Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prior preferences interfere with the associative learning of food values

Alexandra Rich
Yale University https://orcid.org/0000-0002-3346-2426
Sohum Kapadia
Yale School of Medicine
Ryan Henry
Yale School of Medicine
Ohad Dan
Yale School of Medicine
Ifat Levy
ifat.levy@yale.edu
Yale School of Medicine https://orcid.org/0000-0001-7851-9070

Article

Keywords:

Posted Date: February 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8651706/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Communications Psychology on October 9th, 2026. See the published version at https://doi.org/10.1038/s44271-026-00533-5.

Abstract

Addressing unhealthy behaviors requires learning and implementing healthier options. Ecologically, we rarely undergo associative learning in naive environments; our prior preferences confound the learning process. This is perhaps especially true for food, for which we have strong, diverse preferences that may undermine the associative learning policies central to dietary interventions. This preregistered study investigates the flexibility of updating food values using a probabilistic reversal learning task where participants predicted outcomes associated with food stimuli varying in personal preference. The task alternated between “Congruent” blocks, reinforcing pre-existing associations, and “Incongruent” blocks, reversing them. Across independent cohorts, we found that starting condition had a surprising, lasting effect on food-value association updating, with initial reinforcement of food-value priors potentiating learning inflexibility. Rigidity was further evident for liked food items, whereas disliked foods were adaptable to changing value associations. Additionally, associations with positive outcomes were learned more readily than negative ones, suggesting positive reinforcement may be more effective in informing food-based learning. These findings highlight the importance of reinforcement in shaping food-value associations and underscore its relevance to promoting healthier eating habits.

Introduction

Food choices are driven by the subjective value of food; some individuals love olives, while others do not. Some people consider health implications when choosing food, while others base their decisions largely on taste[1]. While high-calorie food is considered highly palatable[2], food valuation goes beyond caloric value. For example, some strongly prefer strawberries to a calorie-dense croissant, or a croissant to a calorically similar granola bar. The subjective, or inherent, value of food items is constructed from previous reward experiences and is mechanistically interconnected with taste, texture, smell, and post-ingestive consequences of eating[3–5]. These preferences are diverse, strong, and resistant to change, yet they are often maladaptive. Understanding how the inherent value of food may be flexibly updated is therefore of high significance for public health.

Studies probing value learning typically use monetary and other rewards to assign value to inherently neutral stimuli (e.g., shapes) and examine the flexible updating of those values which were acquired in the laboratory[6–14]. Prior research has indicated that inherent values can be modified for highly preferred food items[15], yet there is limited understanding of the extent and the flexibility of associative learning processes underlying such changes. Specifically, it is unknown how flexibly we can accomplish a reversal of valence, or updating from positive to negative associations and vice versa, concerning both preferred and non-preferred food items.

Addressing unhealthy behaviors relies upon successful learning and implementation of healthier options. Obesity with binge eating is linked to impaired reward processing and cognitive flexibility, potentially due to alterations in how past reward experiences are integrated into present decisions[16,17]. Flexibly updating subjective food values in favor of healthier choices defines dietary interventions, and

patients with comorbid obesity and binge eating often exhibit lower efficacy rates with these interventions compared to individuals with obesity alone[18,19]. Inherent food values, rather than those learned or assumed from caloric content, motivate food choices in healthy individuals and may play a critical role in disordered eating behavior.

Our primary aim in this study was to investigate the flexibility of updating inherent food values, utilizing a probabilistic reversal-learning task. The task required participants to predict stimuli-coupled outcomes using food stimuli that varied in personal preference from strongly disliked to strongly liked. The design alternated between "Congruent" blocks, reinforcing pre-existing food-value associations, and “Incongruent” blocks, reversing them. Exploratory analyses were performed in an initial cohort before they were preregistered and tested in a second cohort. We hypothesized that inherent preferences would interfere with the associative learning of food values, such that trials Congruent with one’s preferences would exhibit higher prediction accuracy.

We found that, while prediction accuracy was greater on Congruent than Incongruent blocks overall, initial reinforcement of food-value priors was surprisingly relevant and potentiated learning inflexibility. Interestingly, value associations of Disliked food items were suggestible for behavioral changes, whereas participants tracked changing associations inefficiently for Liked foods. Moreover, participants more readily learned about associations drawn between personally salient food items and positive outcomes as opposed to associations with negative outcomes, suggesting that rewarding information about food (e.g., positive health outcomes) may more effectively influence food-based learning. These insights underline the pivotal role of food-value reinforcement and its implications for food-based learning and dietary behavior modification strategies.

Results

Probabilistic reversal learning with personally salient foods

We recruited online participants through Prolific (see Methods) to complete a probabilistic reversal-learning task integrating personally salient food stimuli. As part of a preregistered study (https://osf.io/xjkt7/), participants were recruited in two independent cohorts, and analyses were repeated in each cohort. Participants first rated 65 snack items ranging in nutritional content (Fig. 1a)[20], from which 4 food stimuli, one of each valence level (strongly disliked, disliked, liked, and strongly liked) were randomly selected (Fig. 1b). Participants then completed a learning task in which they were asked to predict outcomes associated with their unique 4 foods (Fig. 1c). Outcomes, ‘+5’ or ‘−5’ arbitrary point values, were determined by the block type in a probabilistic (90/10) reversal design: In "Congruent" blocks, participants’ subjective value ratings of the food items aligned with outcomes 90% of the time, such that liked items were associated with ‘+5’ point outcomes and disliked items were associated with ‘−5’ point outcomes. In "Incongruent" blocks, this relationship was reversed, requiring flexible updating of food-value associations. Participants completed 6 counterbalanced blocks, 3 of each block type, with half of the participants randomly assigned to each of the two start orders. Based on pilot testing (see

Methods, Supplemental Fig. 1), each block contained jittered 32–36 trials, such that each food item was presented 8–9 times per block. Participants received a bonus payment if their predictions were correct on 75% or more of trials.

Importantly, task instructions were explicit: Participants were informed of the probabilistic nature of the task and that associations between food items and outcomes would “switch a number of times throughout the task.” While participants knew that a changing relationship between the food items and point outcomes existed, they were blind to the basis of selection for the 4 food items and the Congruent and Incongruent policy of the task. Participants were also not informed of when a reversal would occur, as blocks were completed in succession without breaks.

Participants who did not have at least one food item in each non-neutral Likert category were excluded. This resulted in the complete sample included in analysis, representing both cohorts of data, of 279 participants (Age: 37.62 ± 12.33 years (mean ± s.d.), BMI: 27.18 ± 6.44 kg/m^2 (mean ± s.d.), 49.46% female). Stimuli ratings obtained from the full recruited sample (N = 415) indicate that the food items exhibited heterogeneous ratings within and across individuals (Supplemental Fig. 2a). Moreover, average ratings for food items were not correlated with caloric or nutritional value (ps > 0.05; Supplemental Fig. 2b). These data both challenge the utility of nutritional content alone to ascertain food value and exhibit heterogeneity indicative of the expected individual differences in valuation. This variability emphasizes that we are capturing subjective value rather than objective properties of the food items.

Initial reinforcement of food-value priors potentiates learning inflexibility

To quantify performance across the course of the task, we conducted trial- and block-wise analyses of average prediction accuracy (Fig. 2). Participants across both independent cohorts of data can predict associations that challenge food-value priors, though there was a significant main effect of block type on prediction accuracy ($F(1, 554) = 33.78$, $p < 0.001$). A Mann-Whitney U test indicated prediction accuracy was greater on Congruent blocks (*Mean*=.68±.15, s.d.) than Incongruent blocks (*Mean*=.61±.16, s.d.) overall ($U = 48886.5$, $Z = 5.23$, $p < 0.001$). Surprisingly, initial block identity had a significant impact on learning quality ($F(1, 554) = 8.17$, $p = 0.004$). Overall accuracy was higher for participants who started in a Congruent block (*Mean*=.66±.13, s.d.) compared to those who started in an Incongruent block (*Mean*=.62±.14, s.d.; $U = 11359.5$, $Z = 2.44$, $p = 0.01$). There was a significant interaction between block type and start block ($F(1, 554) = 31.53$, $p < 0.001$), which was replicated across both cohorts of data (Cohort A, $F(1, 208) = 13.62$, $p < 0.001$, Fig. 2a, left panel; Cohort B, $F(1, 342) = 17.84$, $p < 0.001$), Fig. 2a, right panel). Moreover, such analyses that take the entire block into account are conservative because they average out the performance across both the learning and adapted phases, potentially minimizing observable differences between conditions due to the inclusion of early trial variability: Results were replicated when only the last 15 trials of each block were included ($F(1, 554) = 33.78$, $p < 0.001$ for block type; $F(1, 554) = 8.17$, $p = 0.004$ for start group; $F(1, 554) = 31.53$, $p < 0.001$ for the interaction; Supplemental Fig. 3).

Post hoc comparisons using Tukey's HSD test indicated that participants starting in a Congruent block, where food-value priors were reinforced, exhibited greater prediction accuracy on Congruent (*Mean*=.73±.14, s.d.) than Incongruent blocks (*Mean*=.59±.17, s.d.), 95% CI [-.190, – .098], $p < 0.001$ (Fig. 2b). Accuracy on Congruent blocks in the Congruent-start group was also greater than the accuracy of both Congruent (*Mean*=.62±.15, s.d., 95% CI [-.156,-.062], $p < 0.001$) and Incongruent blocks (*Mean*=.62±.14, s.d., CI[-.155,-.062], $p < 0.001$) in the Incongruent-start group. That is to say, participants who experienced initial reinforcement of their priors were able to achieve a higher average prediction accuracy on later blocks that reinforced their priors (Fig. 2c), above and beyond all other combinations of start group and block type. Interestingly, while start groups did not differ in their average accuracy on Incongruent blocks, variance of prediction accuracy on Incongruent blocks in the Congruent-start group was significantly higher than in the Incongruent-start group (Fligner-Killeen test, $X^2(277) = 7.07$, $p = 0.008$). This indicates that the effect of starting in the Congruent condition on Incongruent block performance is heterogeneous: For some subjects, starting in the Congruent condition impaired learning on Incongruent blocks, but for others it did not. Given this violation of our prior assumption of equal variances, Games-Howell post hoc testing was performed, which validated all aforementioned Tukey HSD findings ($ps < 0.001$).

Participants predict reward associations better than punishment associations, regardless of food preferences

To determine whether prediction accuracy for rewards (‘+5’ outcome) differed from accuracy for punishments (‘–5’ outcome), we conducted a series of Mann-Whitney U tests (Fig. 3). These tests compared accuracy across items associated with rewards and punishments within each combination of start group and block type. Prediction accuracy was similar among strongly liked and liked items, as well as among strongly disliked and disliked items; accordingly, the four stimuli were grouped into Liked (strongly liked, liked) and Disliked (strongly disliked, disliked) categories for this analysis. Regardless of block type and start group, accuracy for reward-associated food (*Mean*=.69±.12, s.d.) was significantly higher than for punishment-associated food (*Mean*=.61±.17, s.d.), $U = 49085.0$, $p < 0.001$. Thus, on Congruent blocks, accuracy for Liked items was higher than for Disliked items, while on Incongruent blocks, accuracy for Disliked items was higher than for Liked items.

