Supplementary Materials

Contents

In this section we describe the results of seven pilot experiments that were conducted to create a viable stimulus set for both sweetness and screen-size ratings. Analysis scripts and data are available online via Open Science Framework (Gillies et al., 2022). A list of all the stimuli used (from both Blechert et al., 2019 and Foroni et al., 2013) can also be found via Open Science Framework (Gillies et al., 2022)

General Methods: Pilot Experiments

Participants

Each participant provided electronic consent to the protocol approved by the Research Ethics Boards of the University of Toronto prior to participation.

Pilot 1

Twenty participants were recruited from the University of Toronto, Mississauga, and were given course credit for participating. The sample had a mean age of 22.1 years. There were 14 females and 6 males, 16 were right-handed, three were left-handed, and one was ambidextrous. All participants had normal or corrected-to-normal vision.

Pilot 2

Twenty participants were recruited from the University of Toronto, Mississauga, and were given course credit for participating. The sample had a mean age of 18.6 years. There were 10 females and 10 males, 17 were right-handed, one was left-handed, and two were ambidextrous. All participants had normal or corrected-to-normal vision, with 11 wearing glasses, one wearing contacts, and the rest requiring neither.

Pilot 3

Nine participants were recruited from the University of Toronto, Mississauga, and were given course credit for participating. The sample had a mean age of 22.89 years. There were six females and three males, 8 were right-handed and one was ambidextrous. All had normal or corrected-to-normal vision except for one participant who had amblyopia (as this experiment did not require high visual acuity, the participant was not excluded). Five wore glasses, one contacts, one neither, and one declined to answer.

Pilot 4

Six participants were recruited from the University of Toronto, Mississauga, and were given course credit for participating. The sample had a mean age of 20.33 years. There were three males and three females. All were right-handed. All had normal or corrected-to-normal vision except one participant who stated they had a colour perception deficiency (as this experiment did not require intact colour vision, the participant was not excluded). Four wore glasses, and two wore neither glasses nor contacts.

Pilot 5

Eighteen participants were recruited from the University of Toronto, Mississauga and the University of Toronto, Scarborough, and were given course credit for participating. The sample had a mean age of 19.22 years. There were 13 females and five males. All were right-handed. All had normal or corrected-to-normal vision. Nine wore glasses, and nine wore neither glasses nor contacts.

Pilot 6

Participants were recruited via Prolific (Prolific, 2021), an online on-demand self-service data collection platform. Participants were pre-screened via prolific to ensure the following: they currently resided in the United States or Canada, were between the ages of 18 and 40, were fluent in English, had no head injuries, had no ongoing mental health conditions or illness, had no cognitive impairments or dementia, and had normal or corrected-to-normal vision.

Participants were paid an hourly rate of $12.85 CAD. Because the experiment lasted approximately 10-15 minutes, most participants earned a total of $4.29 CAD. Participants who completed any previous experiment were not permitted to participate in this experiment.

A total of 21 participants were recruited, with an average age of 28.43 years. There were 12 males and nine females; 20 participants were right-handed, and one was left-handed. All participants had normal or corrected-to-normal vision, with one person wearing contact lenses, 11 wearing glasses, and the rest requiring neither.

Pilot 7

Participants were recruited via Prolific (Prolific, 2021) using the same pre-screening criteria and payment details as described in Pilot 6. Participants who completed any previous experiment were not permitted to participate in this experiment.

A total of 20 participants were recruited, with an average age of 27 years. There were 7 males and thirteen females; 19 were right-handed, and one was left-handed. All had normal or corrected-to-normal vision, with 12 wearing glasses, 3 contact lenses, and the rest requiring neither.

Apparatus

Data were collected online due to the COVID-19 pandemic. Participants read the consent form and answered demographic questions on Qualtrics (Qualtrics, 2020). After giving consent, they were directed to Pavlovia (Pierce et al., 2019), which was the platform used to run the experiments. All experiments were coded using the Psychopy3 Experiment Builder (Pierce et al., 2013). Participants were only permitted to do the experiment on a desktop or laptop computer. Both Macs and Windows computers with various screen-sizes were used.

