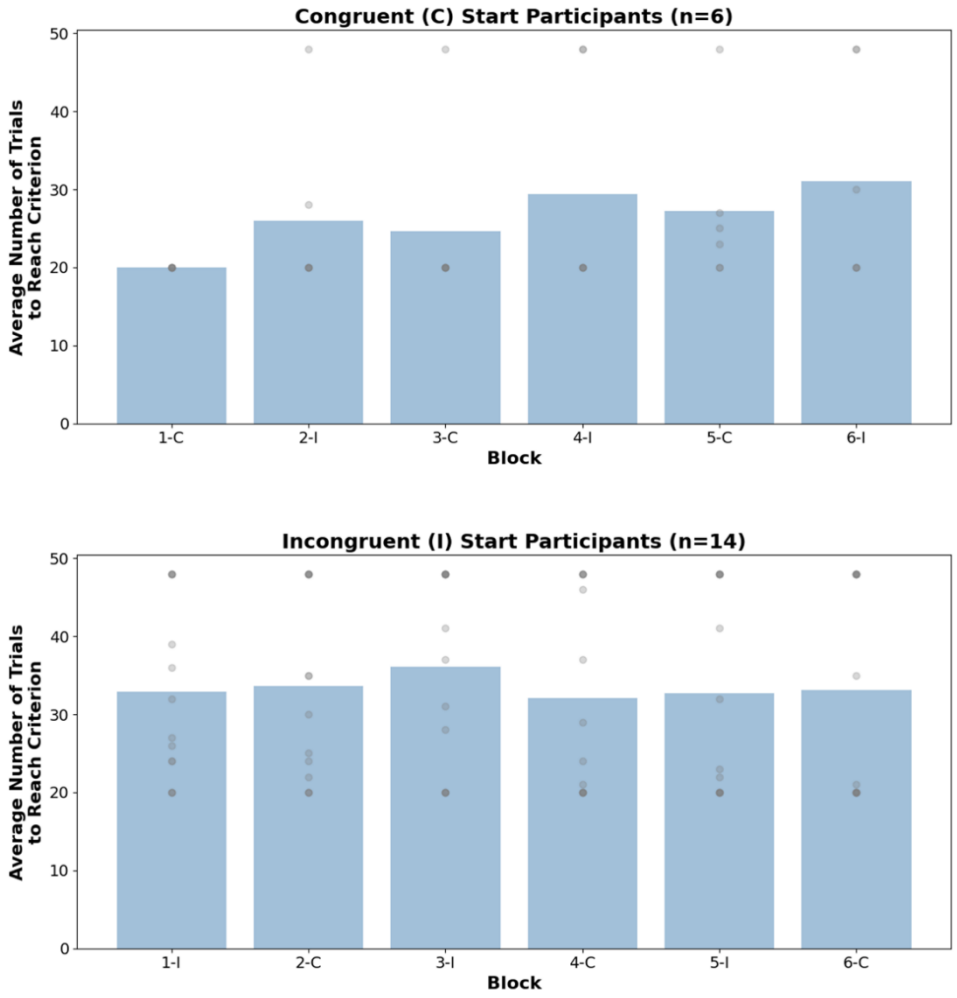
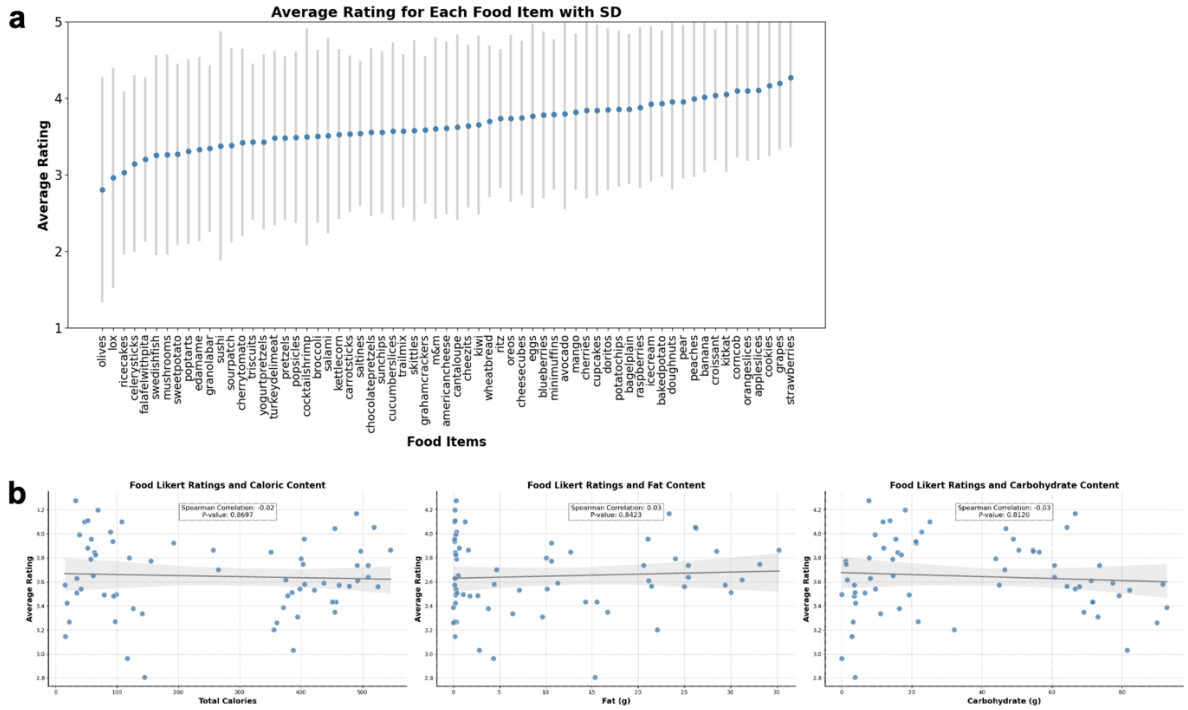