Associations of disliked items are suggestible to change

To evaluate the effect of food item rating, and its interactions with start group and block type, on prediction accuracy, we performed a three-way ANOVA (Fig. 4). As expected, there were significant main effects of block type ($F(1, 1108) = 87.22$, $p < 0.001$), start group ($F(1, 1108) = 12.53$, $p < 0.001$), and their interaction ($F(1, 1108) = 37.68$, $p < 0.001$). There was no significant main effect of food item rating (Supplemental Fig. 4, $p = 0.37$). There was, however, a significant interaction between block type and food item rating ($F(1, 1108) = 19.51$, $p < 0.001$). Post hoc comparisons using Tukey's HSD test revealed that for Liked items, prediction accuracy on Congruent blocks was greater than on Incongruent blocks (Fig. 4a), illustrating a significant influence of positive priors on associative learning flexibility. This was

true of both start groups (Congruent-start: Congruent blocks (*M*=.78±.12), Incongruent (*Mean*=.55±.22, s.d.), 95% CI [-.28, − .18], $p < 0.001$; Incongruent-start: Congruent (*Mean*=.68±.15, s.d.), Incongruent (*Mean*=.57±.20, s.d.), 95% CI [-.16, − .05], $p < 0.001$. In contrast to these findings for Liked items, associative learning involving Disliked food items appeared suggestible, such that participants performed better on block types that aligned with their start block (Fig. 4b): Congruent-start subjects predicted negative outcomes in association with Disliked items (*Mean*=.69±.19, s.d.) with higher accuracy than positive outcomes (*Mean*=.63±.17, s.d.), 95% CI [-.11,-.01], $p = 0.02$, and Incongruent-start subjects predicted positive outcomes in association with Disliked items (*Mean*=.68±.13, s.d.) with higher accuracy than negative outcomes (*Mean*=.57±.20, s.d.), 95% CI [.05,.16], $p < 0.001$. That is to say, participants demonstrated notable susceptibility to start block effects when learning outcomes associated with Disliked food items, whereas Liked food item associations remain inflexible.

To ensure that participants were learning and not simply reporting their preferences, a Wilcoxon signed-rank test was used to determine the accuracy in each block relative to chance level. Below-chance performance on Incongruent blocks, in particular, would suggest that participants are reporting their preferences rather than learning. Over the course of the task, prediction accuracy was largely maintained for both Liked (Fig. 5a) and Disliked items (Fig. 5b), but above-chance on both block types (*ps* < 0.05). Interestingly, when strongly liked and liked items are analyzed separately, participants merely achieve at-chance accuracy on some Incongruent blocks, particularly later in the task and when starting in a Congruent block (Supplemental Fig. 5). However, on average, participants did not demonstrate significantly below-chance performance for strongly liked and liked items on Incongruent blocks, suggesting a more limited but supportable capacity for learning a policy that conflicts with one's positive priors.

Differential learning slopes and thresholds by start group and block type

To investigate within-block learning dynamics, we fitted a sigmoid function to average prediction accuracy across subjects for each experimental condition, with some upper asymptote, or threshold, and some slope (see Methods; Fig. 6). A higher threshold would indicate the subject is able to achieve a higher prediction accuracy on average in that condition across the task, while slope quantifies some rate at which the subject is able to achieve that threshold. Concerning the sigmoid shape, learning within blocks is evident for both start groups across all block type and food item rating combinations. However, a two-sample t-test revealed a significant difference in learning slopes (k) between start groups ($t(1114) = 2.53$, $p = 0.01$), such that Congruent-start subjects had a significantly greater slope (k = 2.19) compared to Incongruent-start participants (k = 1.87). Within start groups, slope remained unaffected by condition; however, different thresholds of prediction accuracy were reached as a function of block type, the interaction between block type and food item rating, and the interaction between block type and start group (Supplemental Fig. 6). We conducted an ANOVA to compare the learning thresholds (L) across all conditions. There was a significant main effect of block type on learning threshold ($F(1, 1108) = 42.59$, $p < 0.001$). There were also significant interactions between block type and start group ($F(1, 1108) = 17.66$, $p < 0.001$) and between block type and food item rating ($F(1, 1108) = 70.58$, $p < 0.001$). Post hoc analysis

using Tukey's HSD identified that, for Incongruent-start participants, greater thresholds were achieved for + 5 outcome-associated conditions, consistent with the finding that participants better predicted reward associations. This effect is reflected within-block among Congruent-start participant learning thresholds as well but is complicated by their significantly greater performance on average on Congruent block types, regardless of image rating or associated outcome.

Start condition has cumulative effects on behavioral measures of valuation

To probe potential changes in food item valuation related to task manipulation, we asked participants to rate their unique 4 non-neutral items on scales of willingness to pay, enjoyment, and expected satisfaction before and after performing the learning task. For participants who completed post-task ratings (n = 274), we conducted paired t-tests to compare ratings before and after the task for each food item across each dimension (Fig. 7a-c). We found that for Incongruent-start participants (n = 131), Disliked items experienced a significant increase in average rating across all three scales of willingness to pay ($t(261)=-2.92$, $p = 0.004$), enjoyment ($t(261)=-2.79$, $p = 0.006$), and expected satisfaction ($t(261)=-2.81$, $p = 0.005$) following task completion. Incongruent-start participants also reported decreased enjoyment ($t(261) = 2.27$, $p = 0.02$) and expected satisfaction ($t(261) = 2.36$, $p = 0.02$) following task completion for Liked items, though this observed decrease was not robust across all measures of valuation. This was not the case for Congruent-start participants (n = 143), who only demonstrated an increase in expected satisfaction for Disliked items ($t(285)=-2.70$, $p = 0.007$), and no significant changes among Liked items. Taken together, these results suggest that starting in a condition that challenges one's priors is associated with cumulative changes in behavioral measures of valuation. Interestingly, this appears more robust in the case of increased valuation of Disliked items, which is consistent with the finding that associative learning involving Disliked food items was particularly prone to manipulation.

To assess if certain demographic variables related to learning, we performed exploratory correlational analyses probing the role of binge-eating severity and Body Mass Index (BMI, kg/m^2) in task performance. Binge-eating severity was assessed using Binge-Eating Scale (BES) Global Scores[21]. Correlations with average accuracy across each condition (Congruent/Incongruent block type and Liked/Disliked food item rating combinations) were tested within each start group. To minimize Type I and Type II error rates, we transformed average accuracy and BES Global Score data to rankit scores before performing Pearson correlations[22]. Interestingly, for Congruent-start participants, prediction accuracy for Disliked items on Incongruent blocks was negatively correlated with global BES scores ($r(140)=-.197$, $p = 0.019$; Supplemental Fig. 7a), such that higher levels of binge eating were associated with lower prediction accuracy when primed by one's priors and learning from positive outcomes in the context of Disliked food items. This relationship remained significant in a linear regression controlling for age, gender, and BMI (Supplemental Fig. 7b): BES Global scores were a significant predictor of accuracy (β=-0.004, 95% CI [-0.008, -0.001], $p = 0.013$), in addition to female gender ($\beta = 0.061$, 95% CI [0.004, 0.117], $p = 0.037$). These exploratory results suggest that reward learning, particularly when primed by one's priors or learning about non-preferred foods, may be more complicated for those who binge eat.

Discussion

In this study, we used a probabilistic reversal-learning task to investigate the flexibility of updating inherent food values. Participants were asked to predict point outcomes that followed the presentation of personally liked or disliked food stimuli, with conditions alternating between "Congruent" and “Incongruent” blocks; pre-existing food-value associations were either reinforced or challenged, respectively. While participants were clearly performing above chance, overall learning was low, despite a simple design, including a 90/10 probability distribution, and explicit instructions about the structure of the task. We discovered that inherent preferences interfered with the associative learning of food values, such that blocks Congruent with one’s preferences exhibited higher prediction accuracy than Incongruent blocks. This result was driven by Congruent-start participants, who experienced initial reinforcement of their food-value priors. Moreover, regardless of start group, trials including Liked food items were characterized by greater accuracy in the Congruent condition. Importantly, this was not the case for Disliked food items, for which participants demonstrated notable susceptibility to start block effects when learning outcomes. This suggestible nature of food-value learning for Disliked food items was supported by pre-post task behavioral changes in subjective valuation measures in the Incongruent-start group. We also explored the influence of positive and negative outcome assignments on food-based learning. Despite the somewhat arbitrary nature of the task outcomes (i.e., non-monetary points), our analysis revealed a positivity bias: Participants more readily learned about associations drawn between personally salient food items and positive outcomes as opposed to associations with negative outcomes, regardless of food valence. Altogether, these results emphasize the influence of food preferences and positive outcomes on flexible learning of a policy.

These results contribute to a more ecologically relevant and nuanced interpretation of food-based learning literature, which has often relied upon caloric or nutritional value to assume food preference and has not probed the flexibility of associative learning with salient food cues. A significant body of work has investigated how one’s preferences are learned or influenced through internal, external, cultural, and contextual factors[23–34]. However, our study aims to characterize the opposite relationship: How do our preferences shape our ability to learn about our changing environments? Our findings suggest that inherent preferences interact with intervention design to substantially influence learning quality.

Initial reinforcement vs. initial challenge of prior preferences

The observed order effect, while not hypothesized and surprising in its strength, may overlap with several known cognitive phenomena. One potential mechanism is the anchoring effect, where judgments are persistently biased toward an initially presented value[35]. Like confirmation biases, anchoring serves as confirmatory hypothesis testing, where a particularly plausible association is tested for correctness[36–40]. Participants may be going into the experiment with the plausible (and personally salient) hypothesis that outcomes will follow their personal preferences. For Congruent-start participants, these hypotheses are proven correct in the first block, strengthening the “anchor” of Liked items to positive outcomes, and Disliked items to negative outcomes. Conversely, Incongruent-start participants receive initial evidence

against their hypotheses, and thus may be less biased in future judgements, consistent with the lack of block type effect in these participants.

More broadly, the order effect may reflect priming, where exposure to a stimulus influences subsequent stimulus responses[41–44]. In the Congruent-start condition, individuals experience associative priming, or an activation of their preexisting association of Liked items with positive outcomes and Disliked items with negative outcomes. The influence of this activation on learning may be furthered by its interaction with confirmation, confidence, and egocentric biases. People tend to favor advice that confirms their beliefs, regardless of accuracy[45–50]. A block that confirmed participants' beliefs may have also led to increased confidence in these beliefs, which also influences learning and decision making[49,51,52]. Moreover, decision makers disproportionately discount others' opinions in favor of their own, a phenomenon known as egocentric discounting[53]. The initial reinforcement of prior preferences of Congruent-start participants may have strengthened this tendency.

