendation algorithm, filter bubble

Supplemental materials: https://doi.org/10.1037/xge0001763.supp

This article was published Online First June 23, 2025.

Aleksandra Krogulska served as action editor.

Giwon Bahg https://orcid.org/0000-0002-9068-4535

Vladimir M. Sloutsky https://orcid.org/0000-0003-4054-3360

Brandon M. Turner https://orcid.org/0000-0003-0966-9301

All data and analysis code are available via the Open Science Framework (https://osf.io/n2mxu/). This study was not preregistered. An earlier version of this article was uploaded as a preprint (https://osf.io/preprints/psyarxiv/yahvf). Portions of this work were previously presented at the 2021 and 2022 virtual annual meetings of the Society for Mathematical Psychology. Portions of this work were submitted by Giwon Bahg in partial fulfillment of the requirements for the PhD degree in psychology at The Ohio State University.

The authors declare that there are no conflicts of interest with respect to the authorship or the publication of this article. This study was supported by the National Science Foundation CAREER award, Grant BCS-1847603 (awarded to Brandon M. Turner), and the National Institutes of Health, Grant R01HD078545 (awarded to Vladimir M. Sloutsky). The authors thank Ryan Day for the development of the online task environment, Matthew Galdo for his early contribution to task development, and Simon Lilburn for the constructive feedback on this article.

Giwon Bahg played a lead role in conceptualization, data curation, formal analysis, investigation, project administration, resources, software, validation, visualization, writing–original draft, and writing–review and editing. Vladimir M. Sloutsky played a supporting role in conceptualization, supervision, and task design and an equal role in funding acquisition and writing–review and editing. Brandon M. Turner played an equal role in conceptualization, funding acquisition, project administration, supervision, and writing–review and editing.

Correspondence concerning this article should be addressed to Giwon Bahg, Department of Psychology, Vanderbilt University, 111 21st Avenue South, 301 Wilson Hall, Nashville, TN 37240, United States. Email: giwon.bahg.1@gmail.com

Many companies providing web-based services (e.g., Google, Facebook, YouTube) have used algorithms for recommending information individually fine-tuned to their users since the mid-2000s. These so-called "personalization" algorithms learn your interests and preferences (e.g., types of video clips you frequently watch), find what kind of content is accessed by other users whose behavior is similar to yours, and suggest it to you. By doing so, you are expected to spend more time on the service, resulting in revenue benefits for the companies.

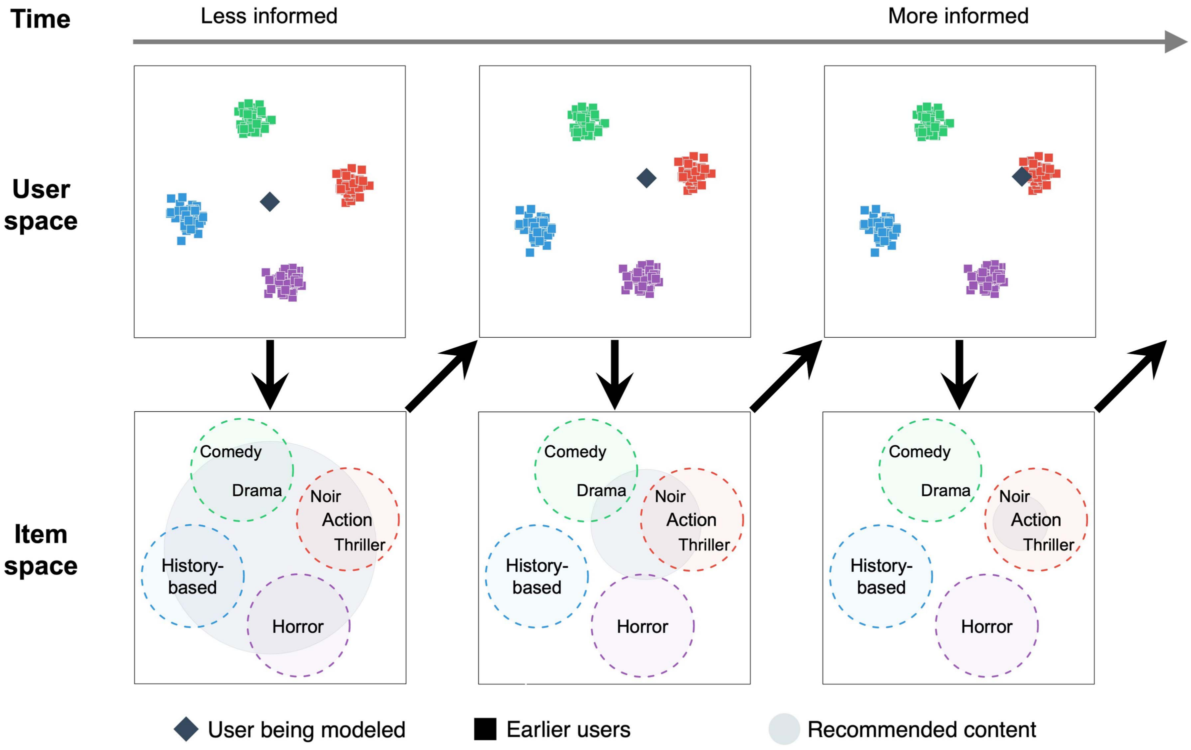
Personalization helps users receive the information and produce advertisements that matter most to them, enabling a much smoother and more efficient navigation of a vast sea of information. However, when an algorithm is designed to increase some consumption metric, such as keeping users engaged with content, it is at least possible that the information we receive is not a representative sample of the world, and this biased sample can lead to a severely distorted impression of reality.

Figure 1 illustrates how the interaction between the personalization algorithm and the user can overrepresent items having shared characteristics. Let us imagine that a person who has never watched movies from a certain country wants to try them, and an on-demand streaming service recommends movies using a personalization algorithm. Individual movies are distributed in the "Item space" (bottom row), roughly based on their genres. Users watching different movies may also be grouped based on their behavior in the streaming service ("User space"; top row). The service recommends some popular movies. The person accidentally chooses an action-thriller film (say, just because it was the very first item on the list) and pleasantly finishes it. The algorithm "behind the curtain" searches for more movies watched by other people who finished the first movie. These movies are mostly from similar genres like action, thriller, and neo-noir; the person watches most of them. The recommended list is filled with more movies of these genres because of other people who watched similar movies.

If the person in the story wanted to expand one's interest in movies to other countries and find something they might like, we can say that the algorithm fulfilled its purpose. However, if this person's goal, whether explicit or implicit, was in fact to understand the overall landscape of movies in this country, the algorithmic recommendation ends up seriously biasing one's understanding. This person is likely to miss other great movies in different genres (e.g., comedy, drama, horror, history-based). This person may also draw an unfounded and overstretching inference about popular culture and society based on one's biased representation (e.g., "what are

Figure 1

A Schematic of Personalization Algorithm Informed by Other Users' Behavior (Known as Collaborative Filtering)



Note. A personalization algorithm recommends items that are most likely to be "consumed" (e.g., watch video clips, purchase items) by individual users. When the algorithm does not have enough information about a user's preferences ("User space" row, first column), it relies on weak constraints provided by previous users, and therefore, all items are likely to be suggested relatively equally ("Item space" row, first column). The algorithm gradually updates its knowledge about the current target user's preferences based on one's behavior. Although initial choices may be incidental, the algorithm identifies previous users that share similar patterns of information consumption ("User space" row, second and third columns) and a subset of the item space they have consumed. The algorithm constrains newly presented recommendations for the current user to this region ("Item space" row, second and third columns). See the online article for the color version of this figure.

the social factors that make neo-noir and crime films particularly popular in this country?").

Our study asked whether biased experiences caused by personalization algorithms may cause problems when humans are genuinely trying to learn about the world. The debate centered around the concept of "filter bubbles" has discussed the lack of information diversity as a cause of intellectual isolation in web-based environments, where previously held beliefs are only further reinforced by personalized information that supports the existing belief (e.g., Pariser, 2011; Yesilada & Lewandowsky, 2022).

Earlier studies on personalization mainly focused on whether personalization algorithms guide users to information selectively (e.g., Hannak et al., 2013; Krafft et al., 2019; Le et al., 2019; Möller et al., 2018) or whether cognitive models might better serve the recommendations provided by curation algorithms (e.g., Bourgin et al., 2021). Regarding the selectivity in the personalization algorithms' recommendations, the mixed results reported in this literature are likely due to potential differences in study design: whether a study relied on real-world data or was experimental, whether the queries made for personalization were appropriate for inducing fine-tuned outputs, whether the measurement of information diversity was domain-specific (e.g., politics) and so forth. Also, semantically loaded domains (e.g., news consumption; Thurman & Schifferes, 2012) investigated in this literature may have potential confounding factors. For example, the behavior of both service providers and users in providing or consuming recommendations may interact with their beliefs about the value of news (e.g., objectivity vs. value-ladenness; Figdor, 2010) and attitudes already established about topics.

More importantly, how personalization affects basic cognitive mechanisms and people's understanding of the world has been less emphasized. A few controlled lab experiments focusing on cognitive processing and personalization (Ho & Tam, 2005; Ho et al., 2011; Tam & Ho, 2006) only assumed a single event of personalized recommendation (but see Ho & Bodoff, 2014, using multisession personalized recommendations for studying preferential choices and attitudes toward personalization algorithms). A crucial concern about personalization is that it can develop and reinforce false beliefs by guiding its users to a knowledge space of highly correlated perspectives. This view implies that personalization is a temporal process and it cannot be fully understood through a single-time event.

To address the concerns about temporal changes in one's belief and representation, we investigated the effect of personalization on basic cognitive mechanisms by focusing on a domain of category learning, securing continuous user-algorithm interactivity similar to the real world by adapting a recommendation algorithm proposed for content-sharing services (Covington et al., 2016). Our experiment used synthetic stimuli to minimize the confounding effect of prior expectations or beliefs on learning. Our hypothesis is that personalization algorithms may overrepresent a learner's "preferred" category or stimulus dimension and, in turn, underrepresent other categories and dimensions that contain valuable information that would enrich a learner's representation of the category structure.

Method

This study was approved by the Institutional Review Board of The Ohio State University.

Participants

Three hundred forty-three participants (346 if we count duplicated participations) were recruited via an online experiment platform Prolific (https://prolific.co) and were paid $10 when they completed the experiment. We conducted the online experiment expecting a convenience sample. We did not collect detailed demographic information as it was not part of our Institutional Review Board-approved protocol. Participants aged 18 or older, fluent in English, and with normal or corrected-to-normal vision were eligible.

We aimed for 20 participants per condition, resulting in 300 participants in total (20 Participants × Three Types of Category Structures × Five Levels of Learning Sequence Manipulation). Our long-term goal when designing the study was to fit a computational cognitive model to each individual participant's data and analyze parameter estimates exploratively, rather than to compare observed data directly between conditions via conventional analysis of variance or generalized linear models. For this reason, the number of participants per condition was not based on power analysis. We considered that 20 participants per condition would be enough to find within-group similarity and between-group difference qualitatively in the parameter estimates.

We excluded the data from 146 experimental sessions, including three sessions finished by three duplicated participants. Forty-six of them were removed due to data quality, and 100 of them due to task difficulty. The data from 200 participants (20 Participants × Two Types of Category Structures × Five Levels of Learning Sequence Manipulation) were analyzed.

During data collection, 46 participants were excluded due to poor quality of data from the learning phase. Our task required participants to click on parts of stimuli to uncover and sample features. The aforementioned participants did not click stimuli for 14 or more trials during the learning phase. Due to the basic structure of the learning-phase task, we could not guarantee that participants who did not sample any features too frequently learned categories appropriately.

After data collection, we decided to exclude one of the category structure conditions (named "crosscutting"; corresponding to 100 sessions in total) entirely from the main analysis due to the failure to control task difficulty. Within this category structure, the accuracy of category representation in the control condition was just as low as in the other learning-sequence conditions. This suggests that categories were so easily confused that participants were unlikely to learn the categories accurately, even in the ideal situation. A detailed discussion on difficulty manipulation is in the Supplemental Material (see also Supplemental Figure S7).