Supplemental Fig 1. On average, 31 trials are required to reach 75% accuracy within each block. Pilot data from online participants (n=20) in which participants had to achieve 75% accuracy within the last 10 trials of each block (after a minimum of 20 trials) before a block reversal occurred. The average number of trials required by participants to reach this learning criterion for a block reversal is shown for each block. The Congruent-start group (starting in the Congruent block; n=6) is shown in the top panel, whereas the Incongruent-start group (starting in the Incongruent block; n=14) is shown in the bottom panel. Data is further categorized by block type (Congruent and Incongruent). For Congruent-start participants, 26.36±3.86 trials were required on average, across blocks of both types, to achieve learning criterion for a reversal; for

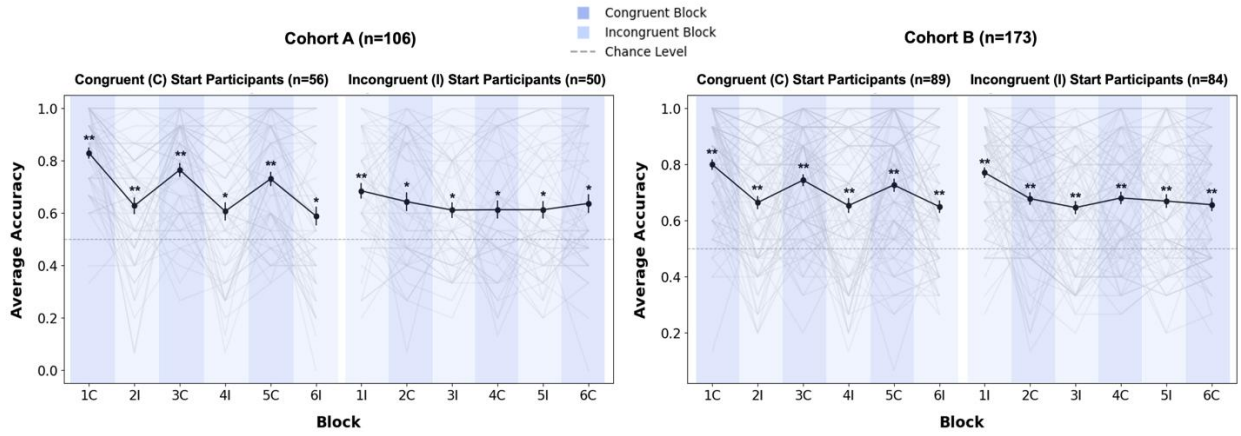
Incongruent-start participants, 33.42±1.40 trials were required, across blocks of both types. For all participants, regardless of starting group, the average number of trials required to meet learning criterion for a reversal was 31.30±1.31.