Initial reinforcement of prior preferences may hinder flexible learning

Congruent-start participants made more accurate predictions overall, compared to Incongruent-start participants. A simplistic interpretation of this result is that initial reinforcement of prior preferences is beneficial, consistent with prior work[54]. A more careful inspection, however, reveals that the higher accuracy of Congruent-start participants was driven by high performance specifically on Congruent blocks. In other words, Congruent-start participants stuck more strongly with their original preferences, an undesirable behavior if the goal is flexible learning and potential for behavioral change. The behavior of these participants may reflect the initial reinforcement of expressing prepotent (or impulsive) responses, which is generally easier to learn than the inhibition of such responses[55]. Conversely, although Incongruent-start participants exhibit lower overall accuracy, their similar accuracy across both block types–those that challenge their priors and those that align with their priors–demonstrates learning flexibility that is not evident among Congruent-start participants. This flexibility is also consistent with the Incongruent-start participants' reports of cumulative change in value following the task. Thus, our findings suggest that priming of one's priors potentiates learning inflexibility.

Differential effects of learning associations of Liked and Disliked items

The significant difference in prediction accuracy between Congruent and Incongruent blocks appears to be driven by Liked items. This is consistent with prior literature concerning the considerable influence of positive priors on learning and choice behavior. For example, attention has a stronger amplifying effect on value when choosing between already liked compared to neutral or disliked items[56]. The influence of positive priors persists across the social domain, where feedback of shared positive preferences with a partner proves beneficial to learning[54].

Interestingly, for Disliked items, participants performed better in the experimental condition they started in: The Congruent-start group performed better in Congruent blocks, while the Incongruent-start group

showed better performance in Incongruent blocks. This suggestibility of Disliked items to the task environment was further supported by changes in food item ratings after task performance among Incongruent-start participants, for whom subjective value ratings of willingness to pay, enjoyability, and expected satisfaction increased for Disliked food items post-task.

The relative rigidity we observed for associations of Liked food items may seem at odds with previous research that demonstrated value update for highly liked foods. In cued-approach training[15], participants watch highly valued food stimuli and are instructed to press a button upon hearing an auditory cue. Following the training, participants were more likely to choose items that were paired with a button press, consistent with an increase in the subjective value of paired items. Participants were also willing to pay more for paired, compared to unpaired, items following the training, although willingness to pay was lower on average for all highly valued items compared to their initial valuations. Subsequent studies used the cued-approach training to devalue the subjective appeal of energy-dense liked foods and shift subsequent choices toward healthier alternatives[57]. Our task, however, demands more from participants: Rather than slight implicit modulation of value within the reward domain, our task requires explicit reversals of inherent valence, from reward to punishment and vice versa.

Taken together, our results reveal an asymmetry in food value plasticity, suggesting that it may be easier to add a relatively disliked food item to one's diet[58,59] rather than remove a liked food item.

Carrots may be more effective than sticks

Regardless of start condition, subjects predicted rewards ('+5' outcome) better than punishments ('–5' outcome). Of note, these rewards and punishments had no real value to participants, whose explicit goal was to maximize prediction accuracy. Bonus payment was based on correct predictions on 75% or more of trials, regardless of the point outcome of any given trial. Despite their arbitrary nature, participants were affected differently by positive and negative outcomes. This is consistent with the considerable body of literature demonstrating positivity biases in learning, where individuals tend to learn more effectively from positive than negative outcomes[49,52,60,61]. To our knowledge, this is the first demonstration of this positivity bias as it relates to outcome assignments to personally salient food items. Considered in the context of food-based learning, it could be that emphasizing information about the rewarding aspects of food, like health benefits, may more effectively influence associative learning than information about potential harm, like negative health consequences.

Clinical implications

Exploratory analyses uncovered that higher levels of binge eating were associated with lower prediction accuracy when primed by one's priors and learning associations between Disliked food items and rewards. Existing literature demonstrates that reward learning is more complicated for those who binge eat. This study suggests that further investigation is warranted into the inability of individuals who binge eat to efficiently update the value of non-preferred stimuli when paired with a positive outcome. Moreover, we find that priming learning with one's priors may be of relevance. Indeed, cue reactivity

predicts eating behavior, and directing thoughts towards food before eating can induce overeating among binge eaters[62–64]. The effect we observed, however, is weak and reflects multiple comparisons, so more research is needed to explore the potential influence of prior preferences on reward learning in binge eating. This could be particularly relevant to obesity with binge eating, which is associated with increased risk of severe health consequences and lower efficacy rates in dietary interventions[18,19,65–68]: impaired subjective value updating may be a cognitive feature implicit in the maintenance of this comorbidity. As new behavioral and pharmaceutical intervention approaches rapidly emerge that demonstrate potential to alter food-related preferences[69–73], it is necessary to characterize the mechanism underlying flexible food value updating as it relates to metabolic and disordered eating pathology.

Limitations and future directions

While our study focused on food, we cannot speak to the specificity of our findings to food. Future research should explore the possible generalization of these findings to other items for which participants exhibit prior value associations (e.g., sports teams). Moreover, our study diverges from previous work in suggesting that one's prior preferences may hinder learning. It may be that context is particularly important: In cases like ours where participants passively learn and report on frequently changing associations, it may be that prior preferences are detrimental to learning, but in other contexts, such as those that are social or involving decision-making[54], it could be the opposite. Importantly, in terms of overall performance, we cannot conclude if Congruent-start participants perform better than baseline or if Incongruent-start participants are worse than baseline. Therefore, a valuable extension of this work could investigate whether, when food items are replaced with items the same participants view neutrally (e.g., shapes), participants achieve overall performance similar to the Congruent-start Congruent condition, similar to all other conditions, or in between. Another limitation of this work is the simplicity of food item selection, accomplished through a Likert scale. Future studies should extend this by integrating perceived health, nutritional content, taste, willingness to pay, and other continuous scales into personalized food stimuli selection. This may allow for greater reconciliation of findings with metabolic literature, such as work showing food preference may be changed through long-term adherence to low-fat and low-carbohydrate diets[74]. Finally, our study is lacking a measure of participant confidence in their predictions, which could be relevant to understanding the observed order effect.

Conclusion

Addressing unhealthy behaviors centers on the successful learning and implementation of healthier options. This makes associative learning a fundamental aspect of adapting to our changing environments, where reliance on innate knowledge is insufficient[75]. Such learning is also capable of producing complex and flexible behavior, be it in humans or artificial intelligence systems[76–78]. Ecologically, we rarely undergo associative learning in entirely naive environments; our prior preferences confound the learning process. This is especially true regarding our strong preferences for food, which

can undermine the associative learning policies central to dietary interventions. Whereas diet adherence is difficult for the general population[79–82], and interventions particularly lack efficacy for individuals who engage in disordered eating[18,65], it is essential to uncover the mechanisms that underlie associative learning with personally salient food items. Revealing these mechanisms can enhance the design of more successful dietary interventions, ultimately leading to better health outcomes.

Methods

Participants and procedures

In a preregistered study (https://osf.io/xjkt7/), participants (N = 415) were recruited online through Prolific in two cohorts. In fall 2024, 173 participants were recruited as part of Cohort A, of which 109 completed the study. Analyses performed in this cohort were preregistered before the remainder of the sample was recruited: In spring 2025, 304 participants were recruited as part of Cohort B, of which 189 completed the study. Across both cohorts, 298 participants completed the study.

First, participants rated 65 food images with no time constraints using a Likert scale. From these ratings, personalized repositories of four food images were generated for each participant, including one food item from each rating category: strongly dislike, dislike, like, and strongly like. Participants who did not have at least one item in each category were excluded from the remainder of the study. Eligible participants in both cohorts (n = 298) rated the subset of their four selected food items across three additional dimensions: willingness to pay, enjoyment, and expected satisfaction. Participants then engaged in a probabilistic reversal learning task involving these food items, during which they were asked to repeatedly predict stimulus-outcome pairs. Following the task, the four food items were rated again on a Likert scale, willingness to pay, enjoyment, and expected satisfaction. Finally, participants completed a series of surveys, including demographic questions and various psychological assessments.

Assessments

Sample characterization. Participants who completed the probabilistic reversal learning task (n = 298) completed several surveys to characterize the sample, including demographics (age, gender, race, ethnicity, height, weight) and psychological assessments (Supplemental Table 1). Surveys included in analyses were self-reported demographics and the Binge Eating Scale (BES)[21]. Participants also filled out additional surveys, including the Hunger Vital Sign[83], Generalized Anxiety Disorder 7 (GAD-7)[84], Patient Health Questionnaire 8 (PHQ-8)[85], Eating Disorder Examination Questionnaire (EDE-Q)[86], and Questionnaire on Eating and Weight Pattens-5 (QEWP-5)[87], for use in other research. Additionally, though not examined for this manuscript, participants were asked to report current or prior usage of GLP-1 agonist medications, such as semaglutide (e.g., Ozempic, Wegovy) or dulaglutide (e.g., Trulicity). Attention checks were embedded within these surveys to ensure participant engagement and data quality.

Food item rating. Food stimuli were 65 images from the Food Folio stimulus set by the Columbia Center for Eating Disorders (Fig. 1a)[20]. The food images selected were snack items ranging in caloric and nutritional value to encompass foods high, medium, and low in calories, fat, carbohydrates, and protein. To obtain each participant's inherent preferences for food stimuli, they were first asked to "rate snack images" (Fig. 1b). Specifically, participants were instructed, "Rate how much you like these snacks, considering them in the context of all food items. Answer as honestly as possible: using the entire range of the rating scale is encouraged." Participants (N = 415) indicated ratings using a sliding Likert scale described by the following points: strongly dislike, dislike, like, strongly like. Food stimuli were each presented once. These ratings were used to create personalized repositories consisting of four food items per participant (one strongly dislike, one dislike, one like, and one strongly like). Participants who did not have at least one item in each category were excluded from the remainder of the study. Before and after the probabilistic reversal learning task, eligible participants (n = 298) answered four specific questions for each of these four items, assessing their liking, willingness to pay, enjoyment, and expected satisfaction. These subset ratings provided additional insights into their preferences and expectations related to the selected food items.

Probabilistic reversal learning. Eligible participants (n = 298) completed a probabilistic reversal learning task in which they made predictions about positive or negative outcome assignments to their personally salient food items (Fig. 1c). Participants received the following instructions: "In this task, you will be shown foods related to positive ('+5') or negative ('–5') outcomes. For a certain number of trials, positive outcomes will often, but not always, be related to the same food items. This relationship will switch a number of times throughout the task. Pay close attention to these relationships: After we show you a food item, we will ask you to predict whether the outcome will be positive or negative. When this screen appears, use the 'left' and 'right' arrow keys to give your answers. You have a time limit of 3 seconds to make your prediction." Prior to the task, 6 practice trials were completed using two food items that participants had not seen previously. A comprehension check regarding the task structure and mechanics was performed, for which 253 participants (84.90%) scored 75% or above, before the correct responses were revealed for all participants to review.

