AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

© 2023 American Psychological Association
ISSN: 0096-3445

2024, Vol. 153, No. 1, 56–69
https://doi.org/10.1037/xge0001470

Testing the Flexibility of Ensemble Coding: Limitations in Cross-Modal Ensemble Perception

Greer Gillies[1], Keisuke Fukuda[2], and Jonathan S. Cant[3]
[1] Department of Psychology, University of Toronto
[2] Department of Psychology, University of Toronto, Mississauga
[3] Department of Psychology, University of Toronto, Scarborough

Ensemble coding (the brain's ability to rapidly extract summary statistics from groups of items) has been demonstrated across a range of low-level (e.g., average color) to high-level (e.g., average facial expression) visual features, and even on information that cannot be gleaned solely from retinal input (e.g., object lifelikeness). There is also evidence that ensemble coding can interact with other cognitive systems such as long-term memory (LTM), as observers are able to derive the average cost of items. We extended this line of research to examine if different sensory modalities can interact during ensemble coding. Participants made judgments about the average sweetness of groups of different visually presented foods. We found that, when viewed simultaneously, observers were limited in the number of items they could incorporate into their cross-modal ensemble percepts. We speculate that this capacity limit is caused by the cross-modal translation of visual percepts into taste representations stored in LTM. This was supported by findings that (a) participants could use similar stimuli to form capacity-unlimited ensemble representations of average screen size and (b) participants could extract the average sweetness of displays when items were viewed in sequence, with no capacity limitation (suggesting that spatial attention constrains the number of necessary visual cues an observer can integrate in a given moment to trigger cross-modal retrieval of taste). Together, the results of our study demonstrate that there are limits to the flexibility of ensemble coding, especially when multiple cognitive systems need to interact to compress sensory information into an ensemble representation.

Public Significance Statement
Ensemble coding (the brain's ability to rapidly extract summary statistics from groups of similar objects) has been found to operate across a range of visual features from low-level (e.g., average color) to high-level (e.g., average facial expression). Beyond concrete physical features, ensemble coding has also been found to operate on more abstract concepts (e.g., average lifelikeness and average cost). Though ensemble coding can be applied to both physical and abstract impressions, the computational limitations of this ability are unclear. We tested the flexibility of ensemble coding by examining interactions between sensory systems (e.g., vision and taste) by asking observers to produce an average taste (sweetness) from the visually presented information. Broadly, our results thus demonstrate that there is a limit to the flexibility of ensemble coding, particularly when multiple cognitive systems must interact (e.g., visual perception, taste perception, and long-term memory) to compress sensory input into an ensemble representation.

Keywords: ensemble coding, vision, taste, cross-modal

Supplemental materials: https://doi.org/10.1037/xge0001470.supp

This article was published Online First September 21, 2023.

Greer Gillies https://orcid.org/0000-0003-3570-4000

Part of this work was presented at the Object Perception, Visual Attention, and Memory (OPAM) Conference in 2021 and the Vision Sciences Society (VSS) Conference in 2022. This work was funded by the NaturalSciences and Engineering Research Council (NSERC) Discovery Grant (435647) to Jonathan S. Cant and an NSERC Discovery Grant (5009170) to Keisuke Fukuda. The authors would like to thank the Cant and Fukuda labs, with special thanks to Kristina Knox and the rest of the 2020 University of Toronto Psychology Graduate Cohort. The authors have no conflicts of interest to report. Data and analysis scripts are freely available on Open Science Framework at https://doi.org/10.17605/OSF.IO/GTMDB.

Greer Gillies served as lead for conceptualization, data curation, formal analysis, investigation, methodology, project administration, visualization, writing–original draft, and writing–review and editing. Keisuke Fukuda served in a supporting role for conceptualization, project administration, writing–original draft, and writing–review and editing. Jonathan S. Cant served in a supporting role for conceptualization, project administration, writing–original draft, and writing–review and editing. Keisuke Fukuda and Jonathan S. Cant contributed equally to funding acquisition, software, supervision, validation, and resources.

Correspondence concerning this article should be addressed to Greer Gillies, Department of Psychology, University of Toronto, 1265 Military Trail, Scarborough, ON M1C1A4, Canada. Email: greer.gillies@mail.utoronto.ca

Our perception of the world as rich and highly detailed is at odds with a body of research that suggests that what we are capable of perceiving is limited by constraints on attention and visual working memory (VWM; Cohen et al., 2016; Luck & Vogel, 1997). Why then do we describe our subjective visual experience as complete and representative of reality? Our visual system's sensitivity to visual regularities is one of the factors that underlies our subjective experience (Alvarez, 2011). Because information within scenes is not completely random, with scenes containing groups of similar objects and features (Whitney & Yamanashi Leib, 2018), the brain can utilize what is called ensemble coding—the rapid extraction of summary statistical information from groups of items—to represent a large amount of information as a single summary statistic (e.g., the average color of leaves on trees in a forest). Ensemble coding enables us to circumvent at least some of the capacity limits to attention and VWM (Khayat & Hochstein, 2018) such that we can abstract the basic perceptual and conceptual representation (the gist) of a scene in a fraction of a second (e.g., within a few hundred milliseconds; Whitney & Yamanashi Leib, 2018). In this series of experiments, we investigated potential limits to the cognitive mechanisms underlying ensemble coding, examining if different sensory modalities (i.e., vision and taste) can interact when forming ensemble percepts.