Stimuli

Stimuli were taken from the Food-Pics database (Blechert et al., 2019), and the FoodCast research image database (FRIDa) (Foroni et al., 2013). The pictures were displayed in a white rectangle in the middle of the participants' computer screen on a grey background. The size of each rectangle was .30×.225 times the screen's height.

A clickable rating scale was used to obtain participant ratings. The scale ranged from 0 to 10, and the numbers were presented on the scale below 11 corresponding tick marks. Participants made their ratings by clicking directly on the scale. Scale granularity was set to .50, enabling participants to click in between two whole values to obtain a half value. The scale was presented in white font in the lower third of the participants' screen, and the scale was 1.5 × .05 the screen's height. For sweetness ratings, presented above the scale were instructions in a white font reading "How sweet was that food? Click on the scale to make your sweetness ratings. 0 = not sweet at all, 10 = extremely sweet" (see Supplementary Figure 1). For screen-size ratings, the rating scale was similar, but ranged from 0 (small) to 10 (large) and above the 0 mark was the world "small", above the 5 mark the word "medium" and above the 10 mark the world "large", all in white font (see Supplementary Figure 2). Presented above the scale were instructions in white font reading "What size was that food compared to the white box? Click on the rating scale to make your response. 0 = small, 5 = medium 10 = large." Participants were encouraged to use the entire range of the scale.

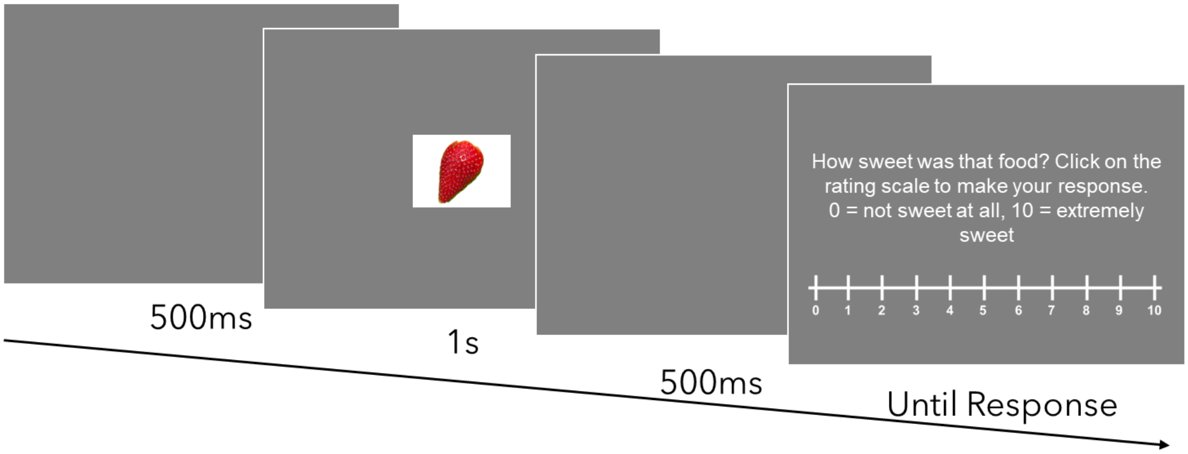
Procedure

Participants were instructed to make sweetness judgements (Pilot Experiments 1, 2, and 6) or screen-size judgements (Pilot Experiments 3, 4, 5, and 7) of different pictures of individual food items (see Supplementary Figures 1 and 2). Participants viewed the food pictures, in random order, for one second. Each picture was followed by a rating scale that participants used to indicate the perceived sweetness or size of the preceding food stimulus. The scale was present until a response was made. In between each trial and stimulus was a 500ms grey screen. Participants were not permitted to skip any trials and had an optional one-minute break every 50 trials.