During the task, the four non-neutral food images from each participant's personalized repository were presented. After a food item was presented, the participant was asked to predict the associated point outcome ('+5' or '–5') that would follow. Outcomes were determined by the block type in a probabilistic 90/10 reversal design. On Congruent blocks, 90% of the time, positively rated items (strongly like, like) were associated with a positive outcome ('+5'), and negatively rated items (strongly dislike, dislike) were associated with a negative outcome ('–5'). On Incongruent blocks, 90% of the time, this association was reversed: positively rated items were associated with a negative outcome, and negatively rated items were associated with a positive outcome. Specifically, the Congruent condition aligned the food item ratings with outcomes, allowing reliance on inherent values, whereas the Incongruent condition introduced conflict, requiring participants to make predictions against their inherent values for success. Participants received a bonus payment if their predictions were correct on 75% or more of trials.

The task consisted of six counterbalanced blocks: Three Congruent and three Incongruent blocks. Each block contained a jittered 32–36 trials, such that each food item was presented 8–9 times per block. The number of trials per block was determined from pilot versions of the task administered both online and in the lab. In the final pilot version, online participants (n = 20) had to achieve 75% accuracy within the last 10 trials of each block (after a minimum of 20 trials) before a reversal occurred: Across start groups and regardless of block type, an average of 31.30 ± 1.31 (mean ± s.d.) trials was required per block to meet this learning criterion (Supplemental Fig. 1). Participants completed 192–216 trials in total. The 90/10 reward probabilities ensured that within each block, 90% of the trials had outcomes following the block logic (e.g., for Congruent blocks, positive outcomes were received for liked items and negative outcomes for disliked items) and 10% had false outcomes (e.g., for Congruent blocks, negative outcomes were received for liked items and positive outcomes for disliked items). Probabilistic events occurred 3–4 times in each block, and these false outcome trials were spaced to ensure no consecutive false outcomes within 5 trials. The 90/10 probability distribution was determined from pilot versions of the task administered both online and in-lab (n_{online}=50, n_{in-lab}=14), in which 80/20 reward probabilities were insufficient to allow for learning above chance level. Attention checks were embedded within the task on a random trial in the third and sixth blocks to ensure participant adherence.

Statistical analysis

To characterize the relationship among subjective food values and caloric/nutritional content, we performed Spearman correlations between food item ratings and caloric, fat, and carbohydrate content. Nutritional content information was obtained from accompanying documentation of the Food Folio stimulus set by the Columbia Center for Eating Disorders.[20] These correlations were performed using rating data from the full recruited sample (N = 415), before exclusion of participants who did not have at least one food item in each non-neutral Likert category.

To assess prediction accuracy, we asked participants on each trial, “Which outcome do you expect?” (possible responses consisting of ‘+5’ and ‘–5’) after the food item was shown but before the actual outcome was revealed. Due to the probabilistic nature of this task, there are different ways to quantify accuracy: Outcome-Based Accuracy is how often participants' predictions match the actual outcomes received, and Policy Adherence Accuracy is how often participants' predictions align with the underlying probabilistic policy. For example, in a Congruent block, predicting a ‘+5’ outcome for a strongly liked or liked food item and a ‘–5’ outcome for a strongly disliked or disliked food item is considered correct, regardless of the outcome of the specific trial. Outcome-Based Accuracy findings provided insight into trial-by-trial predictive performance. However, understanding how well participants have learned the underlying rules of the task is critical to the principal study question concerning associative learning flexibility; therefore, Policy Adherence Accuracy was the primary outcome variable of interest and is synonymous with ‘accuracy’ as described in this study.

Participants who did not respond within the required 3-second timeframe on more than 10% of trials were excluded from analysis, resulting in the final participant pool of n = 279 ($n_{Congruent-start}$=145,

$n_{Incongruent-start}$=134). Analyses were performed in Cohort A (n = 106), before being preregistered and replicated in Cohort B (n = 173). For all analyses, we calculated average accuracy for each subject under the specified conditions (e.g., average Incongruent block accuracy) and then performed mean comparison or correlational analyses based on these averages.

A Shapiro-Wilk test was conducted to determine whether the prediction accuracy data were normally distributed. The results indicated a significant deviation from normality (W = 0.977, $p < 0.001$). Accordingly, non-parametric mean comparison tests (e.g., Wilcoxon signed-rank, Mann-Whitney U) were employed. Two-way and three-way ANOVAs were also conducted, as ANOVA is generally robust to violations of the normality assumption[88]. However, necessary precautions were taken to ensure the homogeneity of variances assumption was met using Levene's test ($ps > 0.05$).

We conducted trial- and block-wise analyses to quantify prediction accuracy over the course of the task and Wilcoxon signed-rank tests to determine performance relative to chance levels for each block. To determine if prediction accuracy differed as a function of block type and to identify order effects, we performed a two-way ANOVA with block type and initial block identity. Significant main and interaction effects were followed by post hoc testing using Tukey's Honest Significant Difference (HSD) test to identify specific group differences, with significance level set at $p = 0.05$. While Levene's test indicated no significant difference in variances across the groups, $F(3, 554) = 2.24$, $p = 0.08$, pairwise Fligner-Killeen tests revealed that the variances of Congruent-start Incongruent block accuracy and Incongruent-start Incongruent block accuracy are significantly different. Accordingly, Games-Howell post hoc testing was performed to identify any deviations from the reported Tukey's HSD findings[89]. Finally, a Mann-Whitney U test was conducted to determine if overall accuracy differed between Congruent-start and Incongruent-start participants.

We additionally conducted Mann-Whitney U Tests to determine if prediction accuracy for reward ('+5' outcome)-associated foods differs from punishment ('–5' outcome)-associated foods, which are defined as a function of block type; for example, liked and strongly liked foods are primarily associated with reward on Congruent blocks but are primarily associated with punishment on Incongruent blocks.

We performed a three-way ANOVA to investigate the main effects of start block, block type, and food item rating on accuracy, as well as the significance of their interactions. Significant main and interaction effects were followed by post hoc testing using Tukey's HSD test to identify specific group differences, with significance level set at $p = 0.05$.

To model within-block learning dynamics, a simple sigmoid function was fitted to the prediction accuracy data for each condition (defined by unique initial block identity, block type, and food item rating combinations). Before fitting, prediction accuracy data were averaged across subjects within each experimental condition. We then fit sigmoid curves to these group-averaged means to estimate the parameters L, k, and x0:

$$y = L / (1 + e^{(-k*(x-x0)}$$

- L represents the upper asymptote or threshold
- k is the growth rate, or slope
- x0 is the trial number at which prediction accuracy reaches half of L

A two-sample t-test was performed to compare the learning slopes (k) between the start groups. Additionally, an ANOVA was conducted to compare the learning thresholds (L) across all conditions. Significant interactions of block type with start group and image item rating were identified and further analyzed using post hoc Tukey's HSD tests.

Paired t-tests were conducted to assess the significance of any changes in ratings (willingness to pay, enjoyment, and expected satisfaction) from pre-task to post-task within each start group. Of note, 5 participants did not complete post-task ratings and were thus excluded from this analysis ($n_{Congruent}$-start = 143, $n_{Incongruent}$-start = 131).

Finally, as an exploratory aim of the preregistered study, we predicted that binge eating severity and BMI may be related to deficits in updating inherent priors. To assess if certain demographic variables relate to learning, we performed exploratory correlational analyses probing the relations of Binge-Eating Scale (BES) Global Scores and BMI with task performance. Correlations with average accuracy across each condition (Congruent/Incongruent block type and Liked/Disliked food item rating combination) were tested within each start group. Before performing Pearson correlations, we executed a rank-based inverse normal transformation of the data to minimize Type I and Type II error[22]. For the significant result identified—that prediction accuracy for Disliked items in Incongruent blocks was negatively correlated with global BES scores among Congruent-start participants—we further analyzed this relationship using a linear regression model. In this model, BES Global Scores served as the independent variable, while average accuracy under the specified conditions acted as the dependent variable. We also controlled for potential confounding variables, including age, gender, and BMI, in our regression analysis. Of note, there were 142 Congruent-start participants with survey data that were included in the regression analysis.

Analyses were conducted using the SciPy and statsmodels libraries and the pingouin package in Python[90,91].

Declarations

Participants provided consent after a detailed explanation of the study was provided, following the institute guidelines approved by the Yale Human Investigation Committee.

Code availability

Analysis codes are available at https://github.com/LevyDecisionNeuroLab/FoodValuation_BehavioralAnalyses. Additional codes are available from the corresponding author on reasonable request.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgements

We thank Liz Goldfarb, Janet Lydecker, Dustin Scheinost, Nachshon Korem, and Sierra Metviner for their very helpful discussions and comments. We also thank Madhav Lavakare and Lukas Nel for their exceptional programming support. This study was funded by NIH grant R01MH133886 to I.L.

Author Information

Authors and Affiliations

Interdepartmental Neuroscience Program, Yale University, New Haven, CT, 06520, USA

Alexandra Rich, Ryan Henry & Ifat Levy

Department of Comparative Medicine, Yale School of Medicine, New Haven, CT, 06520, USA

Alexandra Rich, Sohum Kapadia, Ryan Henry, Ohad J. Dan & Ifat Levy

Yale College, Yale University, New Haven, CT, 06520, USA

Sohum Kapadia

Wu-Tsai Institute, Department of Neuroscience, and Department of Psychology, Yale University, New Haven, CT, 06510, USA

Ifat Levy

Contributions

A.R., R.H., and I.L. designed study, A.R. conducted study, A.R., R.H., S.K., O.D. analyzed data, A.R. and I.L. wrote manuscript, all authors commented and approved the final manuscript.

Corresponding authors

Correspondence to Ifat Levy.

Ethics declarations

Competing interests

The authors declare no competing interests.