Ensemble Coding Across Visual Features

Ensemble coding has been demonstrated for low-level features such as color (e.g., the average color of a group of simple shapes (Kuriki, 2004, Maule & Franklin, 2015; 2016; Webster et al., 2014)), mid-level features such as size (e.g., the average size of a group of simple shapes; Ariely, 2001; Chong & Treisman, 2003), and high-level features such as facial expression (e.g., the average facial expression of a group of faces; Haberman et al., 2009; Haberman & Whitney, 2007, 2009; Li et al., 2016). In the aforementioned studies, observers were able to extract summary statistics under brief viewing durations (e.g., several hundred milliseconds). Furthermore, this ability was not sensitive to set size, as observers could produce a summary statistic that was accurate for as many as 16 simple shapes and complex stimuli (e.g., Haberman & Whitney, 2007; Maule & Franklin, 2015). In contrast, some studies have found that larger set sizes confer a benefit to ensemble coding performance in terms of both accuracy (e.g., Allik et al., 2013; Baek & Chong, 2020; Solomon, 2010) and response time (Robitaille & Harris, 2011). Interestingly, observers typically demonstrate worse accuracy when performing a member-identification task ("Was this item in the display you saw previously?") compared to when they were asked to report an average feature of a display (Ariely, 2001; Haberman & Whitney, 2007, 2009). In summary, individuals can obtain highly accurate summary statistical information for low-, mid-, and high-level features when viewing ensemble displays for a fraction of a second, despite retaining little-to-no information about the individual items that make up those ensembles. This latter point demonstrates the existence of distinct cognitive mechanisms involved in the processing of ensembles and single items (Cant et al., 2015), underscoring the observation that ensemble coding may be a process that is utilized to circumvent limitations to visual attention and VWM.

Ensemble Coding Beyond Concrete Visual Features

Moving beyond simple visual features, Yamanashi Leib et al. (2016) examined ensemble coding for abstract visual impressions, asking if it is possible to obtain summary statistical representations for object lifelikeness (i.e., animacy). They found that observers were sensitive to not only the animacy of random single objects (observers agreed with one another as to how animate the objects were), but they were also sensitive to the average animacy of groups of objects. Object lifelikeness cannot be computed using information from a single feature (e.g., color and size). Rather, object lifelikeness or animacy likely arises from a number of different features interacting together. This was the first study to show that observers could extract summary statistical impressions that were not immediately specified by the basic visual features of the images.

There is also evidence to suggest that ensemble coding can utilize information from other cognitive systems such as long-term memory (LTM). Yamanashi Leib et al. (2020) found that observers were sensitive to the average economic value (i.e., cost) of groups of objects. Like lifelikeness, cost cannot be computed using basic feature information, as the visual features associated with an object's value vary across product categories such that an ensemble value for multiple objects would not be related to the shared features of those objects.

Observers would have needed to retrieve semantic information from LTM about the cost of the objects, given that simple feature information (e.g., color, shape, and size) alone is not enough to make such a judgment. This shows that individuals can also rapidly retrieve information from LTM during ensemble coding. Of note, we are not implying that cost cannot be derived using visual feature information. On the contrary, for the majority of stimuli used by Yamanashi Leib et al. (2020), visual features were more diagnostic for cost estimates than other sensory features (i.e., olfactory, auditory, or somatosensory information would not have been as helpful in reliably deriving estimates of cost). While visual features were used to retrieve semantic information about cost from LTM, our point here is that the same visual features could not have been used across different objects to make the cost judgments.

Other studies have shown that observers can rapidly extract semantic information related to object categories from large sets of images. Khayat and Hochstein (2019) showed observers rapid streams of images (100 ms/item) of a certain category (e.g., "mammals"). Observers were then presented with two images and were asked to select the image that appeared in the stream. The "new" image could be a prototype of the category (e.g., lion), a category member but not a prototype (e.g., squirrel), or a noncategory member (e.g., owl). Participants were more likely to select an item as being "old" if the item was a prototypical member of the category, regardless of whether the test item was new or old. Overall, participants perceived categorical information better than information about individual items within the rapid serial visual presentation stream, and the authors conclude that object categorization may share perceptual–computational mechanisms with summary statistic perception.