We also found that three participants participated in the experiment twice. They have never been assigned to the same condition and, importantly, to the same category structure. These problematic cases always occurred between the "crosscutting" structure excluded from the main analysis and other regular structures. Therefore, their duplicated participation was discarded only from the excluded "crosscutting" structure.

Stimuli

Our experiment is a modification of a category learning task in which participants learn the relationship between items consisting of a set of features and labels that separate the items into distinct groups. Participants were asked to learn how to categorize fictional aliens. The aliens comprised six continuous features, as shown in Figure 2A: location on a horizontal straight line, radius of a circle, brightness on a grayscale spectrum, orientation of a bulb-like shape, curvature of a hyperbola, and spatial frequency. The mathematical procedure used to generate feature values and visualize them is discussed in the Appendix.

We used multiple stimulus environments with different distributions of categories and underlying category structures, considering the diversity of category distributions in the real world (Kemp & Tenenbaum, 2008; Saxe et al., 2019). Using the method proposed by Saxe et al. (2019), we generated categories with three types of underlying structures, and two category structures were considered for the main analysis (Figure 2B; see also Appendix). In the crosscutting structure (excluded from the main text), a hierarchical tree structure first generated four categories branching from two-parent nodes and then crosscut four category nodes to generate eight categories. In the ring structure, the two-dimensional embeddings of the category centroids formed a ring-like shape. In the cluster structure, category centroids were sampled from a multivariate normal distribution without any between-dimension correlation. Instances of categories are samples from a multivariate normal distribution, taking the centroids as mean locations. Category centroids were chosen so that categories would be separated reasonably in a two-dimensional embedding space obtained by multidimensional scaling.

Design

Learning Sequence Manipulation

The core manipulation of our experiment is how to personalize participants' learning experiences. As this manipulation can seem complex and condition labels are mentioned repeatedly throughout the rest of the article, we first describe five levels of learning conditions. During the learning phase, we manipulated the influence of item- and feature-level personalization systematically (see Figure 3A). In the control condition, all learning set items were presented four times, and participants needed to sample all features.

Three personalization conditions used an item-level personalization algorithm with different feature-level manipulations. In Condition PR (Personalized item and Randomized feature order), items were personalized, but features were presented in random order. In Condition PP (Personalized item and Personalized feature order), the order of feature presentation was also personalized by revealing feature information in the dimension with a higher sampling rate earlier. In Condition PA (Personalized item and Active feature sampling), once an item was presented, participants chose dimensions to sample actively. The conceptual description of the personalization algorithm will be provided in a separate section below.

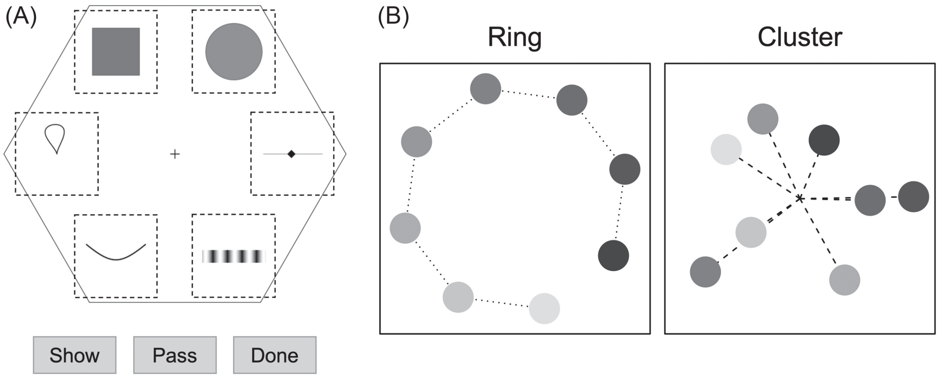
Last, Condition AL (Active Learning) allowed participants to fully control their learning experiences. At an item level, participants chose a category to study at the beginning of each trial. One item was randomly chosen from the selected category and presented. At a feature level, participants selected dimensions to sample directly, as in Condition PA.

Task and Procedure

The experiment consisted of a learning phase and two postlearning tasks. In the learning phase (Figure 4), participants studied categories of fictional crystal-like aliens by sampling feature information. Participants' feature sampling behavior was tracked using a method that is similar to the gaze-contingency paradigm in its motivation but works cumulatively. On each trial, an item was presented with all features hidden behind gray boxes. To learn the relationship between features and the item's category label, participants needed to uncover the boxes differently across conditions.

As illustrated in Figures 3B and 4, participants clicked on either a "Show" button (Conditions Control, PR, and PP) or one of the gray boxes directly (Conditions PA and AL) to reveal a feature, depending on the experimental conditions. In Conditions Control, PR, and PP, a red square cued the location to be revealed if participants clicked on

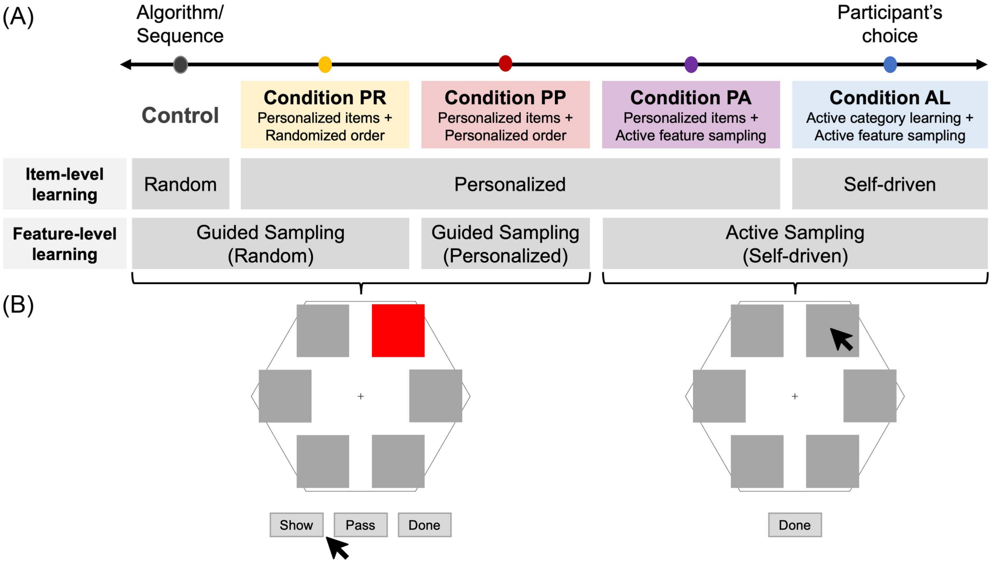
Figure 2
Stimulus and Category Structures



Note. (A) An example stimulus. (B) Simplified visual schematics of category distributions under different underlying structures. Dots represent the category centroids. In the ring structure (left), the locations of eight categories have a spatial dependency. In the cluster structure (right), the centroids of eight categories do not have any underlying regularity or spatial dependency.

Figure 3

Conditions and Feature Sampling During the Learning Phase


(A)
Algorithm/ Sequence
Participant's choice
Control
Condition PR
Personalized items + Randomized order
Condition PP
Personalized items + Personalized order
Condition PA
Personalized items + Active feature sampling
Condition AL
Active category learning + Active feature sampling
Item-level learning
Random
Personalized
Self-driven
Feature-level learning
Guided Sampling (Random)
Guided Sampling (Personalized)
Active Sampling (Self-driven)
(B)
Show
Pass
Done
Done

Note. (A) Manipulation of learning sequences. See the "Learning Sequence Manipulation" section for the details. (B) Depending on the experimental conditions, participants sampled stimulus features by either clicking on the "Show" button to uncover a feature at the location proposed by the experiment program (cued by a red square; left) or clicking directly on the gray square that they want to reveal (right). In the former type of conditions, the "Pass" button allowed participants to skip the proposed dimension. In all conditions, participants finished a trial by clicking on the "Done" button. See the online article for the color version of this figure.

the "Show" button. Participants had to sample all features only in the control condition. Participants were allowed to skip sampling certain dimensions by clicking on the "Pass" button (Conditions PR and PP) or simply not clicking on the gray squares (Conditions PA and AL).

In Conditions Control, PP, PR, and PA, participants clicked the "Done" button when feature sampling was finished, and the category label was presented on the screen. In Condition AL, the label of the selected category was presented on the top until feature sampling was finished by clicking the "Done" button. There was no time limit in the feature sampling step. The complete learning material set consisted of eight categories, each of which had four instances. Category labels were nonword strings consisting of three English alphabets (consonant–vowel–consonant). The length of the learning phase was 128 trials in total.

After finishing the learning phase, participants did (a) a same-different task using a continuous slider and (b) a categorization task. In the same-different task, participants were asked to report their degree of belief that a pair of stimuli came from the same category. Twenty-eight pairs of stimuli from different categories and 16 pairs from the same category were presented, resulting in 44 trials in total. For the different-category trials, category centroids were presented as stimuli. For the same-category trials, a centroid and a sample from the same category constructed the stimulus pair.

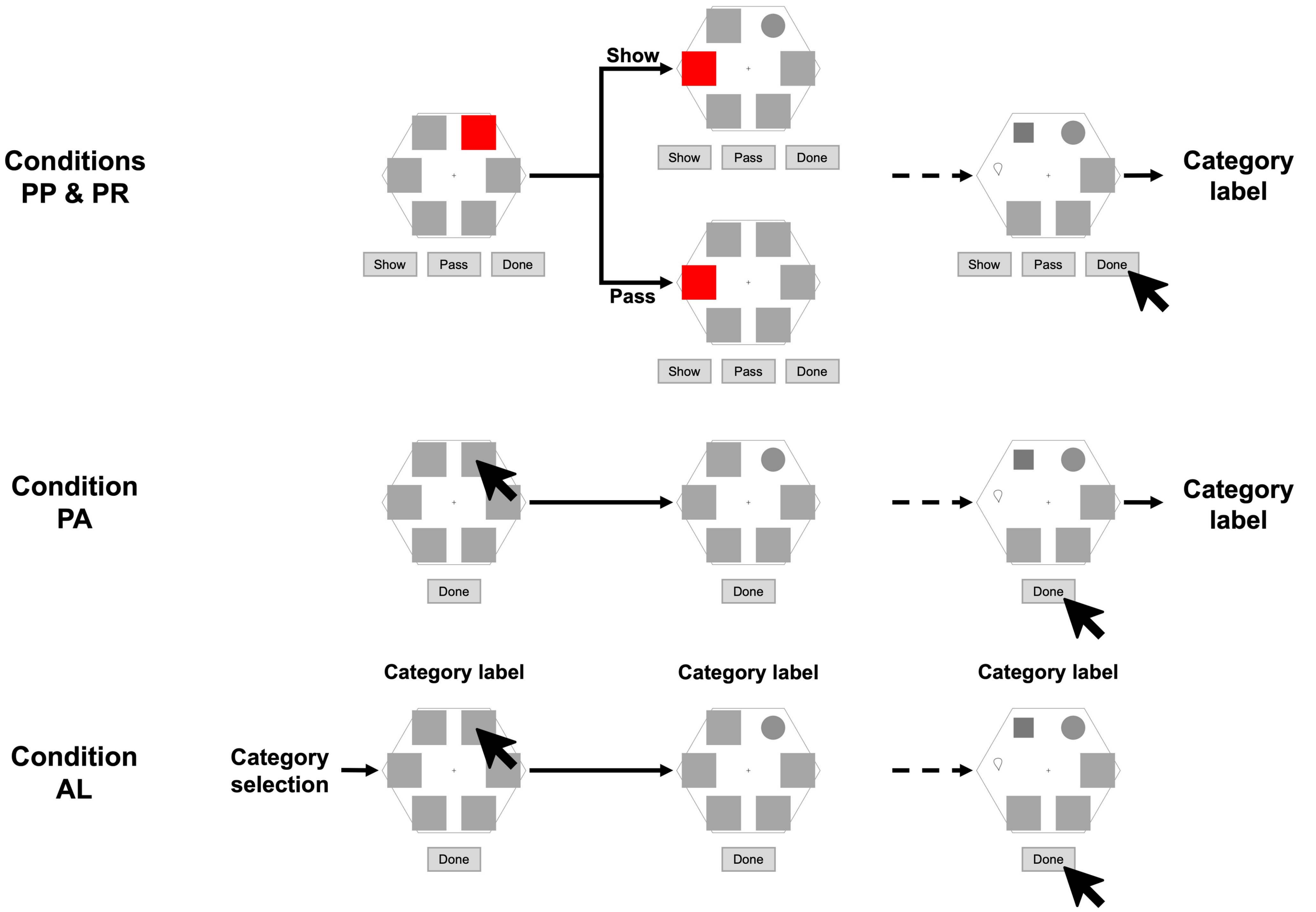
In the postlearning categorization task, participants were asked to classify test items into one of the eight categories using a mouse click. An independent test set comprised four items per category, resulting in 32 items in total. Unlike the learning phase, participants saw all features of aliens from the beginning, and labels of all eight categories were presented to participants as choice options. After making a categorization decision on each trial, participants were asked to answer if they felt confident about their decision on an 11-level Likert scale ranging from 0 to 10. We decided not to counterbalance the order of the two postlearning tasks to minimize any possible carryover effect. Our major concern was that the categorization task showing stimuli and labels from all categories could affect the category representations to be measured in the same-different task.