References

1. Hare, T. A., Camerer, C. F. & Rangel, A. Self-Control in Decision-Making Involves Modulation of the vmPFC Valuation System. *Science* 324, 646–648 (2009).
2. Bjorlie, K. *et al.* Hyper-palatable food consumption during binge-eating episodes: A comparison of intake during binge eating and restricting. https://doi.org/10.1002/eat.23692 doi:10.1002/eat.23692.
3. Capaldi, E. D. Conditioned Food Preferences. in *Psychology of Learning and Motivation* (ed. Medin, D. L.) vol. 28 1–33 (Academic Press, 1992).
4. Berthoud, H.-R., Morrison, C. D., Ackroff, K. & Sclafani, A. Learning of food preferences: mechanisms and implications for obesity & metabolic diseases. *Int. J. Obes.* 45, 2156–2168 (2021).
5. Robinson, M. J. F., Fischer, A. M., Ahuja, A., Lesser, E. N. & Maniates, H. Roles of "Wanting" and "Liking" in Motivating Behavior: Gambling, Food, and Drug Addictions. in *Behavioral Neuroscience of Motivation* (eds. Simpson, E. H. & Balsam, P. D.) 105–136 (Springer International Publishing, Cham, 2016). doi:10.1007/7854_2015_387.
6. O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J. & Andrews, C. Abstract reward and punishment representations in the human orbitofrontal cortex. *Nat. Neurosci.* 4, 95–102 (2001).
7. O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H. & Dolan, R. J. Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron* 38, 329–337 (2003).
8. Reiter, A. M. F., Heinze, H.-J., Schlagenhauf, F. & Deserno, L. Impaired Flexible Reward-Based Decision-Making in Binge Eating Disorder: Evidence from Computational Modeling and Functional Neuroimaging. *Neuropsychopharmacol. Off. Publ. Am. Coll. Neuropsychopharmacol.* 42, 628–637 (2017).
9. Balodis, I. M. *et al.* Monetary Reward Processing in Obese Individuals With and Without Binge Eating Disorder. *Biol. Psychiatry* 73, 877–886 (2013).
10. Svaldi, J., Brand, M. & Tuschen-Caffier, B. Decision-making impairments in women with binge eating disorder. *Appetite* 54, 84–92 (2010).
11. Davis, C., Patte, K., Curtis, C. & Reid, C. Immediate pleasures and future consequences. A neuropsychological study of binge eating and obesity. *Appetite* 54, 208–213 (2010).
12. Miranda-Olivos, R. *et al.* The neural correlates of delay discounting in obesity and binge eating disorder. *J. Behav. Addict.* 10, 498–507 (2021).
13. Zhang, Z., Manson, K. F., Schiller, D. & Levy, I. Impaired Associative Learning with Food Rewards in Obese Women. *Curr. Biol.* 24, 1731–1736 (2014).

14. Zhang (□□□), Z., Mendelsohn, A., Manson, K. F., Schiller, D. & Levy, I. Dissociating Value Representation and Inhibition of Inappropriate Affective Response during Reversal Learning in the Ventromedial Prefrontal Cortex. *eNeuro* 2, ENEURO.0072-15.2015 (2016).
15. Schonberg, T. *et al.* Changing value through cued approach: an automatic mechanism of behavior change. *Nat. Neurosci.* 17, 625–630 (2014).
16. Voon, V. Cognitive biases in binge eating disorder: the hijacking of decision making. *CNS Spectr.* 20, 566–573 (2015).
17. Waltmann, M., Herzog, N., Horstmann, A. & Deserno, L. Loss of control over eating: A systematic review of task based research into impulsive and compulsive processes in binge eating. *Neurosci. Biobehav. Rev.* 129, 330–350 (2021).
18. Da Luz, F. Q., Hay, P., Touyz, S. & Sainsbury, A. Obesity with Comorbid Eating Disorders: Associated Health Risks and Treatment Approaches. *Nutrients* 10, 829 (2018).
19. Cooper, Z., Calugi, S. & Dalle Grave, R. Controlling binge eating and weight: a treatment for binge eating disorder worth researching? *Eat. Weight Disord. - Stud. Anorex. Bulim. Obes.* 25, 1105–1109 (2020).
20. Lloyd, E. C. *et al.* Food Folio by Columbia Center for Eating Disorders: A Freely Available Food Image Database. *Front. Psychol.* 11, (2020).
21. Gormally, J., Black, S., Daston, S. & Rardin, D. The assessment of binge eating severity among obese persons. *Addict. Behav.* 7, 47–55 (1982).
22. Bishara, A. J. & Hittner, J. B. Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychol. Methods* 17, 399–417 (2012).
23. Garvert, M. M., Moutoussis, M., Kurth-Nelson, Z., Behrens, T. E. J. & Dolan, R. J. Learning-Induced Plasticity in Medial Prefrontal Cortex Predicts Preference Malleability. *Neuron* 85, 418–428 (2015).
24. Martino, B. D., Bobadilla-Suarez, S., Nouguchi, T., Sharot, T. & Love, B. C. Social Information Is Integrated into Value and Confidence Judgments According to Its Reliability. *J. Neurosci.* 37, 6066–6074 (2017).
25. Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J. & Frith, C. D. How the Opinion of Others Affects Our Valuation of Objects. *Curr. Biol.* 20, 1165–1170 (2010).
26. Izuma, K. & Adolphs, R. Social Manipulation of Preference in the Human Brain. *Neuron* 78, 563–573 (2013).
27. Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A. & Fernández, G. Reinforcement Learning Signal Predicts Social Conformity. *Neuron* 61, 140–151 (2009).
28. Suzuki, S. *et al.* Learning to Simulate Others' Decisions. *Neuron* 74, 1125–1137 (2012).
29. Zaki, J., Schirmer, J. & Mitchell, J. P. Social Influence Modulates the Neural Computation of Value. *Psychol. Sci.* 22, 894–900 (2011).

30. Moutoussis, M., Dolan, R. J. & Dayan, P. How People Use Social Information to Find out What to Want in the Paradigmatic Case of Inter-temporal Preferences. *PLOS Comput. Biol.* 12, e1004965 (2016).
31. Dietrich, F. & List, C. Where do preferences come from? *Int. J. Game Theory* 42, 613–637 (2013).
32. Warren, C., McGraw, A. P. & Van Boven, L. Values and preferences: defining preference construction. *WIREs Cogn. Sci.* 2, 193–205 (2011).
33. Bernheim, B. D., Braghieri, L., Martínez-Marquina, A. & Zuckerman, D. A Theory of Chosen Preferences. *Am. Econ. Rev.* 111, 720–754 (2021).
34. Falk, A. *et al.* Global Evidence on Economic Preferences*. *Q. J. Econ.* 133, 1645–1692 (2018).
35. Amos Tversky & Daniel Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 1124–1131 (1974).
36. Mussweiler, T. & Strack, F. Hypothesis-Consistent Testing and Semantic Priming in the Anchoring Paradigm: A Selective Accessibility Model. *J. Exp. Soc. Psychol.* 35, 136–164 (1999).
37. Chapman, G. B. & Johnson, E. J. Anchoring, Activation, and the Construction of Values. *Organ. Behav. Hum. Decis. Process.* 79, 115–153 (1999).
38. Mussweiler, T. & Strack, F. The Semantics of Anchoring. *Organ. Behav. Hum. Decis. Process.* 86, 234–255 (2001).
39. Strack, F. & Mussweiler, T. Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *J. Pers. Soc. Psychol.* 73, 437–446 (1997).
40. Furnham, A. & Boo, H. C. A literature review of the anchoring effect. *J. Socio-Econ.* 40, 35–42 (2011).
41. Wagner, A. R. Expectancies and the Priming of STM. in *Cognitive Processes in Animal Behavior* (Routledge, 1978).
42. Tulving, E. & Schacter, D. L. Priming and Human Memory Systems. *Science* 247, 301–306 (1990).
43. Wasserman, E. A. & Miller, R. R. WHAT'S ELEMENTARY ABOUT ASSOCIATIVE LEARNING? *Annu. Rev. Psychol.* 48, 573–607 (1997).
44. Schacter, D. L. & Buckner, R. L. Priming and the Brain. *Neuron* 20, 185–195 (1998).
45. Suen, W. The Self-Perpetuation of Biased Beliefs. *Econ. J.* 114, 377–396 (2004).
46. Marks, J., Copland, E., Loh, E., Sunstein, C. R. & Sharot, T. Epistemic spillovers: Learning others' political views reduces the ability to assess and use their expertise in nonpolitical domains. *Cognition* 188, 74–84 (2019).
47. Charness, G., Oprea, R. & Yuksel, S. How do People Choose Between Biased Information Sources? Evidence from a Laboratory Experiment. *J. Eur. Econ. Assoc.* 19, 1656–1691 (2021).
48. Hart, W. *et al.* Feeling Validated Versus Being Correct. *Psychol. Bull.* 135, 555–588 (2009).
49. Palminteri, S., Lefebvre, G., Kilford, E. J. & Blakemore, S.-J. Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLOS Comput. Biol.* 13, e1005684 (2017).

50. Nickerson, R. S. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Rev. Gen. Psychol.* 2, 175–220 (1998).
51. Dunning, D. Chapter five - The Dunning–Kruger Effect: On Being Ignorant of One's Own Ignorance. in *Advances in Experimental Social Psychology* (eds. Olson, J. M. & Zanna, M. P.) vol. 44 247–296 (Academic Press, 2011).
52. Palminteri, S. & Lebreton, M. The computational roots of positivity and confirmation biases in reinforcement learning. *Trends Cogn. Sci.* 26, 607–621 (2022).
53. Yaniv, I. & Kleinberger, E. Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation. *Organ. Behav. Hum. Decis. Process.* 83, 260–281 (2000).
54. Tarantola, T., Kumaran, D., Dayan, P. & De Martino, B. Prior preferences beneficially influence social and non-social learning. *Nat. Commun.* 8, 817 (2017).
55. Guerrieri, R., Nederkoorn, C. & Jansen, A. Disinhibition is easier learned than inhibition. The effects of (dis)inhibition training on food intake. *Appetite* 59, 96–99 (2012).
56. Smith, S. M. & Krajbich, I. Gaze Amplifies Value in Decision Making. *Psychol. Sci.* 30, 116–128 (2019).
57. Porter, L., Gillison, F. B., Wright, K. A., Verbruggen, F. & Lawrence, N. S. Exploring Strategies to Optimise the Impact of Food-Specific Inhibition Training on Children's Food Choices. *Front. Psychol.* 12, (2021).
58. Koegel, R. L. *et al.* Using Individualized Reinforcers and Hierarchical Exposure to Increase Food Flexibility in Children with Autism Spectrum Disorders. *J. Autism Dev. Disord.* 42, 1574–1581 (2012).
59. Rogers, P. J. & Hill, A. J. Breakdown of dietary restraint following mere exposure to food stimuli: Interrelationships between restraint, hunger, salivation, and food intake. *Addict. Behav.* 14, 387–397 (1989).
60. Bruckner, R., Heekeren, H. R. & Nassar, M. R. Understanding learning through uncertainty and bias. *Commun. Psychol.* 3, 24 (2025).
61. Vidal-Perez, J., Dolan, R. J. & Moran, R. Disinformation elicits learning biases. *eLife* 14, (2025).
62. Boswell, R. G. & Kober, H. Food cue reactivity and craving predict eating and weight gain: a meta-analytic review. *Obes. Rev. Off. J. Int. Assoc. Study Obes.* 17, 159–177 (2016).
63. Jansen, A. A learning model of binge eating: Cue reactivity and cue exposure. *Behav. Res. Ther.* 36, 257–272 (1998).
64. Fedoroff, I. D. C., Polivy, J. & Herman, C. P. The Effect of Pre-exposure to Food Cues on the Eating Behavior of Restrained and Unrestrained Eaters. *Appetite* 28, 33–47 (1997).
65. Poli, R., Maninetti, L., Bodini, P. & Agrimi, E. Obesity, binge eating, obstruction sleep apnea and psychopathological features. *Clin. Neuropsychiatry* 9, 166–170 (2012).
66. Niego, S. H., Kofman, M. D., Weiss, J. J. & Geliebter, A. Binge eating in the bariatric surgery population: a review of the literature. *Int. J. Eat. Disord.* 40, 349–359 (2007).