Creating a Viable Stimulus Set for Sweetness and Screen-Size Ratings

Pilot Experiments 1 and 2: Selecting Sweetness Stimuli

We conducted two pilot experiments to create a database of food pictures that contained a variety of different types of foods (e.g., fruits, vegetables, animal products, grains, snacks, and desserts) across a broad range of perceived levels of sweetness (from not sweet at all to extremely sweet). Only images of single food items were included (e.g., a single pistachio, a loaf of un-sliced bread) and complex foods were not included (e.g., more than one individual food in a picture, such as a sandwich with meat and vegetables). Foods that would not be recognizable to North American participants were also not included (e.g., spaetzle, a Germanic noodle dish). Stimuli were taken from the Food-Pics database (Blechert et al., 2019) and the FoodCast research image database (FRIDa) (Foroni et al., 2013). All pictures were rectangular, with the food photographed against a white background.


How sweet was that food? Click on the rating scale to make your response. 0 = not sweet at all, 10 = extremely sweet
0
1
2
3
4
5
6
7
8
9
10
500ms
1s
500ms
Until Response

Supplementary Figure 1. Trial Sequence for Pilot Experiments 1, 2, and 6. Observers rated the perceived sweetness of pictures of individual food items. Food picture is from FreeFoodPhotos.com).

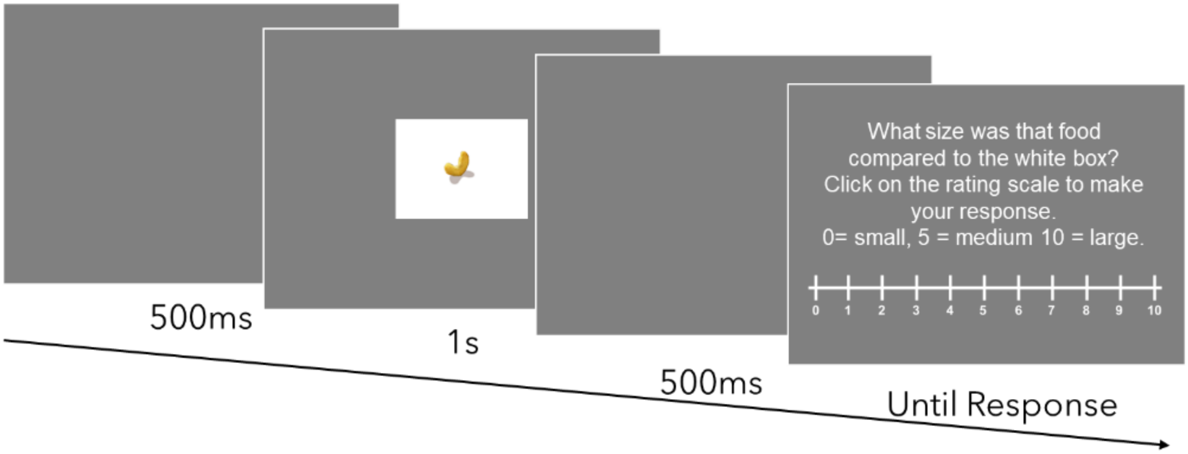
Pilot 1 used 547 images from the Food-Pics database. Pilot 1 was conducted to examine if the Food-Pics database contained enough images of food that spanned a wide range of sweetness. In both pilots (20 participants in each), participants rated the food pictures on a scale from 0 (not sweet at all) to 10 (extremely sweet) (see Supplementary Figure 1). To ensure that we had pictures across a broad range of possible sweetness ratings, images were sorted into five bins based on the average sweetness rating across participants (0-2, 2-4, 4-6, 6-8, 8-10). Within the bins, images were sorted based on the standard deviation of responses (i.e., a low standard deviation meant the participants agreed with one another on how sweet that food was). For Pilot 1, images that contained multiple food items (e.g., a bunch of grapes rather than a single grape) were flagged. Some pictures were later edited in Photoshop (see below), and some were excluded entirely. In sum, 354 images were removed entirely, leaving 192 images for use in Pilot 2.