We emphasize that the learning phase of our experiment deviates from a traditional category learning task in that we did not ask participants to predict category membership. First, knowing all category labels would prevent us from observing the true effect of personalization. In the conditions using the item-level personalization algorithm, there is no guarantee that the algorithm would show all eight categories at least once during the learning phase and constrain participants' pool of knowledge. Therefore, to investigate the effect of personalization, we should avoid offering a complete list of category labels to ensure that participants would not recognize the total number of categories. Second, the absence of categorization decisions during learning increases the ecological validity of the task. The environment in which personalization techniques are applied usually does not test participants' knowledge by asking their learners to predict categories.

In the main text, we did not discuss the same-different task because the results were difficult to interpret. However, there is still

Figure 4

The Learning-Phase Trial Structure of the Noncontrol Conditions



Note. The trial structures of Conditions PP and PR (top row), PA (middle row), and AL (bottom row) are illustrated. Condition PR = Personalized item and Randomized feature order; Condition PP = Personalized item and Personalized feature order; Condition PA = Personalized item and Active feature sampling; Condition AL = Active Learning. See the online article for the color version of this figure.

a possibility that the same-different task affected the performance of the subsequent categorization task. For transparency, we discuss the details and results of the same-different task in the Supplemental Material (see also Supplemental Figure S5).

A Conceptual Overview of the Personalization Algorithm

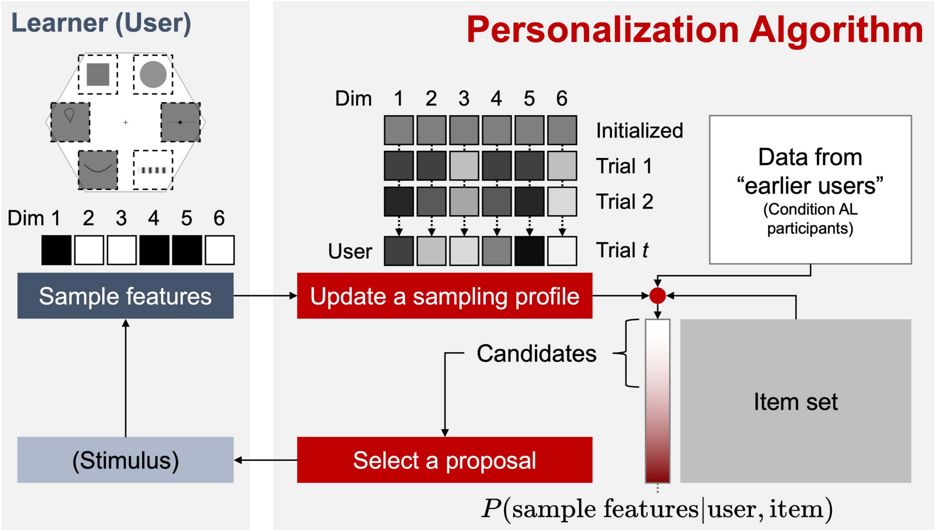
Manipulating items and within-item features presented to participants is the core of our study. Here, we introduce the personalization algorithm used in the study. Readers interested in the mathematical and implementational details of the personalization algorithm are referred to the Supplemental Materials.

To generate item sequences fine-tuned to feature sampling patterns of each participant, we modified a collaborative-filtering recommendation algorithm for YouTube video clips proposed by Covington et al. (2016). Our design choice to adapt the YouTube recommendation algorithm was to secure a reasonable degree of ecological validity in terms of the algorithm's application domain (i.e., content sharing). In general, personalization algorithms aim to recommend information that is most likely to be consumed by users. In the case of video-sharing platforms, content consumption is defined as the platform user's behavior of watching a certain video clip. The original method was trained to learn about the relationship between user information (e.g., watched videos, search queries, age, gender) and video clips that users are likely to watch. Once the personalization method is trained, it takes a user profile as input, computes the probability that the user would watch video clips on the platform, and recommends items with the highest watch probabilities.

Figure 5 describes our simplified version of the personalization algorithm. Just as proposing video clips with the maximum watch probabilities was the goal of Covington et al. (2016), proposing items with the maximum mean sampling probabilities was the goal of our algorithm. The mean sampling probability accounts for the overall expected number of sampled features, so maximizing this quantity means the maximization of content consumption.

Figure 5

Personalization Algorithm


Learner (User)
Dim 1 2 3 4 5 6
Sample features
(Stimulus)
Personalization Algorithm
Dim 1 2 3 4 5 6
Initialized
Trial 1
Trial 2
User
Trial *t*
Data from "earlier users"
(Condition AL participants)
Update a sampling profile
Candidates
Item set
Select a proposal
$P(\text{sample features}|\text{user, item})$

Note. Once a participant samples features of a stimulus, the personalization algorithm updates the sampling profile, chooses candidates with the highest probabilities of feature sampling, and selects one of them as a proposal. The algorithm was trained using the data from the active learning condition (Condition AL). Condition AL = Active Learning. See the online article for the color version of this figure.

The algorithm was trained using participants in the active learning condition (Condition AL) before doing experiments on the personalization conditions (Conditions PR, PP, and PA). Each participant in Condition AL yielded (a) a fully updated user profile vector representing how frequently each participant sampled features from each dimension and (b) the entire history of presented items and the number of sampled dimensions from 128 trials. Participants in Condition AL tended to sample all dimensions or one dimension on each trial, but the number of sampled dimensions was variable across items (Supplemental Figure S4). This variability can introduce biases in the personalization algorithm. We trained the algorithm to find a mapping between a user profile vector and the history of presented items that best predicts the number of sampled dimensions on each trial. In total, 2,560 data points (20 Participants × 128 Trials) were used as a training data set.

In Conditions PR, PP, and PA, the personalization algorithm updated a user profile and proposed items with the highest expected sampling probabilities. Given the most updated user profile, the history of presented items, and the number of sampled dimensions, the algorithm computed the expected feature-sampling probability for all 32 items. The algorithm proposed 12 out of 32 items with the highest expected sampling probability as a candidate set. The item presented on the subsequent trial is one of these candidate items.

We also considered another layer of personalized recommendations at a feature level in Condition PP. This approach recommended that each participant would uncover dimensions that were most frequently sampled earlier. For example, if a stimulus had "color" and "size" dimensions and a participant sampled the color dimension more frequently, the feature-level personalization method suggested sampling the color dimension first.

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study, and the study follows Journal Article Reporting Standards (Appelbaum et al., 2018). All data, analysis code, and research materials are publicly available via the Open Science Framework (https://osf.io/n2mxu). The code for the Bayesian adaptation of the Wilcoxon rank sum test (van Doorn et al., 2020) is publicly available via the Open Science Framework (https://osf.io/gny35). We did not preregister the design and analyses of this study.

Analysis

Sampling Diversity

One of our interests was whether the personalization algorithm developed selective information sampling compared to the control condition. To investigate this question, we quantified the diversity of sampling patterns using the Shannon entropy.

Let $\tilde{\mathbf{u}} = (\tilde{u}_1, \ldots, \tilde{u}_D)$ $(D = 6)$ represent a vector comprising the sampling rate of the dth dimension (i.e., the number of trials at which a participant sampled the dth dimension divided by the total number of trials). The Shannon entropy for sampling diversity is defined as

$$\text{entropy}_{\text{sampling}} = \sum_{d=1}^{D} \tilde{u}_d \log \frac{1}{\tilde{u}_d} = -\sum_{d=1}^{D} \tilde{u}_d \log \tilde{u}_d. \qquad (1)$$

The maximum possible value of entropy for sampling diversity is achieved when all dimensions are equally sampled, that is, $\tilde{u}_d = 1/6$ for all d:

$$\max(\text{entropy}_{\text{sampling}}) = -\sum_{d=1}^{D} \frac{1}{6}\log\frac{1}{6} \approx 1.79176. \tag{2}$$

Representational Accuracy

Whether participants developed accurate category representations was another interest of our study. Here, we focus on the procedure used for the representational accuracy analysis using confusion matrices obtained from the postlearning categorization task. For the same-different task, readers are referred to the Supplemental Materials.

Quantifying the Representational Accuracy

To understand the representational distortion caused by different learning sequences, we compared confusion matrices derived from the postlearning categorization task between conditions. If a participant learned categories perfectly, the resulting confusion matrix would be an (8 × 8) diagonal matrix

$$\mathbf{M}^* = [m^*_{ij}]_{(8\times 8)} = \begin{bmatrix} 4 & 0 & \cdots & 0 \\ 0 & 4 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 4 \end{bmatrix}, \tag{3}$$

where the (i, j) element of the confusion matrix represents the frequency with which a participant classified a test item from Category j to Category i. Any incorrect decisions will cause deviations from $\mathbf{M}^*$ and represent distortions in the category representation that each participant developed during the learning phase.

The representational distance from the ideal responses was defined as the square root of the sum of squared element-wise error (RSSE): Given a confusion matrix obtained from a participant, $\mathbf{M} = [m_{ij}]_{(8\times 8)}$, and the one from the ideal responses, $\mathbf{M}^*$,

$$d^{(\mathbf{M})} = \sqrt{\sum_{i=1}^{8}\sum_{j=1}^{8}(m^*_{ij} - m_{ij})^2}. \tag{4}$$

A large distance means that the category representation is more severely distorted compared to the standard (i.e., $\mathbf{M}^*$).

RSSE-like measures are less frequently used compared to other accuracy-centered performance metrics (e.g., precision, recall, F-score, Cohen's κ) when analyzing confusion matrices. However, the accuracy-centered performance metrics may not capture the details of *mis*classification that reflects incorrect category representations. As how participants incorrectly categorize test items is just as important as categorization accuracy to us, we think that the RSSE may serve our purpose of comparing category representations. Detailed discussions of different performance metrics are in the Supplemental Material (see also Supplemental Figure S7).

Group Comparison

To compare the degree of representational distortion between the control and noncontrol conditions, we used a Bayesian adaptation of the Wilcoxon rank sum test (van Doorn et al., 2020). Here, we hypothesized that the noncontrol conditions would develop inaccurate representations compared to the control condition. In the confusion matrix analysis, the alternative hypothesis was that the representational distance (i.e., $d^{(\mathbf{M})}$) is smaller in the control condition than in the noncontrol conditions. The null hypothesis is that there is no between-group difference in the representational distance.

For planned comparison, we evaluated the Bayes factor using the Savage–Dickey density ratio (SDDR) method for testing a point hypothesis and its alternative (Wagenmakers et al., 2010). First, a Gibbs sampler drew posterior samples of the latent difference parameterized as δ between the control and noncontrol conditions. The Gibbs sampler used 10 chains, each of which sampled 10,000 posterior samples after discarding the first 2,500 samples as a burn-in phase. After sampling the posterior, we approximated the posterior distribution using logspline density estimation and evaluated the SDDR at $\delta = 0$. As the alternative hypothesis in this analysis was order-restricted (i.e., $H_A\!:\!\delta > 0$), we followed the modified procedures for one-sided hypothesis testing discussed in Wagenmakers et al. (2010). Both the prior and posterior distributions of δ were truncated at $(0, \infty)$ and normalized before evaluating the SDDR. We present the SDDR after base-10 log transformation such that positive values support the hypothesis that nonzero between-group difference exists, while negative values support the null hypothesis. According to Jeffreys (1961), a base-10 log-transformed Bayes factor greater than 0.5 and 1 corresponds to "substantial" and "strong" evidence supporting the alternative hypothesis.