67. Wonderlich, S. A., de Zwaan, M., Mitchell, J. E., Peterson, C. & Crow, S. Psychological and dietary treatments of binge eating disorder: Conceptual implications. *Int. J. Eat. Disord.* 34, S58–S73 (2003).
68. Marcus, M. D., Wing, R. R. & Lamparski, D. M. Binge eating and dietary restraint in obese patients. *Addict. Behav.* 10, 163–168 (1985).
69. Itzkovitch, A., Bar Or, M. & Schonberg, T. Cue-approach training for food behavior. *Curr. Opin. Behav. Sci.* 47, 101202 (2022).
70. Stice, E. *et al.* Efficacy of a food response and attention training treatment for obesity: A randomized placebo controlled trial. *Behav. Res. Ther.* 158, 104183 (2022).
71. Yang, Y., Qi, L., Morys, F., Wu, Q. & Chen, H. Food-Specific Inhibition Training for Food Devaluation: A Meta-Analysis. *Nutrients* 14, 1363 (2022).
72. Aldawsari, M. *et al.* The Efficacy of GLP-1 Analogues on Appetite Parameters, Gastric Emptying, Food Preference and Taste Among Adults with Obesity: Systematic Review of Randomized Controlled Trials. *Diabetes Metab. Syndr. Obes. Targets Ther.* 16, 575–595 (2023).
73. Bettadapura, S., Dowling, K., Jablon, K., Al-Humadi, A. W. & le Roux, C. W. Changes in food preferences and ingestive behaviors after glucagon-like peptide-1 analog treatment: techniques and opportunities. *Int. J. Obes.* 49, 418–426 (2025).
74. Martin, C. K. *et al.* Change in Food Cravings, Food Preferences, and Appetite During a Low-Carbohydrate and Low-Fat Diet. *Obesity* 19, 1963–1970 (2011).
75. Lind, J., Ghirlanda, S. & Enquist, M. Social learning through associative processes: a computational theory. *R. Soc. Open Sci.* 6, 181777 (2019).
76. Enquist, M., Lind, J. & Ghirlanda, S. The power of associative learning and the ontogeny of optimal behaviour. *R. Soc. Open Sci.* 3, 160734 (2016).
77. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* 518, 529–533 (2015).
78. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489 (2016).
79. Leme, A. C. B., Hou, S., Fisberg, R. M., Fisberg, M. & Haines, J. Adherence to Food-Based Dietary Guidelines: A Systemic Review of High-Income and Low- and Middle-Income Countries. *Nutrients* 13, 1038 (2021).
80. Moreira, E. A. M., Most, M., Howard, J. & Ravussin, E. Dietary Adherence to Long-Term Controlled Feeding in a Calorie-Restriction Study in Overweight Men and Women. *Nutr. Clin. Pract.* 26, 309–315 (2011).
81. Domosławska-Żylińska, K., Łopatek, M., Krysińska-Pisarek, M. & Sugay, L. Barriers to Adherence to Healthy Diet and Recommended Physical Activity Perceived by the Polish Population. *J. Clin. Med.* 13, 22 (2024).
82. Boujelbane, M. A. *et al.* Regional variations in Mediterranean diet adherence: a sociodemographic and lifestyle analysis across Mediterranean and non-Mediterranean regions within the MEDIET4ALL

project. *Front. Public Health* 13, (2025).

83. Gattu, R. K. *et al.* The Hunger Vital Sign Identifies Household Food Insecurity among Children in Emergency Departments and Primary Care. *Children* 6, 107 (2019).
84. Spitzer, R. L., Kroenke, K., Williams, J. B. W. & Löwe, B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch. Intern. Med.* 166, 1092–1097 (2006).
85. Kroenke, K. *et al.* The PHQ-8 as a measure of current depression in the general population. *J. Affect. Disord.* 114, 163–173 (2009).
86. Luce, K. H. & Crowther, J. H. The reliability of the Eating Disorder Examination-Self-Report Questionnaire Version (EDE-Q). *Int. J. Eat. Disord.* 25, 349–351 (1999).
87. Yanovski, S. Z., Marcus, M. D., Wadden, T. A. & Walsh, B. T. The Questionnaire on Eating and Weight Patterns-5 (QEWP-5): An Updated Screening Instrument for Binge Eating Disorder. *Int. J. Eat. Disord.* 48, 259–261 (2015).
88. Blanca Mena, M. J., Alarcón Postigo, R., Arnau Gras, J., Bono Cabré, R. & Bendayan, R. Non-normal data: Is ANOVA still a valid option? *Artic. Publ. En Rev. Psicol. Soc. Psicol. Quant.* https://diposit.ub.edu/dspace/handle/2445/122126 (2017).
89. Lee, S. & Lee, D. K. What is the proper way to apply the multiple comparison test? *Korean J. Anesthesiol.* 71, 353–360 (2018).
90. SciPy 1.0: fundamental algorithms for scientific computing in Python | Nature Methods. https://www.nature.com/articles/s41592-019-0686-2.
91. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. *SciPy* 2010 https://doi.org/10.25080/Majora-92bf1922-011 (2010) doi:10.25080/Majora-92bf1922-011.

Figures

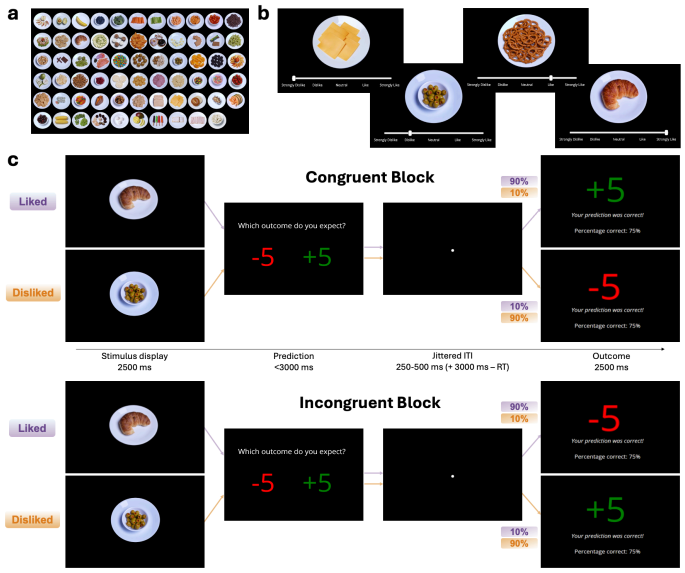


Figure 1

Experimental paradigm. a Food stimuli were 65 snack images ranging in caloric, fat, carbohydrate, and protein content. **b** To obtain each participant's inherent preferences, participants rated food stimuli using a Likert scale. For each participant, 4 non-neutral food items were randomly selected. **c** Participants were asked to predict associations between personally salient food items and point outcomes in a probabilistic reversal learning task. In Congruent blocks, Liked food items, including both strongly liked and liked items, were most often (on 90% of trials) associated with positive outcomes and Disliked food items, including both strongly disliked and disliked items, were most often associated with negative outcomes. In Incongruent blocks, Liked food items were most often associated with negative outcomes and Disliked food items were most often associated with positive outcomes. On 10% of trials in a given block, a probabilistic event occurred such that associations between food items and outcomes were presented in a manner contrary to the aforementioned rules of the current block. On the Outcome screen, participants could view the actual outcome ('+5' or '-5' arbitrary point value), feedback on their prediction accuracy for that trial, and feedback on their total prediction accuracy. The task consisted of 6 counterbalanced blocks with jittered 32-36 trials each. The timeline for each trial is shown.

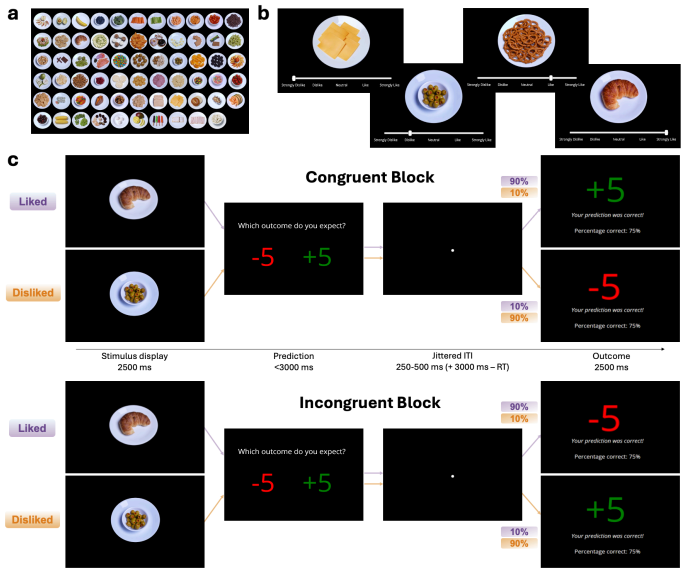


Figure 1

Experimental paradigm. a Food stimuli were 65 snack images ranging in caloric, fat, carbohydrate, and protein content. **b** To obtain each participant's inherent preferences, participants rated food stimuli using a Likert scale. For each participant, 4 non-neutral food items were randomly selected. **c** Participants were asked to predict associations between personally salient food items and point outcomes in a probabilistic reversal learning task. In Congruent blocks, Liked food items, including both strongly liked and liked items, were most often (on 90% of trials) associated with positive outcomes and Disliked food items, including both strongly disliked and disliked items, were most often associated with negative outcomes. In Incongruent blocks, Liked food items were most often associated with negative outcomes and Disliked food items were most often associated with positive outcomes. On 10% of trials in a given block, a probabilistic event occurred such that associations between food items and outcomes were presented in a manner contrary to the aforementioned rules of the current block. On the Outcome screen, participants could view the actual outcome ('+5' or '-5' arbitrary point value), feedback on their prediction accuracy for that trial, and feedback on their total prediction accuracy. The task consisted of 6 counterbalanced blocks with jittered 32-36 trials each. The timeline for each trial is shown.