The results of Pilot 1 revealed that some bins were under-represented (e.g., bin 4-6 had only 27 items) and others were over-represented (e.g., bin 0-2 contained 54 images) or contained too many duplicate food items (e.g., eight pictures of red apples in one bin). Pilot 2 addressed these issues by adding additional pictures from the FRIDa database, which contained more types of fruits (e.g., a blood orange) and sweeter foods than the Food-Pics database. Pilot 2 used 46 images from the FRIDa database. Using the data from Pilot 1, 192 images that 1) had low standard deviations of response compared to the other images within a bin, and 2) were not over-represented in the image database (e.g., not all eight pictures of apples were used) were selected from the food-pics data base. In addition, for Pilot 2, 64 images (58 from Food-Pics and six from FRIDa) were edited in Photoshop (Adobe Photoshop CS, 2004) to ensure that only a single food item was present (e.g., a single grape rather than bunch of grapes), leading to a total of 302 food pictures for Pilot 2 (participants rated both the edited and un-edited images). Images were again binned based on the average response across participants, and we then selected 150 food pictures across a broad range of sweetness ratings (approximately 30 images per bin) with low standard deviations (the average SD response of the images that were included was 1.81, ranging from of .89 to 2.81. The average SD response of the images that were excluded was 2.02, ranging from .78 to 3.37).

The resulting image database contained 127 pictures from the Food-Pics database (Blechert et al., 2019), 34 of which were edited in photoshop (Adobe Photoshop CS, 2004) to ensure that only a single food item was present in each picture. We used an additional 23 pictures from the FRIDa database (Foroni et al., 2013), four of which were edited in Photoshop. These 150 pictures were used in Pilot Experiment 6, and Experiments 1, 2, and 4.

Pilot Experiments 3 through 5: Selecting screen-size stimuli

We conducted three pilot experiments to ensure the food pictures varied across a broad range of screen-sizes, using the same methods as Pilots 1 and 2.


What size was that food compared to the white box? Click on the rating scale to make your response. 0= small, 5 = medium 10 = large.
0
1
2
3
4
5
6
7
8
9
10
500ms
1s
500ms
Until Response

Supplementary Figure 2. Trial Sequence for Pilot Experiments 3, 4, 5, and 7. Observers rated the perceived screen size of pictures of individual food items. Food picture is from Food-Pics (Blechert et al., 2019).

In Pilot 3, nine participants rated the perceived screen-size of 238 food images (the same images used in Pilot 2, minus the 64 un-edited images) on a scale from 0 (small) to 10 (large) (see Supplementary Figure 2), and images were sorted into five bins based on the average screen-size rating, and by the standard deviation of response within each bin. We found that bin five (images with a screen-size rating of 8-10) was underrepresented, with only 16 items.

In Pilot 4, images were edited in Photoshop to manipulate their screen size. Of the 238 images, 18 images from the Food-Pics database were edited, and seven images from the FRIDA image database were edited to change their size (five images were made smaller, 20 were made larger). The results of Pilot 4 (6 participants) revealed that both bins one and five were slightly under-represented (26 and 28 food items, respectively).

In Pilot 5 (18 participants), of the 238 images, an additional 15 images from Food-Pics and four images from FRIDA were edited to change their size (eight were made smaller, 11 were made larger). Of those images, four from Food-Pics and two from FRIDA were previously edited to ensure only a single food was present (i.e., they were edited twice).

All bins contained at least 30 images. We then selected 150 food pictures across a broad range of screen-size ratings with low standard deviations (the average SD of the images that were included was 1.22, ranging from .36 to 1.75. The average SD of the images that were excluded was 1.88, ranging from 1.78 to 2.52).