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 $\text{Cauchy}(0, 1/\sqrt{2})$ and set following the default procedure of van Doorn et al. (2020). The second prior, $\text{Cauchy}(0, \sqrt{2})$, was more diffuse than the first one, which could lower the Bayes factor favoring the alternative hypothesis if the evidence is not strong and robust enough. The result was consistent between the two priors. The analysis using the second prior is reported in the Supplemental Materials.

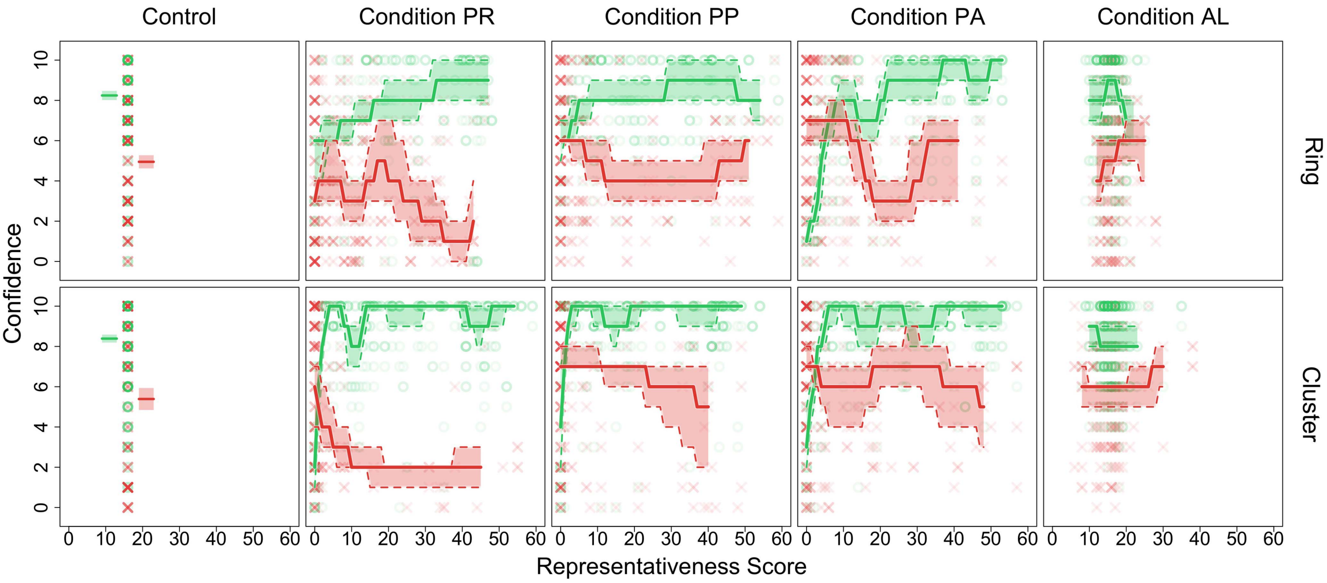
Postlearning Categorization

We used general additive models to describe the smoothed mean trend of accuracy and confidence in the postlearning categorization task (Figure 6). The mean trend lines and their 95% confidence intervals were obtained by fitting general additive models using cubic spline bases for each condition. In the case of confidence ratings, accuracy levels (i.e., correct, incorrect) were also treated independently. The R package mgcv (Wood, 2011) was used for mean trend estimation. The trends of accuracy and confidence were estimated using logistic and cumulative ordinal regression models (Bürkner & Vuorre, 2019; Wood et al., 2016), respectively.

The preliminary smoothing analysis showed that the frequency with which a test item's category was presented during the learning phase ("representativeness score" or S_R from here) is crucial in understanding confidence ratings. To further clarify the relationship between confidence and accuracy conditioned on representativeness scores, we fitted mixed-effect logistic regression models predicting

Figure 6

Postlearning Categorization Confidence Rating (y-Axis) as a Function of the Representativeness Score of the Test Items (x-Axis)



Note. Reported confidence of the individual test items color-coded based on their accuracy is illustrated along with its mean trend for each condition (bold lines). The accuracy of trials was color-coded differently (correct: green empty circles; incorrect: red "×" markers). In both panels, shaded areas indicate the 95% confidence intervals of the estimated trend. The trend lines and associated confidence intervals for the noncontrol conditions are based on the predictions of generalized additive models using cubic spline bases fitted to each condition and accuracy level. Condition PR = personalized item and randomized feature order; Condition PP = personalized item and personalized feature order; Condition PA = personalized item and active feature sampling; Condition AL = active learning. See the online article for the color version of this figure.

item-wise accuracy from confidence ratings. Participants served as a random-effect component.

Fixed-effect variables were confidence ratings (11 levels), learning-sequence conditions (five levels), category structures (two levels; "ring" and "cluster"), representativeness scores (ranging from 0 to 59), and their two- and three-way interactions. Any interactions between the learning-sequence conditions and representativeness scores were not included in the model because they caused a model identifiability issue.

Table 1

Mixed-Effect Models of Postlearning Categorization Accuracy With Varying Degrees in the Complexity of Random-Effect Components

Model index	Random effect term	BF_{x7}	WAIC
1	1	≈0	5440.2
2	C	≈0	5335.4
3	$1 + C$	≈0	5307.1
4	$1 + S_R$	≈0	5107.7
5	$C + S_R$	≈0	5070.4
6	$1 + C + S_R$	≈0	4999.9
7	$1 + C + S_R + C{:}S_R$	1	4850.7

Note. Seven mixed-effect models were compared based on the Bayes factor (BF) and widely applicable information criterion (WAIC). The third column (BF_{x7}) shows the BF supporting Model x compared to Model 7. By definition, the value of BF_{x7} resulting from comparing Model 7 to itself is 1. BF values lower than 1 suggest that Model 7 is preferred over Model x. Models with higher BF_{x7} values and lower WAIC values are preferred. C = confidence ratings; S_R = representativeness scores; $C{:}S_R$ = two-way interaction between confidence ratings and representativeness scores.

We fitted seven model variants with increasing complexity of random effects (Table 1). The R package brms (Bürkner, 2017), which relies on Stan (Stan Development Team, 2024) and the R package bridgesampling (Gronau et al., 2020), was used for model fitting and comparison. We used a normal distribution with a mean of 0 and a standard deviation of 10 as a diffuse prior for all fixed-effect terms. We followed default prior settings in the R package brms for random effect parameters: (a) a nonstandardized half Student's t distribution with a location parameter of 0, a scale parameter of 2.5, and 3 df for standard deviations of each random-effect term and (b) a Lewandowski–Kurowicka–Joe prior with a shape parameter of 1 for correlation between random-effect terms. We sampled 10,000 posterior samples from four chains and discarded the first 5,000 samples from each chain as a warm-up phase, resulting in 20,000 posterior samples.

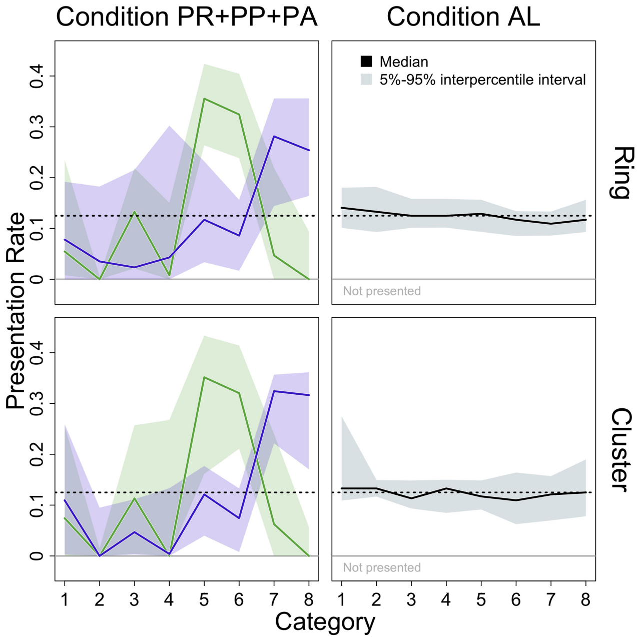
We report the results from Model 7 as it was the best model supported by the Bayes factor and widely applicable information criterion (Watanabe & Opper, 2010). Model 7 had a random intercept, random slopes for confidence ratings (C), representativeness scores (S_R), and their interaction ($C{:}S_R$). The Gelman–Rubin convergence diagnostic ($\hat{R}$; Gelman & Rubin, 1992) was within [0.999, 1.009] across all parameters in Model 7.

Results

Content Diversity

Figure 7 illustrates the category-wise presentation rate per condition. The results show that our algorithm successfully constrained

Figure 7
Personalization Algorithm: Category Presentation



Note. Each panel shows the category presentation rate of noncontrol conditions (top: ring, bottom: cluster; left: Conditions PR, PP, and PA, right: Condition AL). Personalization conditions (PR, PP, PA) were collapsed for brevity. Solid lines represent the median presentation rate. Shaded areas are 5%–95% interpercentile intervals. Personalized conditions (left) show two clusters of participants with different patterns of category presentation, which are differently color-coded. Black dotted lines represent the situation in which all categories are equally sampled. Gray bold lines refer to the zero presentation rate (i.e., a category has never been presented during the learning phase). Condition PR = Personalized item and Randomized feature order; Condition PP = Personalized item and Personalized feature order; Condition PA = Personalized item and Active feature sampling; Condition AL = Active Learning. See the online article for the color version of this figure.

the content diversity (see also Supplemental Figure S2). In both the ring and cluster structures, the personalization algorithm (Condition PR, PP, and PA) limited the diversity of presented categories during the learning phase. Hierarchical clustering revealed two clusters in both structures: one frequently presenting Categories 5 and 6, and the other one focusing on Categories 7 and 8. We also found the similarity in dimension sampling patterns between the two category environments (top left vs. bottom left), although this seems like a coincidence emerged in how our personalization algorithm was trained in each environment. Personalization conditions also show between-participant variability in category presentation, particularly for the categories presented less frequently (i.e., Categories 1–4). By contrast, participants in the active learning condition (Condition AL) sampled all eight categories roughly equally during the learning phase.

Considering how our algorithm works, some categories were presented relatively frequently than others likely because participants were expected to sample the most number of dimensions from them. Our personalization algorithm learned this tendency from Condition AL participants' user profiles and their number of sampled dimensions every trial. Note that, due to their synthetic nature, our stimuli do not have any inherent perceptual or semantic properties leading some categories selectively to higher or lower presentation rates.

Selective Information Sampling

One of our most important questions centered on whether personalization harms the exploration of information. We analyzed the sampling profiles of participants (Figure 8) by calculating the mean number of sampled dimensions per participant (top; see also Supplemental Figure S3) and the Shannon entropy of dimension-wise selectivity (bottom; Equation 1). On the bottom row, lower values of the Shannon entropy suggest that learners did not sample all six dimensions equally, focusing only on a smaller subset.

Participants in Conditions PA sampled the least number of dimensions on average, followed by those in Conditions AL, PP, and PR (top). Conditions PA and AL require active selection of learning materials to various degrees and therefore lead to higher sampling costs in terms of cognitive resources, while Conditions PR and PP simply ask participants to click on the "Sample" button. Higher sampling costs in these conditions seem to guide participants to sample fewer dimensions.