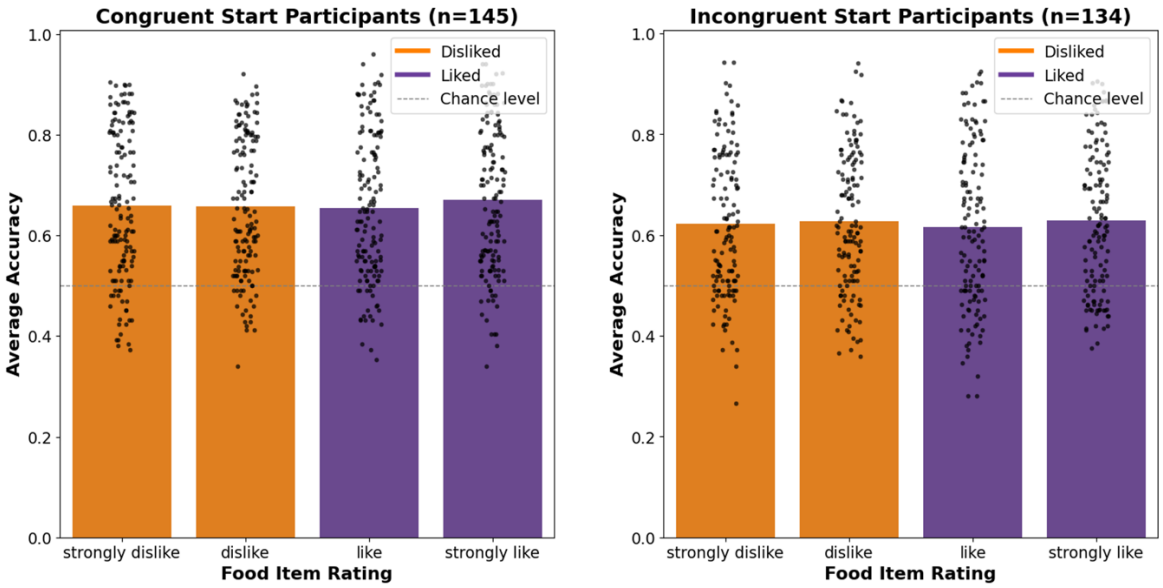
15

16 **Supplemental Fig 2. Food item ratings are heterogenous and not correlated with caloric or**
17 **nutritional content. a** Distribution of Likert ratings for the 65 food stimuli, obtained from the
18 full recruited sample (n=415). Likert values are coded as follows: 1=strongly disliked,
19 2=disliked, 3=neutral, 4=like, 5=strongly like. Values are means ± s.d., as indicated by the
20 surrounding light grey bars. **b** Spearman correlations among average food item rating and
21 caloric content (left panel), fat content (in grams; middle panel), and carbohydrate content (in
22 grams; right panel). Each data point represents a single food item from the 65-item stimuli set.