While ensemble coding can be applied to a wide range of both physical and abstract features, and can involve cognitive systems such as memory, at present it is unclear if there are limits to the flexibility of how ensemble coding is used to compress information from groups of items into useful summary statistical metrics.

The Current Study

As discussed above, findings from Yamanashi Leib et al. (2016, 2020) demonstrated that observers can retrieve semantic information from LTM when computing average lifelikeness and economic value, which, combined with previous research utilizing simpler stimuli (for a review, see Whitney & Yamanashi Leib, 2018), showcases the utility of ensemble coding across a range of visual scenarios. However, our everyday experiences are multisensory rather than unisensory in nature. For example, eating is a multisensory experience (Spence, 2015). Consider a bowl of strawberries. We use our sense of vision to find the most-ripe strawberry (see the online supplemental materials for a discussion of how color influences taste perception), we can also use our sense of tactician to determine ripeness (is the strawberry mushy?). With olfaction, we determine the flavor of the strawberry (via directly smelling the strawberry but also when it is in our mouths via orthonasal olfaction), and with our sense of taste we can determine how sweet and sour the strawberry is. One way to investigate interactions between different senses is to examine if people can use visual information to retrieve representations that belong to another sensory modality stored in LTM during ensemble coding.

To examine if such an ability is possible, we asked observers to make judgments about the average taste (e.g., sweetness) of groups of visually presented food pictures. While some visual features may contribute to taste perception (Spence, 2015; Spence et al., 2010) (e.g., red strawberries are sweeter than pale ones), information conveyed by these features is not consistent (e.g., red foods are not always sweet). Another important factor is experience with these foods stored in LTM. It would be difficult to determine that an orange is sweeter than a grapefruit based solely on visual input, as these foods have similar visual features. Rather, you would have to know that a grapefruit is sour based on previous experience with that food. To expand on this idea, "sweetness" may initially be encoded in a unisensory manner (e.g., taste can be determined in the absence of other senses), but some impressions of taste may be more multisensory experiences (e.g., interactions between vision, olfaction [flavor], and taste). Regardless of how taste information is initially encoded, observers would need to use one sense (vision) to retrieve a stored representation of taste from LTM, moving beyond what has been previously shown possible with ensemble coding.

To preview our results, we found that when pictures of foods were viewed simultaneously, observers were limited in the number of items they could incorporate into their ensemble percepts of average sweetness. This capacity limit is likely driven by the information transformation required (vision to taste), as observers could use the same food stimuli to generate a summary statistic for average screen size (no information transformation required). We found that utilizing a sequential display removed the capacity limit when computing cross-modal ensemble percepts of average sweetness, suggesting that when the information transformation required is sufficiently complex, spatial attention becomes a limiting factor (e.g., the computation of average taste cannot be done in parallel).

Experiment 1

The purpose of Experiment 1 is to evaluate if different sensory modalities can interact during ensemble coding. This experiment requires that participants be shown visual arrays containing pictures of multiple different food items. The perceived sweetness of the food stimuli (see Figure 1) used in this and the subsequent experiments were validated over a series of pilot studies (see the online supplemental materials). Briefly, a separate group of raters viewed 150 individual food pictures and rated them on their perceived sweetness on a scale from 0 (*not sweet at all*) to 10 (*extremely sweet*). Importantly, an intraclass correlation coefficient score of .98 (Cicchetti, 1994) showed that observers were in high agreement with one another.

In this experiment, participants were asked to rate the average sweetness of visual arrays of different foods. Importantly, showing that this form of ensemble coding is possible requires demonstrating that participants can integrate taste information from multiple items present in an ensemble, rather than merely reporting the sweetness of a single item. To accomplish this, we used a subset manipulation identical to that of Yamanashi Leib et al. (2016). Specifically, in some trials, participants were only shown part of the whole ensemble (i.e., one, two, or four of the six items) and were asked to report the average sweetness. Their responses were compared to the predicted sweetness of the full six-item ensemble. Since the subsets are not representative of the average sweetness of the full six-item ensemble, this analysis simulates what might occur if participants use a subsampling strategy (i.e., only use some of the available information) when viewing the full six-item ensemble. That is, if participants employ a subsampling strategy, the correlation between participants' actual ratings of average sweetness and the predicted ratings will plateau at the set sizes equal to the subsampling limit (see Figure 2). However, if participants can integrate information from multiple items, then the correlation would increase as the set size increases, suggesting that participants can form cross-modal representations of ensemble sweetness. To preview our results, we found that participants could integrate up to four out of the available six items into their average sweetness ratings.

Method

Transparency and Openness

Data and analysis scripts from the experiments (including all pilots) are available on Open Science Framework at https://doi.org/10.17605/OSF.IO/GTMDB.