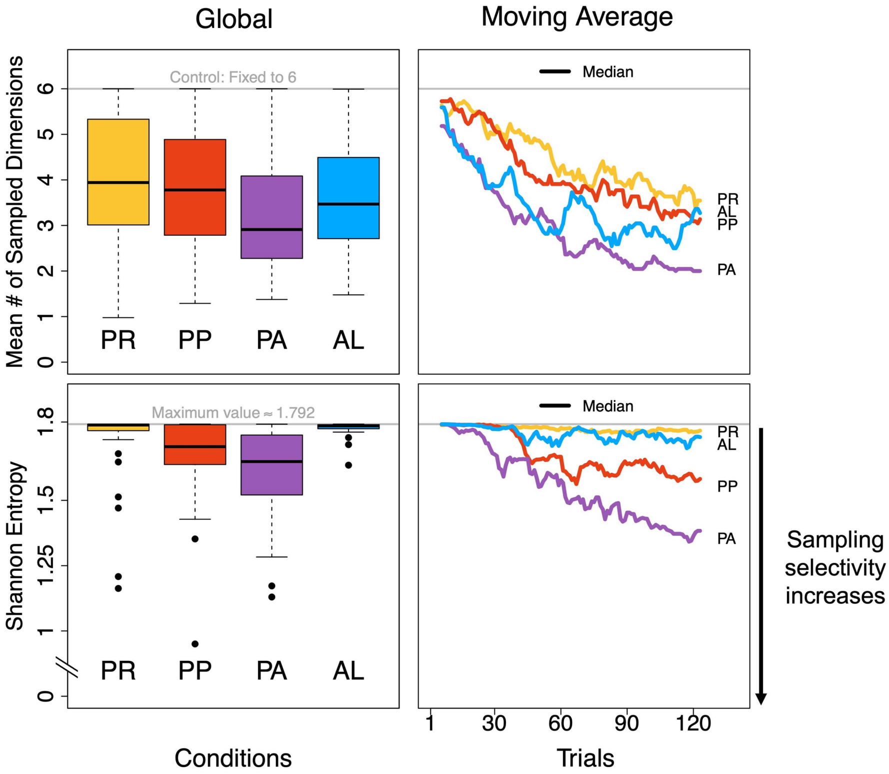
However, the entropy of sampled dimensions decreased over time in Conditions PP and PA, indicating that participants narrowed their exploration. Feature sampling became more selective in these conditions as the learning phase continued (bottom right). By contrast, in Condition AL, when participants could sample freely and without constraint, the entropy stayed relatively high across trials, indicating that participants tended to explore relatively more dimensions. Condition PR, despite its limited item variety, also showed the pattern of equally distributed sampling. This is likely because the randomized feature presentation order led participants to keep sampling to see the information they wanted or simply because the sampling cost was low.

Figure 9 describes how personalization diversifies information sampling patterns. Each panel shows participants' normalized sampling rate in a two-dimensional principal component (PC) space. Compared to Conditions PR and AL in which many participants roughly sampled all six dimensions equally (dark gray areas), participants in Conditions PP and PA deviated from this pattern of equally distributed sampling. For extreme cases on the corners or borders of the PC space, hexagonal shapes show which dimensions were sampled more frequently. These hexagonal figures show that participants in Conditions PP and PA are distributed over the PC space depending on their "preferred" dimensions.

The result shows that personalization of category learning sequences can guide learners to limit the amount and diversity of information learners access. It is noteworthy that personalization of the item-level exposure constrained the feature-level sampling diversity even when learners were able to access all dimensions actively (i.e., Condition PA). Compared to the active learning counterpart in which learners were also able to control items to study (Condition AL), the sampling pattern became more selective when the personalization algorithm controlled the item-level presentation. This comparison reveals that personalizing

Figure 8

Selectivity of the Sampling Profiles During the Learning Phase



Note. The participant-wise mean number of sampled dimensions (top row) and the Shannon entropy of the normalized sampling profile (bottom row). Left and right panels show the corresponding quantity throughout the learning phase (left) and the median moving-window quantities changing over time with a window size of 11 trials. Underlying category structures were collapsed. Conditions are differently color-coded (yellow: PR, red: PP, green: PA, blue: AL). Condition PR = personalized item and randomized feature order; Condition PP = personalized item and personalized feature order; Condition PA = personalized item and active feature sampling; Condition AL = active learning. See the online article for the color version of this figure.

higher level information by computer algorithms is enough to distort how people explore detailed aspects of the environment.

Distortions in Category Representations

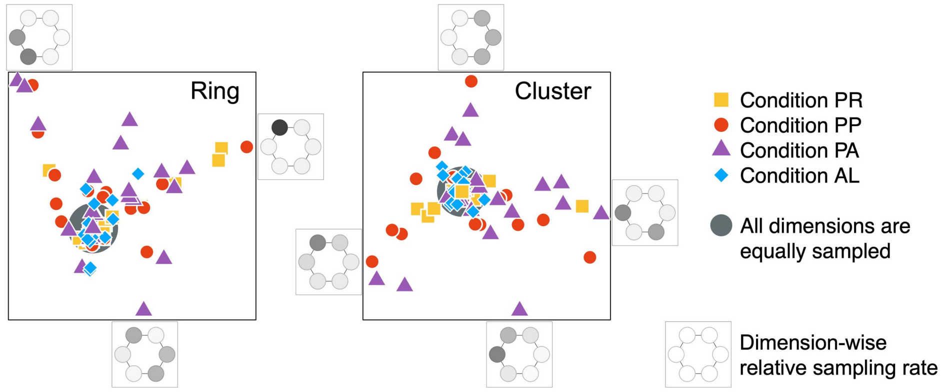
By analyzing how accurate participants were in the postlearning categorization task, we can evaluate the degree to which their category representation was distorted. One way to quantify such representational distortion is to interpret the distance (Equation 4) between confusion matrices obtained from the participants (Figure 10A) and the one assumed by an observer who did not miscategorize any items (Equation 3). A larger distance from the ideal confusion matrix indicates that the developed internal representation was more severely distorted.

We hypothesized that the noncontrol participants would have less accurate category representations compared to those in the control condition. Figure 10B shows the result of planned comparisons using a Bayesian adaptation of the Wilcoxon rank sum test (van Doorn et al., 2020). The Bayes factor associated with all personalization conditions indicated at least "substantial" evidence of representational distortion, with Condition PR (ring) and Conditions PP and PA (cluster) particularly showing "strong" evidence (Jeffreys, 1961). By contrast, active learning conditions tended to support the null hypothesis that there is no representational distortion with respect to the control condition, although the amount of evidence was indecisive in the cluster structure.

The result shows that learners in personalized environments are likely to develop distorted category representations. One could question the source of misclassification: Does it come from random guesses, participants' adherence to learned category structures and labels despite a lack of support, or incorrect generalization based on the belief that misclassification is supported? The major deviation from the no-miscategorization scenario seems to suggest that participants at least did not rely on random guesses. If the responses relied on random guesses, the incorrect responses must be distributed equally across eight categories. However, Figure 10A suggests that this is not the case. In the ring structure, the incorrect responses are dispersed to adjacent categories in feature space (Figure 2B). In the

Figure 9

Between-Participant Variability of the Sampling Profiles During the Learning Phase



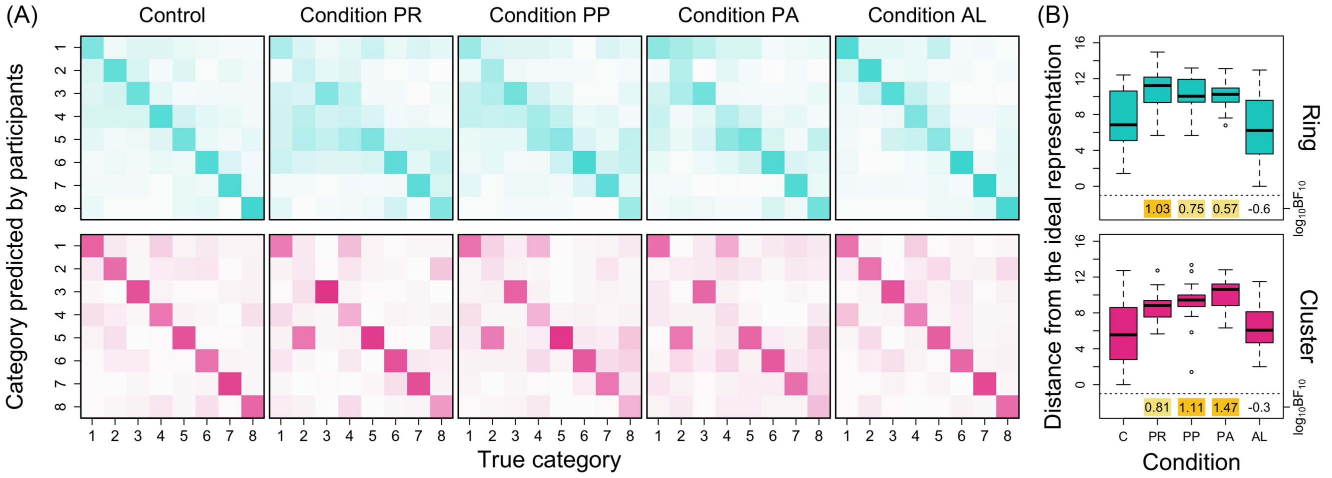
Note. Each panel shows the normalized sampling rate of participants per condition in a two-dimensional principal component space. Hexagonal shapes on the borders visualize the relative sampling rate of feature dimensions at the corresponding location on the principal component space, in which dense-colored dimensions are more frequently sampled. Dark gray circles on the background represent the region in which participants sampled six feature dimensions (approximately) equally. Condition PR = personalized item and randomized feature order; Condition PP = personalized item and personalized feature order; Condition PA = personalized item and active feature sampling; Condition AL = active learning. See the online article for the color version of this figure.

cluster structure, the patterns in misclassification are often strongly biased toward specific category labels (e.g., Category 2 items are often misclassified as Category 5). Both observations imply that the incorrect responses from personalized participants were guided systematically.

Figure 11 illustrates the relationship between the Shannon entropy representing the sampling selectivity of participants (x-axis; Figure 8, bottom left) and the representational distance (y-axis; Figure 10B), collapsing the category structures. Participants with low entropy values (i.e., high sampling selectivity) tend to produce

Figure 10

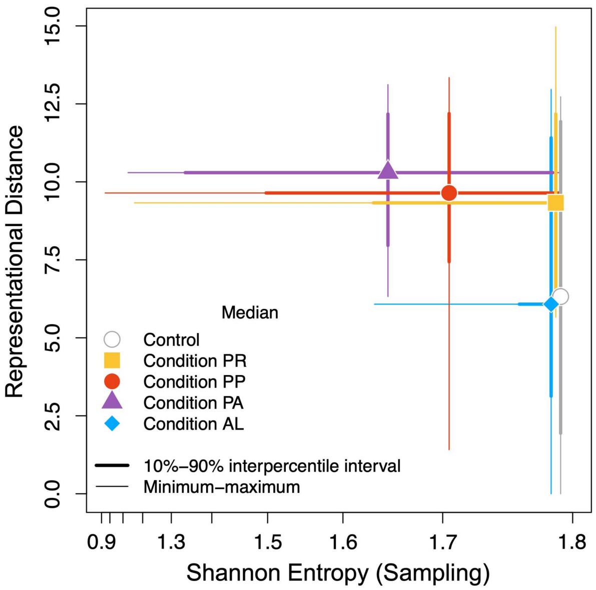
Confusion Matrix Analysis



Note. (A) Condition-wise average confusion matrices. Densely colored cells indicate a higher average response rate. (B) The representational accuracy analysis is based on the squared errors between the ideal and observed confusion matrices. $\log_{10}$ BF_{10} means the base-10 log-transformed Savage–Dickey density ratio in favor of the alternative hypothesis that the noncontrol conditions have a larger representational distance than the control condition. Positive and negative values support the alternative and null hypotheses, respectively. All $\log_{10}$ BF_{10} values were rounded to two decimal places. $\log_{10}$ BF_{10} values greater than 0.5 (light yellow) and 1 (yellow) were highlighted. Condition PR = personalized item and randomized feature order; Condition PP = personalized item and personalized feature order; Condition PA = personalized item and active feature sampling; Condition AL = active learning; BF = Bayes factor. See the online article for the color version of this figure.

Figure 11

Sampling Entropy and Representational Distance



Note. Condition-wise median (points), 10%–90% interpercentile intervals (thick solid lines), and minimum/maximum ranges (thin solid lines) are described for the Shannon entropy of participant-wise sampling profiles (x-axis) and the representational distance computed using confusion matrices (y-axis). Two category structures were collapsed. For visual clarity around the maximum possible sampling entropy ($x \approx 1.792$; Equation 2), the x-axis scale was adjusted by applying power transformation. Condition PR = personalized item and randomized feature order; Condition PP = personalized item and personalized feature order; Condition PA = personalized item and active feature sampling; Condition AL = active learning. See the online article for the color version of this figure.

more incorrect responses, resulting in high representational distance. Condition-wisely, participants in the personalization conditions (particularly, PP and PA) are associated with low entropy and high representational distance.