This yielded a final image database of 150 food pictures, with 122 coming from the Food-Pics database, and the remaining 28 were from the FRIDa database. Of the 122 from Food-Pics, 37 were edited in Photoshop to ensure only a single food was present, 19 were edited to change their size, and four were edited in both ways. Of the 28 pictures from FRIDA, four were edited to ensure only a single food was present, seven were edited to change their size, and two were edited in both ways. These 150 pictures were used in Pilot Experiment 7, and Experiment 3.

Pilot Experiment 6: Individual Sweetness Ratings

In biology, taste is defined as when a certain class of chemical contacts specialized taste receptors on the tongue, palate, and throat (Breslin & Spector, 2008). The sensation of a sweet taste is due to the presence of simple carbohydrates (sugars) on the tongue (Breslin & Spector, 2008). The greater amount of simple carbohydrates a food has, the sweeter it is. Taste perception, therefore, can be argued to be somewhat objective. Prior to showing participants arrays of multiple foods (food ensembles) to examine ensemble coding of sweetness, we must first ensure that there is consistency in how participants rate the sweetness of individual foods. If food sweetness judgements have some level of objectivity, there should be high agreement between participants regarding how the sweetness of foods are rated (e.g., on a scale of 0-10 for sweetness, most participants would rate a strawberry about 7). Participants rated the 150 final food-pictures derived from Pilot Experiment 2 on their perceived sweetness.

Results

We evaluated participants' consistency in rating sweetness by using an intraclass correlation coefficient (ICC) test. Specifically, a two-way mixed model was used to measure consistency across ratings, and the test yielded a score of .98, which is within the "excellent" range (Cicchetti, 1994). Therefore, the foods were rated very similarly across the different participants. Each food picture was assigned a "sweetness score" calculated by averaging the ratings for that food from the 21 participants. The pictures ranged in sweetness from .62 to 9.00.

To ensure that participants were using the whole scale and not just categorizing foods using a binary system of "sweet, not sweet", we sorted the pictures into five bins based on their average rating. Bin 1 had 31 items with an average rating between 0 and 1.99, bin 2 had 26 items between 2 and 3.99, bin 3 had 36 items between 4 and 5.99, bin 4 had 23 items between 6 and 7.99, and bin 5 had 34 items between 8 and 10. Therefore, each bin is roughly equally represented in this image database, meaning that participants were using the full range of the scale and were not using a binary categorization strategy when judging perceived sweetness.

Pilot Experiment 7: Individual Screen-Size Ratings

Size can be considered a low or mid-level visual feature (Whitney & Yamanashi Leib, 2018), and screen-size computations can be performed using information that is directly available on the retina and does not require retrieval of information from long-term memory (LTM). In other studies that examined ensemble coding for average screen-size (e.g., Ariely, 2001; Chong & Treisman, 2003; Corbett et al., 2012), the stimuli belonged to the same category (e.g., all the same simple shape such as circles). The food picture stimuli used in our study are much more variable in terms of visual features. Importantly, the food items used varied significantly in their outline shape. Therefore, it was unknown if observers would agree with one another on the perceived screen-size of the stimuli. Due to this problem, we compared ratings of average screen-size (in Experiment 3) to predicted screen-size ratings generated by a different crowd of raters in Pilot Experiment 7, eschewing the need to obtain the actual pixel size of all the images. To obtain a measure of “screen-size”, we asked participants to rate the screen-size of the food stimuli relative to the white rectangle the food was presented in. This method was optimal as the study was conducted online, and participants used monitors of all different sizes. This way, perceived screen-size would be consistent, even when computer monitors were differently sized.