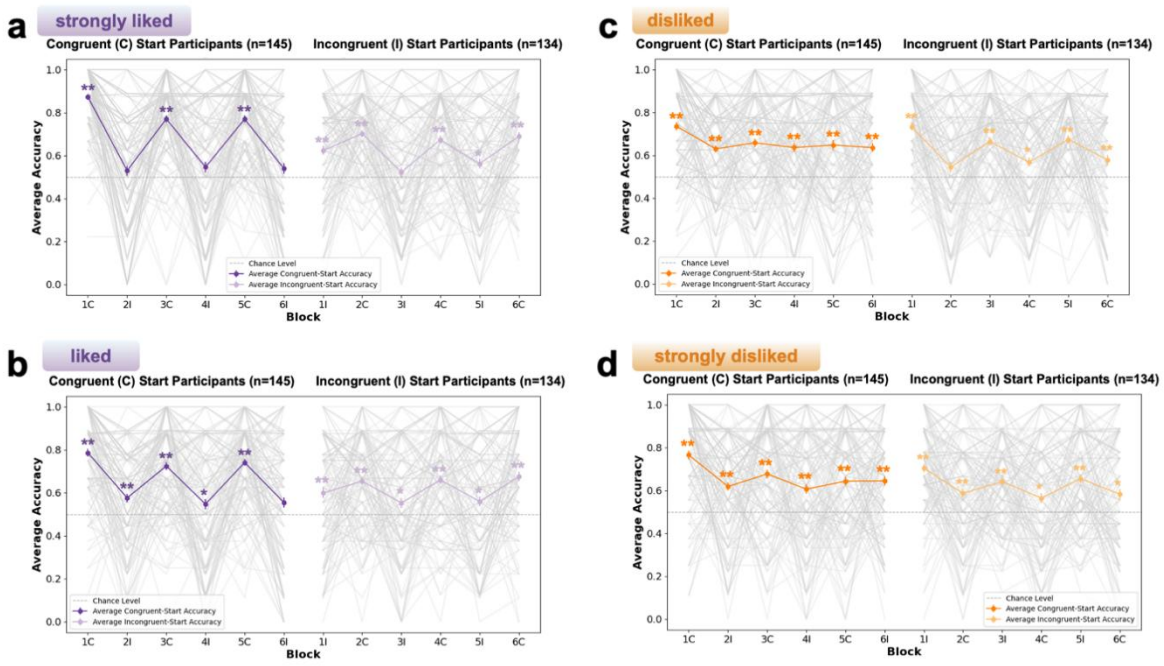
23



Supplemental Fig 3. Average accuracy of last 15 trials per block, by start group and block type, over time. 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 average accuracy of the last 15 trials in each block was significantly above chance level. End-of-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.

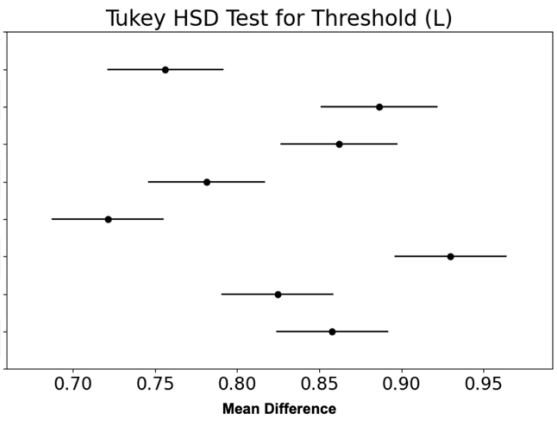


Supplemental Fig 4. Average accuracy does not differ as a function of food image rating alone. Accuracy data are categorized by food item rating. The Congruent-start group (starting in the Congruent block; n=145) is shown in the left panel, whereas the Incongruent-start group (starting in the Incongruent block; n=134) is shown in the right panel. P-values displayed above the bars indicate the significance of the difference between food items of different ratings 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.