Generalization and Confidence

We further investigated the nature of misclassification during the categorization test by examining test items from the categories never presented during the learning phase. If the personalized participants relied strictly on sound metacognitive judgment, it may be possible that participants would associate the unfamiliar test stimuli (i.e., the stimuli from categories never shown during the learning phase) selectively with *unfamiliar* category labels (i.e., labels that have not been presented during the learning phase; Figure 12A). This "novel category" response pattern reflects one's attempt to avoid incorrectly extrapolating one's category knowledge.

Figure 12B shows the histograms of the probability that personalized participants gave the "novel category" responses during the categorization test. Condition AL was excluded from this plot because all active-learning participants selected all eight categories at least once during the learning phase. If a learner does not generalize one's category representations hastily, one may choose the "novel category" responses for test items coming from unobserved categories during the learning phase. However, the response probabilities tend to be low, suggesting that personalized participants tend to classify the test stimuli previously not observed to one of the known categories incorrectly. Only a few participants actively used the novel category options. Readers interested in the condition-wise trends are referred to Supplemental Figure S8.

Our interpretation of "novel category" responses during the test phase may be questionable. In principle, labels are mere names of categories and do not explicitly entail any metacognitive properties (e.g., feeling of knowing). Participants may have been hesitant to use the "novel category" options not because of their metacognitive judgment, but due to other reasons (e.g., familiarity with category labels).

However, confidence ratings suggest that participants tend to be confident about their decisions for unfamiliar test items when they make incorrect categorization decisions. Figure 6 shows the confidence rating as a function of the representativeness scores of the test item, along with their smoothed trends. Here, the term "representativeness" refers to the frequency with which the category of each test item was presented during the learning phase. If the representativeness score of a test item is x, it means that the item's category was presented x times during the learning phase. A zero representativeness score means that a test item's category was never shown during the learning phase.

When participants made correct decisions, participants became more confident as the representativeness score of test items increased. When participants made incorrect decisions, (near-)zero representativeness induced a higher confidence rating. This interaction is counterintuitive because being confident about one's decision when there is no supporting evidence (i.e., zero representativeness) is not optimal regardless of the response accuracy.

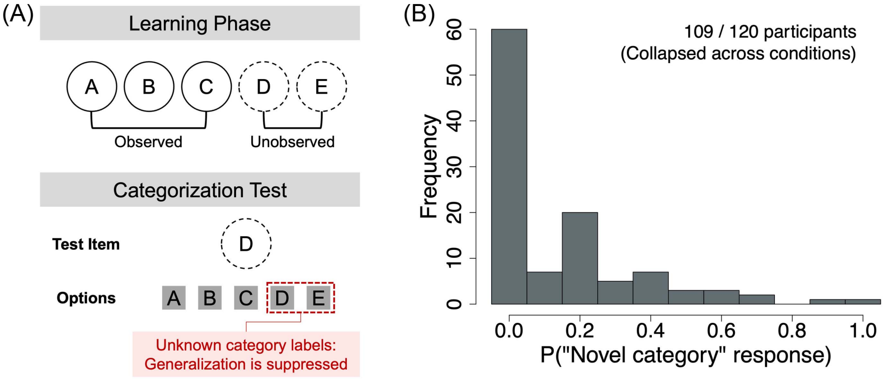
Figure 13 clarifies the trend of overconfidence when test items are from unfamiliar categories by comparing confidence ratings and accuracy directly within the Bayesian mixed-effect logistic regression model. Due to the complexity of the model, we present the condition-wise fixed-effect (thick lines) or participant-wise (thin lines) model predictions. We refer readers interested in fixed-effect coefficients and standard deviation estimates of random effect terms to the Supplemental Materials.

Average model predictions were presented given three different intervals of representativeness scores (S_R): (a) $S_R = 0$, (b) $11 \leq S_R \leq 21$, (c) $S_R \geq 41$. Intervals (a) and (c) show the cases in which test items' categories were never presented ("not presented"; a) or presented very frequently ("over-represented"; c) due to personalization. Interval (b) captures the moderate range, including all Control trials and 90% of Condition AL trials.

When a test item was from the categories that were never presented during the learning phase (top row), higher confidence ratings were associated with lower accuracy in many participants in Conditions PP and PA. However, when the test item's category was presented with moderate frequency (similar to the category presentation frequency in the control condition; middle row), higher confidence ratings were associated with higher accuracy. If the test item's category was presented extremely frequently due to personalization (bottom row), near-perfect categorization accuracy was expected even with a mediocre level of confidence ratings. In summary, Figures 6 and 13 show that participants who learned a limited subset of categories are likely to be overconfident when they attempt to categorize unfamiliar items.

Figure 12

"Novel Category" Responses



Note. (A) A visual illustration of the "novel category" responses. Assume that a learner was not exposed to Categories D and E during the learning phase and, therefore, does not know the presence of these categories. If a test item belongs to one of the unobserved categories (e.g., Category D) and the learner can classify it to one of the unknown categories (i.e., "D" or "E"), this means that the learner refused to generalize one's knowledge to the presented test item incorrectly. (B) A histogram of the probability that participants give the "novel category" responses during the postlearning categorization test. All category structure and learning-sequence conditions were collapsed. The number of participants validly included in the histogram is specified in the top right corner. The control and active learning conditions were not considered because participants observed all eight categories at least once during the learning phase. See the online article for the color version of this figure.

Category knowledge biased by learning-sequence manipulations may have supported incorrect categorization decisions and inflated confidence. Of course, participants' familiarity with category labels might still be an alternative explanation for higher confidence with (near-)zero representativeness. However, classic categorization models based on supervised learning (i.e., category labels are provided to learners) assume that categorization decisions rely on the association between learned materials and category labels (e.g., Kruschke, 1992; Love et al., 2004; Minda & Smith, 2001; Nosofsky, 1986). Even if participants choose a category label due to familiarity, evidence supporting such decisions may not be strong unless novel items are aligned well with the previously learned categorization scheme. Higher confidence ratings for (near-)zero representativeness items suggest the possibility that participants thought their decision was reasonably supported by generalizing biased knowledge.

Discussion

Our results showed that personalization algorithms can severely constrain the knowledge space and distort the representation of category structure. Learners in personalized environments sampled information more selectively than those in the control environments. Selective information sampling employed by learners in personalized environments makes them possibly accurate in frequently exposed categories but inaccurate in unexposed categories. Despite the lack of exposure to some categories, learners in personalized environments were not well calibrated: their confidence for unfamiliar test items was high despite low accuracy.

We investigated the influence of personalization in an unconventional environment—a category learning task that does not involve preferences. Despite the difference between optimizing for learning and optimizing for preferences, we expected this setting to show the emergence of biased and distorted representations purely from personalized experiences, without involving semantics or any belief/value systems. Also, the category learning literature has shown that people reallocate attention based on whether a dimension is helpful in discriminating between categories (e.g., Ashby & Maddox, 2005; Galdo et al., 2022; Kruschke, 1992; Love et al., 2004). This diagnosticity can be seen as a "value-free" analog of preferences.

Personalization, Attention, and Information Search

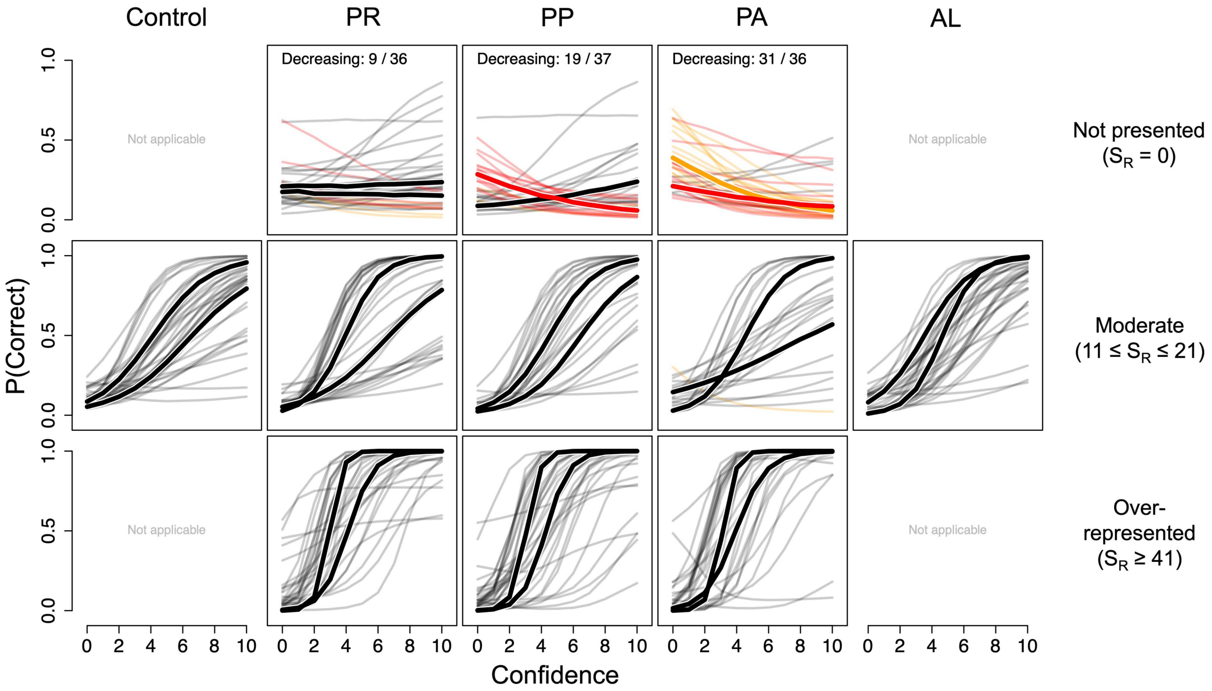
The comparison between the active learning condition and the other three personalized learning conditions emphasizes how personalization induces the development of selective attention (Figure 8). Active learners (Condition AL) continued to sample all six dimensions almost equally, although the number of sampled dimensions per trial slightly decreased over time. However, learners in personalized environments, especially Conditions PP and PA, developed highly selective sampling profiles, focusing on a small set of dimensions.

The contrast in sampling selectivity between Conditions PA and AL indicates that controlling the higher level information (i.e., items; Figure 7) can influence lower level information search (i.e., dimensions; Figure 8). Both conditions had the same degree of freedom in feature sampling while differing in the item selection method. Compared to Condition AL participants who tended to sample all dimensions approximately equally, Condition PA participants developed significantly biased and highly individualized sampling profiles. This result implies that participants came to different conclusions about the informativeness of dimensions due to personalized item sequences.

The category learning literature has shown that learners reallocate their attentional resources to the diagnostic aspects of stimulus information over the course of learning (e.g., Ashby & Maddox, 2005; Galdo et al., 2022; Kruschke, 1992; Love et al., 2004). As participants

Figure 13

Mixed-Effect Modeling: Predictions



Note. Model predictions of correct-response probabilities (*y*-axis) from confidence ratings (*x*-axis) are plotted. The results from two category structures were collapsed. Columns represent learning-sequence conditions. Rows represent three subsets of data based on different ranges of representativeness scores (S_R). Curves represent the participant-wise (thin) or condition-wise fixed-effect (thick) model predictions. Conditions Control and AL were not plotted on the top and bottom rows because there were no applicable trials. Curves showing a (practically) monotonically decreasing trend were colored in orange ("ring") or red ("cluster"). To address minor deviations that are likely due to Monte Carlo noises, we allowed only one increment in correct-response probabilities from 10 possible consecutive increments in confidence ratings when identifying decreasing-trend curves. On the top row, the numbers of participants generating predictions and those yielding decreasing-trend lines per condition are specified in the top left corner of each panel. Condition PR = personalized item and randomized feature order; Condition PP = personalized item and personalized feature order; Condition PA = personalized item and active feature sampling; Condition AL = active learning. See the online article for the color version of this figure.

in the personalization and nonpersonalization groups were exposed to different subsets of categories, different senses of diagnosticity would be developed between the two groups. Having said that, the fact that Condition AL participants sampled all dimensions almost equally throughout the learning phase deviates from the classic prediction that Condition AL participants will find their own optimal subset of dimensions for distinguishing all eight categories. In addition to the explicit goal of learning given to participants, allowing Condition AL participants to choose categories to study may have prevented biased exploration as this approach is unusual in typical category learning experiments.