Results

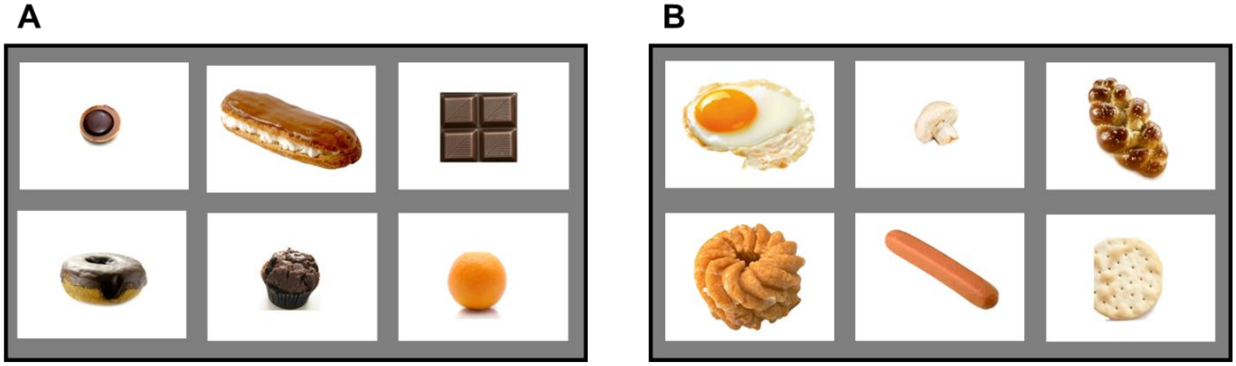
Participants rated the perceived screen-size of the final 150 food-pictures derived from Pilot Experiment 5. Using the same analysis as Pilot 6, we found that the food pictures were rated very similarly on their perceived screen-size (*ICC* = .99), which, like the ratings of perceived sweetness, is within the “excellent” range (Cicchetti, 1994). The pictures ranged in screen-size ratings from .2 to 9.65. To examine how participants used the scale, we sorted the images into five bins based on their average rating. Bin 1 had 28 images with an average rating between 0 and 1.99, bin 2 had 28 images with an average rating between 2 and 3.99, bin 3 had 35 images with an average rating between 4 and 5.99, bin 4 had 26 images with an average rating between 6 and 7.99, and bin 5 had 33 images with an average rating between 8 and 10. Thus, each bin was roughly equally represented in the image database, showing that participants were using the full range of the provided scale. Ensembles presented in Experiment 3 were generated using the same method as described in Pilot 6.

Vision and Taste

A question that arises from the results of this work is whether or not these “cross-modal” ensemble judgements can be computed using low-level visual information, specifically colour. Indeed, colour is one of the most important sensory cues regarding how people set taste and flavour expectations (Spence, 2015; Spence et al., 2010). The question is if colour alone can provide enough taste information to guide observers’ judgements of average taste.

There is some evidence that colour names are more likely to be associated with certain taste labels (e.g., “red” and “sweet”, “yellow” and “sour”) (O’Mahony, 1982). Some studies suggest that colour impacts the intensity of certain taste experiences. For example, a sweet beverage dyed red will be perceived as sweeter than a beverage without dye (see Spence, 2015, for a review). However, to date, no one has studied how one colour may contribute to perceived taste across taste categories (e.g., salty vs. sweet). For example, does a salty beverage dyed red taste sweeter than a salty beverage with no colorant? Previous studies also only show that vision can interact with taste (e.g., colour can moderate taste intensity), but no study to our knowledge has shown that colour alone can act as a salient taste cue (e.g., this is red therefore it must be sweet).

Within a single food category (such as “apples”), colour may be used as a cue to determine sweetness. For example, a red apple is likely to be sweeter than a green apple. Indeed, in our food picture database, the green apple was rated as less sweet than the red apple (sweetness scores of 4.45 and 5.31, respectively). However, colour cannot be used in the same manor across different food categories. A reddish salami was rated as not sweet, for example. If observers were using colour alone, their average ratings should not have approximated the predicted sweetness ratings, given the variability of the colour-sweetness relationships across the foods. Indeed, based on the predicted average sweetness values calculated from the results of Pilot Experiment 6, both the sweetest and least-sweetest ensembles contained foods of similar colours (i.e., mostly brown or tan, with both containing one food that is orange and circular: a mandarin for the sweetest ensemble, and an egg with a yolk for the least sweet ensemble; see Supplementary Figure 3). The average response on the set-size six display for the sweetest ensemble was 8.36 (predicted sweetness was 7.85), and the average response on the least sweet ensemble was 2.34 (predicted sweetness was 2.67), showing that even when displays had similar colours, the participants' average sweetness ratings were quite different. This demonstrates that colour alone could not have been reliably used to generate accurate ratings of perceived sweetness, and thus our results are not likely confounded by differences in colour across the food pictures that we used.