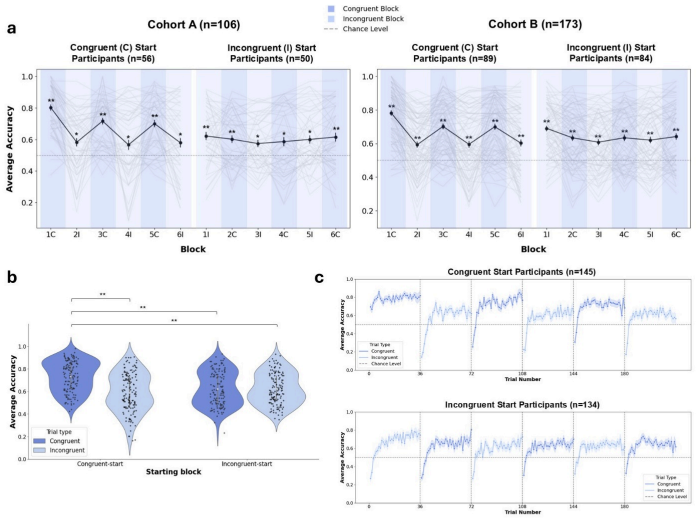


Figure 2

Average accuracy by start group and block type, over time. a Data were collected at two separate time points, indicated here as Cohort A (n=106) and Cohort B (n=173). Accuracy data are shown block-wise, as a function of block type (Congruent and Incongruent) and the condition participants started in (Congruent-start and Incongruent-start). Values are means ± s.e.m. A Wilcoxon signed-rank test was used to determine if the accuracy in each block was significantly above chance level. Block performance significantly greater than chance level is denoted by ** ($p<0.001$) and * ($p<0.05$). The grey dashed line at y=0.5 denotes the chance level prediction accuracy. Grey solid lines represent average accuracy for individual participants. Error bars represent the standard error of the mean (SEM). **b** Average accuracy as a function of both block type and start group ($n_{Congruent\text{-}start}=145$, $n_{Incongruent\text{-}start}=134$). Light grey values inside each violin are means ± interquartile range (25th to 75th percentile), as indicated by the surrounding grey bars. The p-values indicated are derived from post hoc analyses using Tukey's Honest Significant Difference (HSD) test following a two-way ANOVA. Significant differences between groups are denoted by ** ($p<0.001$). **c** Average accuracy is depicted trial-wise within each block. Data is categorized by block type and starting group. Each block contained jittered 32-36 trials. Values are means ± s.e.m., as indicated by the surrounding error cloud. The grey dashed line at y=0.5 denotes the chance level prediction accuracy.

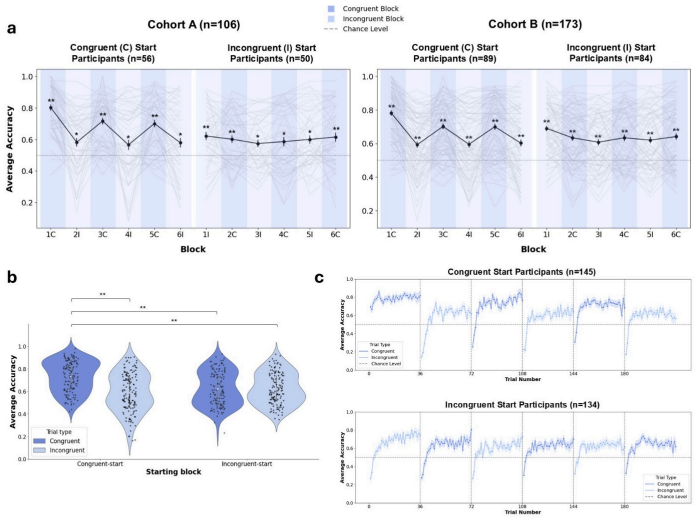


Figure 2

Average accuracy by start group and block type, over time. a Data were collected at two separate time points, indicated here as Cohort A (n=106) and Cohort B (n=173). Accuracy data are shown block-wise, as a function of block type (Congruent and Incongruent) and the condition participants started in (Congruent-start and Incongruent-start). Values are means ± s.e.m. A Wilcoxon signed-rank test was used to determine if the accuracy in each block was significantly above chance level. Block performance significantly greater than chance level is denoted by ** ($p<0.001$) and * ($p<0.05$). The grey dashed line at y=0.5 denotes the chance level prediction accuracy. Grey solid lines represent average accuracy for individual participants. Error bars represent the standard error of the mean (SEM). **b** Average accuracy as a function of both block type and start group ($n_{Congruent\text{-}start}=145$, $n_{Incongruent\text{-}start}=134$). Light grey values inside each violin are means ± interquartile range (25th to 75th percentile), as indicated by the surrounding grey bars. The p-values indicated are derived from post hoc analyses using Tukey's Honest Significant Difference (HSD) test following a two-way ANOVA. Significant differences between groups are denoted by ** ($p<0.001$). **c** Average accuracy is depicted trial-wise within each block. Data is categorized by block type and starting group. Each block contained jittered 32-36 trials. Values are means ± s.e.m., as indicated by the surrounding error cloud. The grey dashed line at y=0.5 denotes the chance level prediction accuracy.

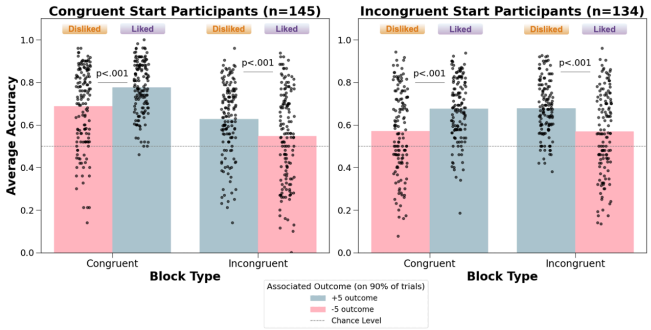


Figure 3

Average accuracy is higher for reward associations, regardless of block type, food item rating, or start condition. Accuracy data are categorized by block type (Congruent and Incongruent) and food item rating (Liked and Disliked). The Congruent-start group (starting in the Congruent block; n=145) is shown in the left panel, whereas the Incongruent-start group (starting in the Incongruent block; n=134) is shown in the right panel. The associated outcome (on 90% of trials) for items of each image rating and block type combination is indicated by bar colors. P-values displayed above the bars indicate the significance of the difference between Liked and Disliked items within each block, as determined by the Mann-Whitney U test. The grey dashed line at y=0.5 denotes the chance level prediction accuracy.

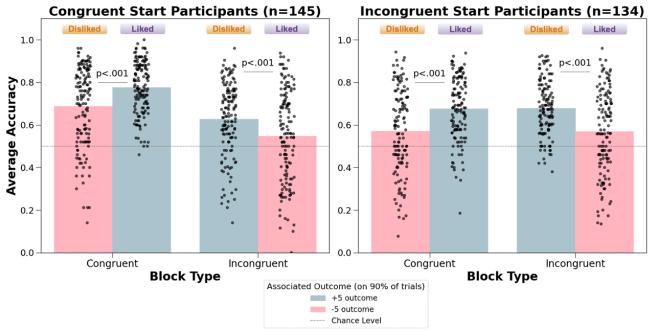


Figure 3

Average accuracy is higher for reward associations, regardless of block type, food item rating, or start condition. Accuracy data are categorized by block type (Congruent and Incongruent) and food item rating (Liked and Disliked). The Congruent-start group (starting in the Congruent block; n=145) is shown in the left panel, whereas the Incongruent-start group (starting in the Incongruent block; n=134) is shown in the right panel. The associated outcome (on 90% of trials) for items of each image rating and block type combination is indicated by bar colors. P-values displayed above the bars indicate the significance of the difference between Liked and Disliked items within each block, as determined by the Mann-Whitney U test. The grey dashed line at y=0.5 denotes the chance level prediction accuracy.

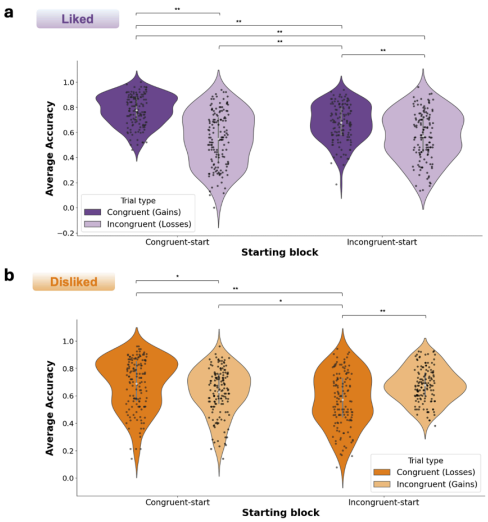


Figure 4

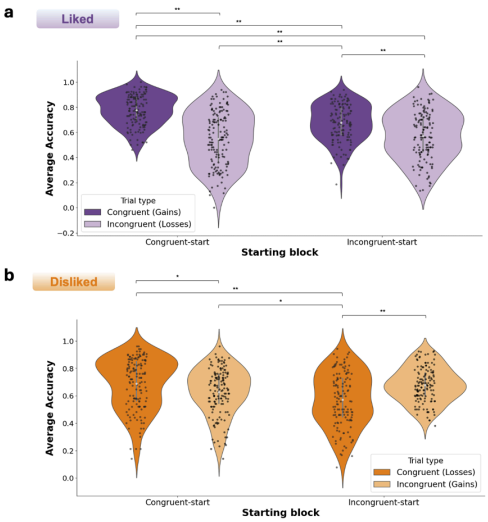


Figure 4

Average accuracy by food item rating, start group, and block type. Average prediction accuracy is shown by food item rating, for both (**a**) Liked items (liked and strongly liked) and (**b**) Disliked items (disliked and strongly disliked). Within each panel, data are categorized by block type (Congruent and Incongruent) for participants in the Congruent-start (starting in the Congruent block; n=145) and Incongruent-start (starting in the Incongruent block; n=134) groups. Light grey values inside each violin are means ± interquartile range (25th to 75th percentile), as indicated by the surrounding grey bars. P-values are derived from post hoc analyses using Tukey's Honest Significant Difference (HSD) test following a three-way ANOVA. Significant differences between groups are denoted by ** ($p<0.001$) and * ($p<0.05$).

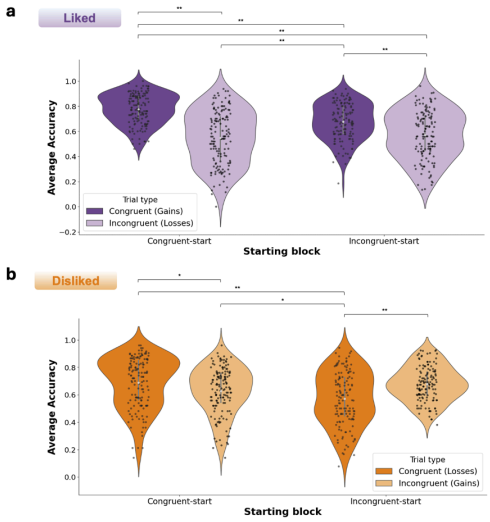


Figure 4

Average accuracy by food item rating, start group, and block type. Average prediction accuracy is shown by food item rating, for both (**a**) Liked items (liked and strongly liked) and (**b**) Disliked items (disliked and strongly disliked). Within each panel, data are categorized by block type (Congruent and Incongruent) for participants in the Congruent-start (starting in the Congruent block; n=145) and Incongruent-start (starting in the Incongruent block; n=134) groups. Light grey values inside each violin are means ± interquartile range (25th to 75th percentile), as indicated by the surrounding grey bars. P-values are derived from post hoc analyses using Tukey's Honest Significant Difference (HSD) test following a three-way ANOVA. Significant differences between groups are denoted by ** ($p<0.001$) and * ($p<0.05$).