Postlearning Categorization and Metacognition

The analysis of the postlearning categorization shows that the distorted category representations resulting from personalized learning (Figure 10) could be due to misguided generalization. Participants in the personalized conditions could have considered using the "novel category" cues (Figure 12) more actively when test items came from the categories they have never or rarely observed. The failure to use the "novel category" response strategy and the inflated confidence in incorrect categorization suggest that participants generalized their biased categorical knowledge to unfamiliar test items and did not realize the lack of decision evidence.

Despite the differences in task materials, a study by Hampton et al. (2012) may also support our understanding of the "novel category" options in categorization-related judgments. In that experiment, participants were asked to judge whether given statements were true or false across multiple domains (factual knowledge, category membership, and autobiography). The experiment was conducted twice, 1 week apart, using the same set of statements. Compared to the condition in which only "100% sure" and "100% false" responses were allowed, participants in the condition with the third "unsure" option did not provide more consistent categorization judgment, unlike factual knowledge judgment on which the three-option condition showed improved response consistency. Particularly, participants with three options failed to use the "unsure" option consistently for categorization judgment, compared to factual knowledge judgment. This result implies that it is difficult to be aware of the lack of evidence in categorization.

Although the response uncertainty cue in categorization is not easy to use (Hampton et al., 2012), we might still be able to evaluate different aspects of decision qualities in our study setting by using the "unsure" response option. Despite our interpretation based on the relationship between the strategy use and the inflated confidence, the "novel category" responses may still be influenced by label familiarity. The "unsure" option directly asks participants for metacognitive judgment. Therefore, using it in place of or simultaneously with the "novel category" responses may help us further investigate how incorrect categorization of previously unseen items is related to metacognition, generalization, and label familiarity.

Other Domains of Personalization

In this study, we focused on personalization algorithms that aim to maximize content consumption, as found in online content providers. However, such commercial settings are not the only application of personalization. In education studies, fine-tuning learning experiences has been studied extensively under the theme of adaptive or personalized learning (Bernacki et al., 2021; Martin et al., 2020).

However, we did not pursue direct comparisons between personalization algorithms for commercial settings and those for education. Personalization in education involves various information sources, design features, goals, scales of application, and ownership (Bernacki et al., 2021; Walkington & Bernacki, 2020). What is curated in educational personalization is not limited to content; it also includes tasks, content formats and presentations, instructions, and task rigors. The heterogeneity in personalized education makes it unrealistic to explore the space of algorithm designs comprehensively in this article.

More importantly, formal educational settings have an underlying consensus on clear academic objectives to be achieved, and personalization in education inherits these goals. By contrast, our interest lies in situations where such shared goals or standards are provided only vaguely or even do not exist.

Our interest was in the influence of personalization on basic cognitive processes such as attention and learning. Category learning offers a good test bed because it induces attentional reallocation to informative feature dimensions over time (e.g., Ashby & Maddox, 2005; Galdo et al., 2022; Kruschke, 1992; Love et al., 2004). If participants develop biased and individually distinctive category representations due to personalized learning experiences, we can expect that personalization will not only develop biased beliefs about categories but also modulate attentional reallocation and influence information sampling behavior.

Other studies (e.g., Bodó et al., 2019; Hutmacher & Appel, 2023) have approached human behavior under personalization at the level of attitude (e.g., attitudes toward personalization algorithms) and motivation (e.g., whether the behavior is motivated by one's will and interest). However, while taking attitude and motivation for granted, these studies do not address more fundamental building blocks of information search, such as how the importance, relevance, or diagnosticity of information is evaluated in different learning experiences, and how one's scope of attention and information search changes over such experiences.

Due to our interest in basic cognitive processes, the task using synthetic stimuli serves our purpose well. Other contexts of content personalization (e.g., news, search engine outputs) may improve ecological validity of the task and allow us to explore richer interactions between content and cognition while still following the same principle of curated recommendation. However, these advantages come at the cost of semantically loaded problem contexts that heavily interact with the knowledge and belief systems that people have already developed. Also, the use of more naturalistic and complex media may make quantifying behavior and interpreting results tricky, although not impossible. The modeling tradition in the category learning literature helps us formalize the problem, apply personalization to well-controlled lab-based experiments, and interpret the results more straightforwardly.

Interactive Relationship Between the Algorithms and Users

Earlier studies investigating news consumption claim that filter bubble-like intellectual isolation is mostly due to self-driven confirmatory search induced by people's internal biases, rather than an algorithm (Ekström et al., 2022; Yom-Tov et al., 2014). However, we argue that the relationship between the personalization algorithm and learners interacting with it is not one-directional but rather interactive and that personalization can contribute to the development of initial biases in one's belief system. Our experiment used stimuli with artificial features for which participants could not have any prior knowledge or preferences. Therefore, selectivity in information sampling observed in the personalization conditions can be attributed to the interaction with personalization algorithms, which means that personalization can cause biases in information sampling. We do not intend to hastily generalize our observations to different personalization platforms and contexts (e.g., news providers, search engines), as algorithms being used may control the manner of personalization differently across services. Having said that, our results suggest that the principle of personalization algorithms and the people using them influence each other, forming a positive feedback loop.

The patterns of category choices and feature information sampling in the active learning condition seem to align with the view that people pursue diversity and comprehensiveness in learning and information search. For example, people were positive about news personalization when it was expected to improve the diversity and depth of news (Bodó et al., 2019). However, this interpretation has a limitation in both media studies and our study due to the underlying contexts and goal settings. In real-world news consumption, the pursuit of diversity, not only the attitude toward personalization, may be influenced by one's prior experience with and understanding of media. Being aware of value-ladenness in news consumption (i.e., objective news is difficult to achieve because news producers' observations rely on their own perspectives; Figdor, 2010) may lead (at least some) news platform users to pursue balancedness and comprehensiveness. Also, attitudes toward news consumption may rely on the users' previous experiences with media and their content. Users' prior experience would become a confounding factor when interpreting the influence of algorithms. Our study, on the other hand, used synthetic stimuli to remove confounding from semantics but provided a goal (i.e., learning alien categories) and offered category-level choice options (i.e., labels) explicitly. Also, a list of category options provided to Condition AL participants shows a clear boundary of the feature space to be explored. Both aspects may have encouraged unbiased exploration. The nature of active exploration in the personalization context requires further investigation.

A previous study suggested that the influence of personalization may depend on users' internal motivation (Hutmacher & Appel, 2023). However, it is still possible that such motivation is formed partially through previous experiences with media. Although we do not measure motivation directly, we showed that personalized learning sequences may form selective and individualized information sampling behavior at the level of feature dimensions. Whether personalization builds more biased and selective information sampling (and its underlying motivation) may be investigated further by allowing participants to control where to explore in the feature space after experiencing different learning sequences. We cannot use the setting of the active learning condition (i.e., choosing a category label to study) for this purpose because it will immediately show the discrepancy between one's personalized learning experiences and the complete set of knowledge, leading participants to pursue exploration. This concern calls for different ways of goal setting and exploration for appropriate task development.

Last, we concentrated on the core principle of personalized recommendations—constraining information based on users' interests. However, personalization considering fairness or metalevel goals like users' interest in diversity has become a meaningful research topic (e.g., Celis & Vishnoi, 2017; Eskandanian et al., 2017; Liu et al., 2019; Zhou et al., 2010). As the lab-based experimental studies on personalization involve more real-world contexts and become more semantically loaded, detailed comparisons between personalization algorithms depending on their immediate (i.e., maximizing the probability of content consumption and service use) and metalevel (e.g., diversity pursuit) goals will be important.

Constraints on Generality

We intentionally removed any semantic aspects from our task to separate the influence of human-algorithm interactions from background knowledge and real-world experiences. Although our study might have social implications, the appearance of the experiment follows that of typical lab-based category learning experiments. Also, as the central manipulation in our study was the interaction between humans and computer algorithms, the main premise was that participants could use computers. Our results may be generalized to the population of English-speaking people who are familiar with computers and web-based environments.

The fact that our results come from participants who were willing to complete our cognitively demanding task needs further consideration. We consider that some demanding conditions (e.g., Conditions PA and AL that require active information sampling) were necessary in our study, considering how people interact with online environments. Meanwhile, explicit goal setting (i.e., participants were asked to learn categories) is typical in lab-based experiments but never happens in content-sharing services. A similar task aiming at implicit learning may improve generalizability to real-world contexts.

As stated above, we do not hastily suggest that our results can be generalized to any web-based environments, problem contexts, and service platforms. We showed that the principle of personalized recommendations can hinder optimal learning. Having said that, its influence might also interact with interaction modality (e.g., visual vs. language-based), the level of semantics involved in a problem domain, and users' attitudes toward personalization algorithms and a specific problem domain itself. Future studies incorporating these variables will help elucidate the full landscape of how personalization affects human learning and cognition.

Conclusion

The impact of personalization is concerning because its detrimental effect on selective attention can persist longer and be hard to fix. Many studies have argued that once selective attention is developed, it makes further exploration and redistribution of attentional resources difficult despite their need for correcting false beliefs and representations (Best et al., 2013; Blanco & Sloutsky, 2019; Hoffman & Rehder, 2010; Rich & Gureckis, 2018; Turner et al., 2021). Furthermore, due to the importance of concepts and categories in cognition, the adverse effect of personalization may expand to other cognitive actions. For example, people might use their misguided representation as a basis for stereotyping (Operario & Fiske, 2001; Vinacke, 1957), which would be reinforced by selective sampling and further classification. Personalization algorithms might potentially lead people using them to a trap of biased understanding that continues reinforcing itself.

References

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

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

Bernacki, M. L., Greene, M. J., & Lobczowski, N. G. (2021). A systematic review of research on personalized learning: Personalized by whom, to what, how, and for what purpose(s)? *Educational Psychology Review*, *33*(4), 1675–1715. https://doi.org/10.1007/s10648-021-09615-8

Best, C. A., Yim, H., & Sloutsky, V. M. (2013). The cost of selective attention in category learning: Developmental differences between adults and infants. *Journal of Experimental Child Psychology*, *116*(2), 105–119. https://doi.org/10.1016/j.jecp.2013.05.002

Blanco, N. J., & Sloutsky, V. M. (2019). Adaptive flexibility in category learning? Young children exhibit smaller costs of selective attention than adults. *Developmental Psychology*, *55*(10), 2060–2076. https://doi.org/10.1037/dev0000777

Bodó, B., Helberger, N., Eskens, S., & Möller, J. (2019). Interested in diversity: The role of user attitudes, algorithmic feedback loops, and policy in news personalization. *Digital Journalism*, *7*(2), 206–229. https://doi.org/10.1080/21670811.2018.1521292

Bourgin, D. D., Abbott, J. T., & Griffiths, T. L. (2021). Recommendation as generalization: Using big data to evaluate cognitive models. *Journal of Experimental Psychology: General*, *150*(7), 1398–1409. https://doi.org/10.1037/xge0000995

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Celis, L. E., & Vishnoi, N. K. (2017). *Fair personalization*. arXiv. https://doi.org/10.48550/arXiv.1707.02260

Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for YouTube recommendations. *Proceedings of the 10th ACM conference on recommender systems* (pp. 191–198). Association for Computing Machinery.

Ekström, A. G., Niehorster, D. C., & Olsson, E. J. (2022). Self-imposed filter bubbles: Selective attention and exposure in online search. *Computers in*

Human Behavior Reports, *7*, Article 100226. https://doi.org/10.1016/j.chbr.2022.100226

Eskandanian, F., Mobasher, B., & Burke, R. (2017). A clustering approach for personalizing diversity in collaborative recommender systems. *Proceedings of the 25th conference on user modeling, adaptation and personalization* (pp. 280–284). Association for Computing Machinery.