Supplementary Figure 3. Ensembles with Similar Colours but Different Sweetness Values.
A) One of the sweetest and B) one of the least sweet ensembles used in the Experiments. Pictures are from the Food-Pics (Blechert et al., 2019). The rest of the food pictures (orange, chocolate, muffin, and cracker) are freely available illustrative examples of the stimuli from FreeFoodPictures.com but are not images used in the actual study.

Lastly, as colour is an important cue used to perceive the identity of a food item, this task would likely be rendered extremely difficult if colour cues were removed altogether. Although colour alone does not determine the sweetness of a food item, this does not mean that colour is uninformative when considering the sweetness of a food. Consider, for example, the difficulty in rating the perceived sweetness of a greyscale strawberry. Moreover, the subtle differences between foods such as oranges and grapefruits may also be lost if colour cues are removed, as other diagnostic visual features (i.e., shape, texture) are very similar between the two. In summary, although colour is an important visual cue related to perceived sweetness, differences in colour cues alone likely cannot explain the results of our study, given how colour contributes to sweetness perception differentially across food categories.

References

Adobe Photoshop CS. (2004). Berkeley, CA: Peachpit Press.

Blechert, J., Lender, A., Polk, S., Busch, N. A., & Ohla, K. (2019). Food-picsextended—an image database for experimental research on eating and appetite: additional images, normative ratings and an updated review. *Frontiers in psychology*, *10*, 307. https://doi.org/10.3389/fpsyg.2019.00307

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision research*, *43*(4), 393-404. https://doi.org/10.1016/S0042-6989(02)00596-5

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, *6*(4), 284. https://doi.org/10.1037/1040-3590.6.4.284

Corbett, J. E., Wurnitsch, N., Schwartz, A., & Whitney, D. (2012). An aftereffect of adaptation to mean size. *Visual cognition*, *20*(2), 211-231. https://doi.org/10.1080/13506285.2012.657261

Foroni, F., Pergola, G., Argiris, G., & Rumiati, R. I. (2013). The FoodCast research image database (FRIDa). *Frontiers in human neuroscience*, *7*, 51. https://doi.org/10.3389/fnhum.2013.00051

Free Food Photos, https://www.freefoodphotos.com.

Gillies, G., Fukuda, K., & Cant, J. (2022, October 19). Cross-Modal Ensemble Data and Analysis. https://doi.org/10.17605/OSF.IO/GTMDB

O'Mahony, M. (1983). Gustatory responses to nongustatory stimuli. *Perception*, *12*(5), 627-633. https://doi.org/10.1068/p120627

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior research methods*, *51*(1), 195-203. https://doi.org/10.3758/s13428-018-01193-y

Prolific. (2021). Oxford, UK. Available at: https://www.prolific.co

Qualtrics. (2020). Provo, Utah, USA. Available at: https://www.qualtrics.com

Spence, C. (2015). On the psychological impact of food colour. *Flavour*, *4*(1), 1-16. https://doi.org/10.1186/s13411-015-0031-3

Spence, C., Levitan, C. A., Shankar, M. U., & Zampini, M. (2010). Does food color influence taste and flavor perception in humans?. *Chemosensory Perception*, *3*(1), 68-84. https://doi.org/10.1007/s12078-010-9067-z

Whitney, D., & Yamanashi Yamanashi Leib, A. (2018). Ensemble perception. *Annual review of psychology*, *69*, 105-129. https://doi.org/10.1146/annurev-psych-010416-044232