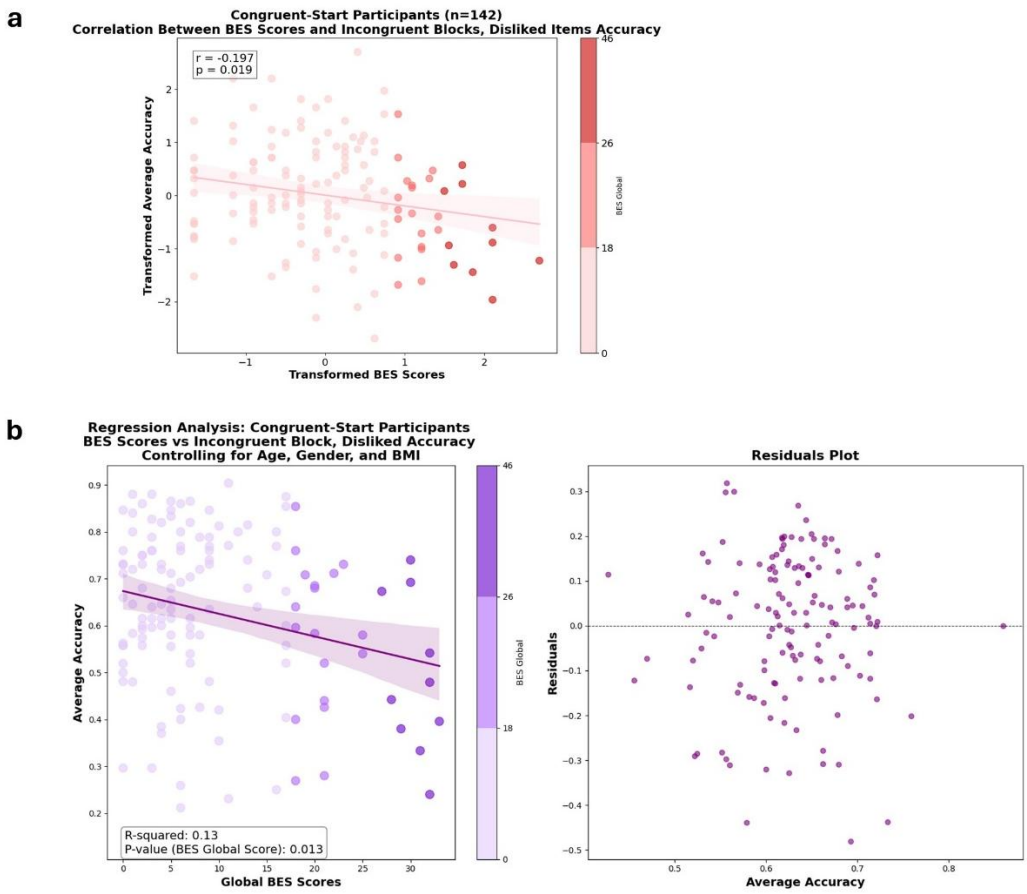


Supplemental Fig 5. Block-wise average accuracy by food item rating, start group, and block type. Average prediction accuracy is shown block-wise by collapsed food item rating, for (**a**) strongly liked, (**b**) liked, (**c**) disliked, and (**d**) strongly disliked food items. Accuracy data are additionally categorized by block type (Congruent and Incongruent) and the condition participants started in (Congruent-start and Incongruent-start). 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.

Starting group	Food item rating	Block type
Incongruent-start	Liked	Incongruent
Incongruent-start	Liked	Congruent
Incongruent-start	Disliked	Incongruent
Incongruent-start	Disliked	Congruent
Congruent-start	Liked	Incongruent
Congruent-start	Liked	Congruent
Congruent-start	Disliked	Incongruent
Congruent-start	Disliked	Congruent



Supplemental Fig 6. Sigmoid function fit to within-block average accuracy by food item rating, start group, and block type. An ANOVA was performed to investigate how model-derived thresholds (Ls) vary across task conditions. Tukey's HSD post hoc test results are displayed to showcase comparisons of threshold values across conditions. Values are mean differences ± 95% CI.



Supplemental Fig 7. For Congruent-start participants, binge eating severity is negatively associated with accuracy for Disliked items on Incongruent blocks. a Accuracy data of Congruent-start participants with survey data (n=142) on Incongruent blocks for Disliked items and Binge Eating Scale (BES) Global Scores underwent rank-based inverse normal transformation, followed by Pearson correlation. BES Global Scores are color-categorized according to severity, with increasingly dark shades of the color scale indicating Low, Mild to Moderate, and Severe binge eating, respectively. **b** Linear regression analysis examining the relationship between BES Global Scores and accuracy on Incongruent blocks with Disliked items among Congruent-start participants. The left panel shows a scatter plot of BES Global Scores against average accuracy, with a regression line, R-squared value, and p-value for the

BES Global Score coefficient indicating the relationship while controlling for age, gender, and BMI (kg/m^2). BES Global Scores are color-categorized according to severity, with increasingly dark shades of the color scale indicating Low, Mild to Moderate, and Severe binge eating, respectively. The right panel displays a residuals plot.

	Cohort A (n=106)	Cohort B (n=173)	Complete Sample (n=279)
Age			
Mean ± SD	34.74 ± 11.48	39.34 ± 12.53	37.62 ± 12.33
Gender			
Female	46	92	138
Male	49	76	125
Non-binary or Gender-fluid	5	1	6
Prefer not to say	1	0	1
Unknown	5	4	9
Race			
White	63	130	193
Black or African American	17	18	35
Asian	12	8	20
American Indian/Alaskan Native	1	1	2
Native Hawaiian or Other Pacific Islander	1	0	1
Other or Mixed	3	12	15
Prefer not to say	4	0	4
Unknown	5	4	9
Ethnicity			
Hispanic or Latino	9	18	27
Not Hispanic or Latino	92	151	243
Unknown	5	4	9
BMI (kg/m^2)			
Mean ± SD	26.64 ± 6.55	27.49 ± 6.38	27.18 ± 6.44
Binge Eating Scale Global Score			
Mean ± SD	10.32 ± 9.33	8.57 ± 8.32	9.22 ± 8.73

79 **Supplemental Table 1. Sample demographics.** Sample sizes reported for Cohort A (n=106),

80 Cohort B (n=173), and the Complete Sample (n=279) represent the number of participants who

completed the probabilistic reversal learning task and survived data quality exclusions. In the Complete Sample, 9 participants (”Unknown”) did not provide survey data.