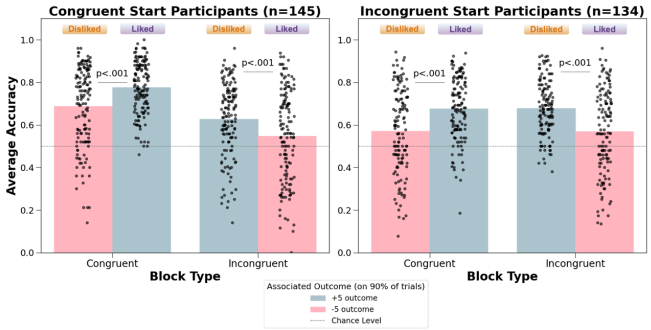


Figure 5

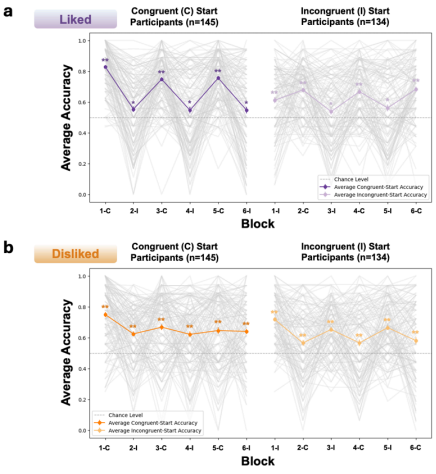


Figure 5

Average accuracy by food item rating, start group, and block type, over time. Average prediction accuracy is shown block-wise by food item rating, for both (**a**) Liked items (liked and strongly liked) and (**b**) Disliked items (disliked and strongly disliked). Accuracy data are additionally categorized by block type (Congruent and Incongruent) and the condition participants started in (Congruent-start and Incongruent-start; $n_{Congruent\text{-}start}$=145, $n_{Incongruent\text{-}start}$=134). Within each block, each food item (strongly liked, liked, disliked, and strongly disliked) was shown on approximately 8-9 trials. Values are means ± s.e.m. A Wilcoxon signed-rank test was used to determine if the accuracy in each block was significantly above chance level. Block performance significantly greater than chance level is denoted by ** ($p<0.001$) and * ($p<0.05$). The grey dashed line at y=0.5 denotes the chance level prediction accuracy. Grey solid lines represent average accuracy for individual participants. Error bars represent the standard error of the mean (SEM).

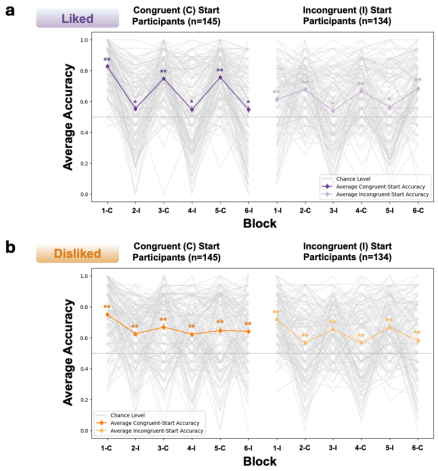


Figure 5

Average accuracy by food item rating, start group, and block type, over time. Average prediction accuracy is shown block-wise by food item rating, for both (**a**) Liked items (liked and strongly liked) and (**b**) Disliked items (disliked and strongly disliked). Accuracy data are additionally categorized by block type (Congruent and Incongruent) and the condition participants started in (Congruent-start and Incongruent-start; $n_{Congruent\text{-}start}$=145, $n_{Incongruent\text{-}start}$=134). Within each block, each food item (strongly liked, liked, disliked, and strongly disliked) was shown on approximately 8-9 trials. Values are means ± s.e.m. A Wilcoxon signed-rank test was used to determine if the accuracy in each block was significantly above chance level. Block performance significantly greater than chance level is denoted by ** ($p<0.001$)

and * ($p<0.05$). The grey dashed line at y=0.5 denotes the chance level prediction accuracy. Grey solid lines represent average accuracy for individual participants. Error bars represent the standard error of the mean (SEM).

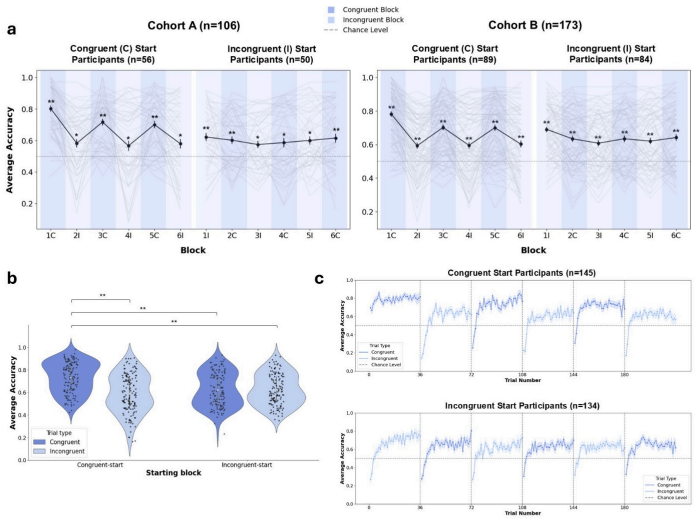


Figure 6

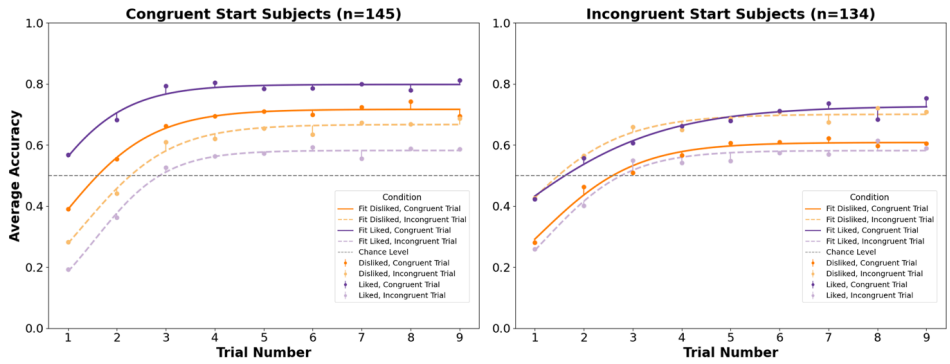


Figure 6

Sigmoid function fit to within-block average accuracy by food item rating, start group, and block type. The fit of a sigmoid function to the average accuracy data within blocks across task conditions, defined by the condition participants started in (Congruent-start and Incongruent-start; $n_{Congruent\text{-}start}$=145, $n_{Incongruent\text{-}start}$=134), food item rating (Liked and Disliked), and block type (Congruent and Incongruent). The plot on the left represents subjects who began with a Congruent block, while those on the right represent subjects who began with an Incongruent block. Each data point corresponds to the mean accuracy for trials, with error bars indicating asymmetrical residual errors from the sigmoid fit, calculated as the difference between fitted values and actual means. The fitted sigmoid curves model the average accuracy progression over the trials, capturing learning or performance trends. Within each block, each item (Strongly Liked, Liked, Disliked, and Strongly Disliked) was shown on approximately 8-9 trials; accordingly, Trial Number represents a simulated trial number to accommodate the aggregation of these data across blocks and food items of the same rating subgroup (Liked, Disliked). The grey dashed line at y=0.5 denotes the chance level prediction accuracy.

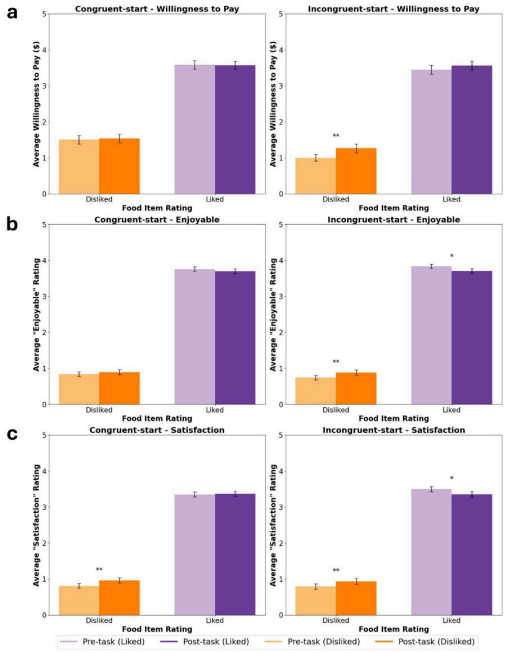


Figure 7

Changes in willingness to pay, enjoyment, and satisfaction for food items by start group and item rating. (**a**) Average willingness to pay, (**b**) enjoyment, and (**c**) satisfaction ratings for Liked and Disliked food items are shown, both before and after task completion, for participants who completed post-task ratings (n=274). The Congruent-start group (starting in the Congruent block; n=143) is shown in the left panels, whereas the Incongruent-start group (starting in the Incongruent block; n=131) is shown in the right panels. Values are means ± s.e.m. Significant differences between pre-task and post-task ratings are denoted by ** ($p<0.01$) and * ($p<0.05$). Error bars represent the standard error of the mean (SEM).

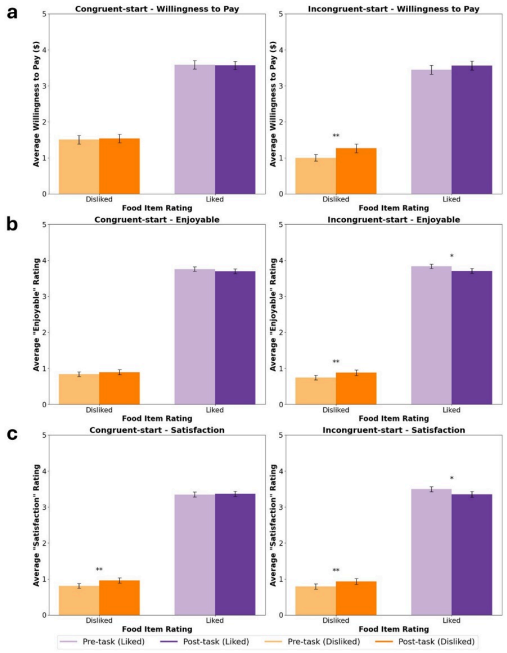


Figure 7

Changes in willingness to pay, enjoyment, and satisfaction for food items by start group and item rating. (**a**) Average willingness to pay, (**b**) enjoyment, and (**c**) satisfaction ratings for Liked and Disliked food items are shown, both before and after task completion, for participants who completed post-task ratings (n=274). The Congruent-start group (starting in the Congruent block; n=143) is shown in the left panels, whereas the Incongruent-start group (starting in the Incongruent block; n=131) is shown in the right panels. Values are means ± s.e.m. Significant differences between pre-task and post-task ratings are denoted by ** ($p<0.01$) and * ($p<0.05$). Error bars represent the standard error of the mean (SEM).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalFigures.pdf