Figdor, C. (2010). Objectivity in the news: Finding a way forward. *Journal of Mass Media Ethics*, *25*(1), 19–33. https://doi.org/10.1080/08900521003638383

Galdo, M., Weichart, E. R., Sloutsky, V. M., & Turner, B. M. (2022). The quest for simplicity in human learning: Identifying the constraints on attention. *Cognitive Psychology*, *138*, Article 101508. https://doi.org/10.1016/j.cogpsych.2022.101508

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, *92*(10), 1–29. https://doi.org/10.18637/jss.v092.i10

Hampton, J. A., Aina, B., Andersson, J. M., Mirza, H. Z., & Parmar, S. (2012). The Rumsfeld effect: The unknown. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 340–355. https://doi.org/10.1037/a0025376

Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., & Wilson, C. (2013). Measuring personalization of web search. *Proceedings of the 22nd international conference on world wide web* (pp. 527–538). Association for Computing Machinery.

Ho, S. Y., & Bodoff, D. (2014). The effects of web personalization on user attitude and behavior. *Management Information Systems Quarterly*, *38*(2), 497–A10. https://doi.org/10.25300/MISQ/2014/38.2.08

Ho, S. Y., Bodoff, D., & Tam, K. Y. (2011). Timing of adaptive web personalization and its effects on online consumer behavior. *Information Systems Research*, *22*(3), 660–679. https://doi.org/10.1287/isre.1090.0262

Ho, S. Y., & Tam, K. Y. (2005). An empirical examination of the effects of web personalization at different stages of decision making. *International Journal of Human–Computer Interaction*, *19*(1), 95–112. https://doi.org/10.1207/s15327590ijhc1901_7

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General*, *139*(2), 319–340. https://doi.org/10.1037/a0019042

Hutmacher, F., & Appel, M. (2023). The psychology of personalization in digital environments: From motivation to well-being—A theoretical integration. *Review of General Psychology*, *27*(1), 26–40. https://doi.org/10.1177/10892680221105663

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

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(31), 10687–10692. https://doi.org/10.1073/pnas.0802631105

Krafft, T. D., Gamer, M., & Zweig, K. A. (2019). What did you see? A study to measure personalization in Google's search engine. *EPJ Data Science*, *8*(1), Article 38. https://doi.org/10.1140/epjds/s13688-019-0217-5

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44. https://doi.org/10.1037/0033-295X.99.1.22

Le, H., Maragh, R., Ekdale, B., High, A., Havens, T., & Shafiq, Z. (2019). Measuring political personalization of google news search. *The world wide web conference* (pp. 2957–2963). Association for Computing Machinery.

Liu, Q., Reiner, A. H., Frigessi, A., & Scheel, I. (2019). Diverse personalized recommendations with uncertainty from implicit preference data with the bayesian mallows model. *Knowledge-Based Systems*, *186*, Article 104960. https://doi.org/10.1016/j.knosys.2019.104960

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332. https://doi.org/10.1037/0033-295X.111.2.309

Martin, F., Chen, Y., Moore, R. L., & Westine, C. D. (2020). Systematic review of adaptive learning research designs, context, strategies, and technologies from 2009 to 2018. *Educational Technology Research and Development*, *68*(4), 1903–1929. https://doi.org/10.1007/s11423-020-09793-2

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775–799. https://doi.org/10.1037/0278-7393.27.3.775

Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm: An empirical assessment of multiple recommender systems and their impact on content diversity. *Information Communication and Society*, *21*(7), 959–977. https://doi.org/10.1080/1369118X.2018.1444076

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–61. https://doi.org/10.1037/0096-3445.115.1.39

Operario, D., & Fiske, S. T. (2001). *Stereotypes: Content, structures, processes, and context*. Blackwell Publishing.

Pariser, E. (2011). *The filter bubble: What the internet is hiding from you*. Penguin.

Rich, A. S., & Gureckis, T. M. (2018). The limits of learning: Exploration, generalization, and the development of learning traps. *Journal of Experimental Psychology: General*, *147*(11), 1553–1570. https://doi.org/10.1037/xge0000466

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(23), 11537–11546. https://doi.org/10.1073/pnas.1820226116

Stan Development Team. (2024). *RStan: The R interface to Stan* (R package Version 2.32.6) [Computer software]. https://mc-stan.org

Tam, K. Y., & Ho, S. Y. (2006). Understanding the impact of web personalization on user information processing and decision outcomes. *Management Information Systems Quarterly*, *30*(4), 865–890. https://doi.org/10.2307/25148757

Thurman, N., & Schifferes, S. (2012). The future of personalization at news websites: Lessons from a longitudinal study. *Journalism Studies*, *13*(5–6), 775–790. https://doi.org/10.1080/1461670X.2012.664341

Turner, B. M., Kvam, P. D., Unger, L., Sloutsky, V., Ralston, R., & Blanco, N. J. (2021). *Cognitive inertia: How loops among attention, representation, and decision making distort reality*. PsyArXiv. https://doi.org/10.31234/osf.io/8zvey

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. https://doi.org/10.1080/02664763.2019.1709053

Vinacke, W. E. (1957). Stereotypes as social concepts. *The Journal of Social Psychology*, *46*(2), 229–243. https://doi.org/10.1080/00224545.1957.9714322

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Walkington, C., & Bernacki, M. L. (2020). Appraising research on personalized learning: Definitions, theoretical alignment, advancements, and future directions. *Journal of Research on Technology in Education*, *52*(3), 235–252. https://doi.org/10.1080/15391523.2020.1747757

Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(12), 3571–3594.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*(516), 1548–1563. https://doi.org/10.1080/01621459.2016.1180986

Yesilada, M., & Lewandowsky, S. (2022). Systematic review: YouTube recommendations and problematic content. *Internet Policy Review*, *11*(1), Article 1652. https://doi.org/10.14763/2022.1.1652

Yom-Tov, E., Dumais, S., & Guo, Q. (2014). Promoting civil discourse through search engine diversity. *Social Science Computer Review*, *32*(2), 145–154. https://doi.org/10.1177/0894439313506838

Zhou, T., Kuscsik, Z., Liu, J.-G., Medo, M., Wakeling, J. R., & Zhang, Y.-C. (2010). Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(10), 4511–4515. https://doi.org/10.1073/pnas.1000488107

Appendix

Stimulus Feature Generation

We used a procedure to impose underlying structures in the distribution of stimulus categories (Kemp & Tenenbaum, 2008; Saxe et al., 2019). We describe the method for generating a covariance matrix (Saxe et al., 2019) from which we sampled learning- and test-phase items.

For the ring structure, we first generated an adjacency matrix $\mathbf{A} = [a_{ij}]_{(8\times 8)}$. $a_{ij} = 1/e_{ij}$ ($e_{ij} \neq 0$) means that Categories i and j are nearby to one another, or in terms of graph theory, that Categories i and j ($i, j = 1, 2, \ldots, 8$) are connected. Using $e_{ij} = 1/0.7$ for all i and j, we set

$$\mathbf{A} = \begin{bmatrix} 0 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0.7 \\ 0.7 & 0 & 0.7 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.7 & 0 & 0.7 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.7 & 0 & 0.7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.7 & 0 & 0.7 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.7 & 0 & 0.7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.7 & 0 & 0.7 \\ 0.7 & 0 & 0 & 0 & 0 & 0 & 0.7 & 0 \end{bmatrix}, \tag{A1}$$

which makes Categories 1 and 2, 2 and 3, … , 7 and 8, and finally 8 and 1 are "adjacent." Then, a category-level covariance matrix $\mathbf{\Phi}^*$ was defined as $\mathbf{\Phi}^* = \{(\mathbf{D} - \mathbf{A}) + \frac{1}{\sigma^2}\mathbf{I}\}^{-1}$ where $\mathbf{D} = [d_{ij}]_{(8\times 8)}$ is a diagonal matrix with $d_{ii} = \sum_{j=1}^{8} a_{ij}$, $\mathbf{I}$ is an identity matrix, and $\sigma^2 = 1/0.09$. The values of e_{ij} and σ^2 were set following Saxe et al. (2019).

To sample all items from the covariance matrix $\mathbf{\Phi}^*$ simultaneously so that they share the same centroids, we defined a $K \times 8$ matrix $\mathbf{M}$ where K is the total number of samples we draw. If the kth sample ($k = 1, \ldots, K$) is from Category i, the kth row of $\mathbf{M}$ is filled with zeros except for its ith element being one. The actual item-level covariance matrix we used to get K items is $\mathbf{\Phi} = \mathbf{M}\mathbf{\Phi}^*\mathbf{M}^T$ where a superscript T is a transpose operator. We arbitrarily sampled 40 items per category (i.e., $K = 320 = 40 \times 8$). The first four items per category served as learning-phase items, and the second four items as test-phase items.

For the cluster structure, the item-level covariance matrix $\mathbf{\Phi}$ was set as a block-diagonal matrix consisting of diagonal blocks $\mathbf{\Phi}_i$ ($i = 1, \ldots, 8$). The inverse of the category-level block of $\mathbf{\Phi}$, $\mathbf{\Phi}_i^{-1}$, is defined as $\mathbf{\Phi}_i^{-1} = c_1^{(i)}\mathbf{1} + c_2^{(i)}\mathbf{I}$ where $\mathbf{1}$ is a $K_i \times K_i$ matrix filled with ones, and K_i is the number of items we sample from Category i. c_1 and c_2 are constants defined as

$$c_1^{(i)} = \frac{\sigma^2}{K_i + 1} + \frac{K_i - 1}{\left(\frac{1}{e_i} + \frac{1}{\sigma^2}\right)K_i} + \frac{1}{\left(\frac{K_i + 1}{e_i} + \frac{1}{\sigma^2}\right)K_i(K_i + 1)},$$
$$c_2^{(i)} = \frac{\sigma^2}{K_i + 1} - \frac{1}{\left(\frac{1}{e_i} + \frac{1}{\sigma^2}\right)K_i} + \frac{1}{\left(\frac{K_i + 1}{e_i} + \frac{1}{\sigma^2}\right)K_i(K_i + 1)}. \tag{A2}$$

As in the ring environment, K_i was set to 40. e_i was set to 0.24, following Saxe et al. (2019). Unlike other parameters, which we adopted from the original method, we set $\sigma^2 = 0.04$.

For the crosscutting structure, we assumed that four upper-level categories are generated from a hierarchical tree structure in which a root branches into two sublevels, each of which branches again into two sublevels. Last, each of the four upper-level categories was divided into two groups, assuming that the same separation factor applies to all four categories. For example, we can divide a large "animals" category into "mammals" and "birds" sublevels, which can be divided again into "cats," "monkeys," "pigeons," and "penguins." Each category can be separated into "male" and "female" groups, resulting in eight categories. This structure was defined based on a matrix $\mathbf{\Sigma}_{yx}$ that describes whether each category (columns) has a certain feature (row). Following an example case used in Saxe et al. (2019), we set $\mathbf{\Sigma}_{yx}$ as

$$\mathbf{\Sigma}_{yx} = \frac{1}{8}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1.1 & 1.1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.1 & 1.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.1 & 1.1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1.1 & 1.1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}. \tag{A3}$$

(*Appendix continues*)

The item-level covariance matrix was obtained by $\mathbf{\Phi} = \mathbf{M}\mathbf{\Sigma}_{yx}^{T}\mathbf{\Sigma}_{yx}\mathbf{M} + \sigma^2\mathbf{I}$ where $\sigma^2 = 0.001$.

When raw feature values for each category structure were obtained, the values were shifted and scaled so that minimum and maximum feature values within a set of sampled items would be 0 and 1, respectively. To generate visual stimuli, the transformed feature values *f* were used as follows. For the location on a straight line, the value of *f* was mapped onto a horizontal line that takes zero and one as its left and right extremes, respectively. The radius was determined as $(0.5 - 0.05)f + 0.05$. As for the brightness, we translated brightness on a grayscale into the degree of transparency or "α level" of a black rectangular for ease of feature manipulation. If $f = 0$, the color of the rectangular is not affected and looks black on a monitor. If $f = 1$, the color of the rectangular becomes transparent and therefore looks white on a monitor. For the orientation, a shape of dripping water was rotated $(90 + 180f)$ degrees counterclockwise. For the curvature, we used a rectangular hyperbola defined as $y = (10^f)^2/(2x)$, $x \in [0, 10]$. Last, the spatial frequency was defined as the frequency of a sine wave, which was set as $\{(10 - 0.5)f + 0.5\}$.

Received March 17, 2024
Revision received February 14, 2025
Accepted February 25, 2025 ■