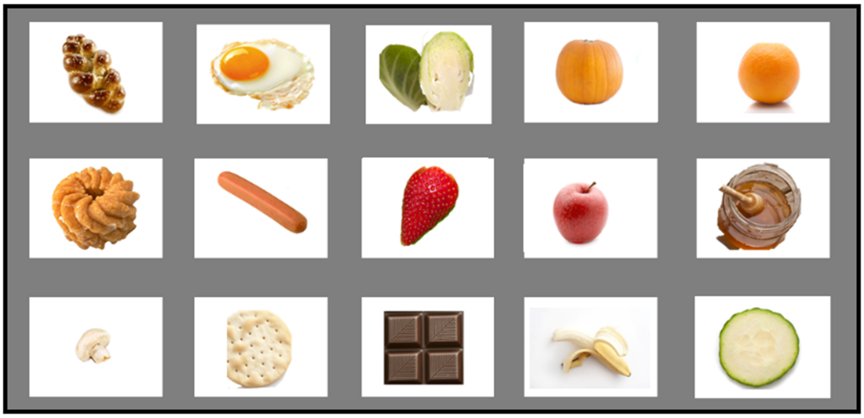
Constraints on Generality

The selected sample was individuals from North America (the United States and Canada). This sample was selected as we needed to ensure that observers were familiar with the foods used in the study. As such, using the same image database on a different population may not yield the same pattern of results (e.g., participants may not be familiar with the foods, and show low agreement on perceived sweetness). However, we do not have reason to believe that the underlying cognitive mechanisms would differ across cultures. To study cross-modal ensemble coding for taste from vision, it is necessary to use an image database that contains food pictures that the observers are familiar with.

Participants

Participants were recruited via Prolific (Prolific, 2021), an online on-demand self-service data collection platform. Participants were prescreened 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

Figure 1
Sample Stimuli Used in the Average Sweetness Experiments

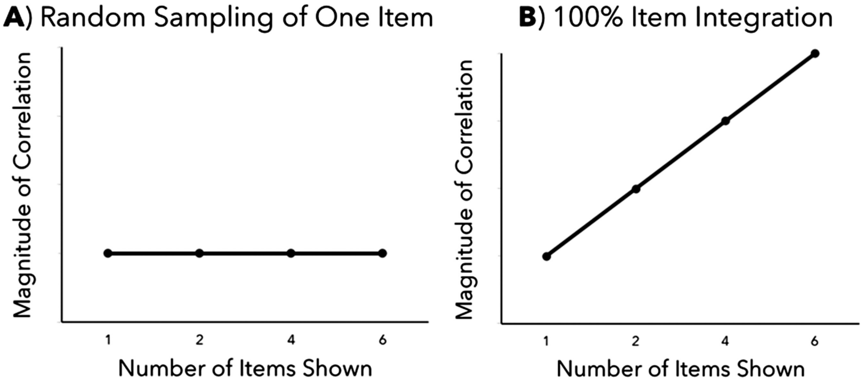


Note. An example of the food pictures used in Experiments 1, 2, and 4. Pictures are from the Food-Pics (Blechert et al., 2019) database (bread, egg, donut, hotdog, mushroom). The rest of the food pictures are freely available illustrative examples of the stimuli (https://www.freefoodphotos.com) but are not images used in the actual study. The Food-Pics are adapted from "Food-Pics_Extended–An Image Database for Experimental Research on Eating and Appetite: Additional Images, Normative Ratings and an Updated Review," by J. Blechert, A. Lender, S. Polk, N. A. Busch, and K. Ohla, 2019, *Frontiers in Psychology*, *10*, p. 307 (https://doi.org/10.3389/fpsyg.2019.00307). CC BY-NC. See the online article for the color version of this figure.

ongoing mental health condition 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 ~10–15 min, most participants earned a total of 4.29 CAD. Each participant provided electronic consent to the protocol approved by the Research Ethics Boards of the University of Toronto prior to participation.

A total of 21 participants were recruited, and one was excluded (see the "Participant Exclusion Criteria" section), leaving a final sample size of 20. This sample size is akin to what was used in both Yamanashi Leib et al. (2016, 2020). In addition, an a priori power analysis was conducted using G*Power 3 (Faul et al., 2007) to test the difference between two independent group means using a two-tailed test, a large effect size ($d = 0.80$), and

Figure 2
Predictions



Note. These graphs show potential outcomes for the magnitude of the correlation between participants' responses and the predicted sweetness ratings for Experiments 1 through 4. (A) The pattern of results that would occur if participants sampled only one item from an array. This pattern would show that observers are unable to integrate information from multiple items when making their ensemble judgments (random sampling one item). (B) The pattern of results would occur if participants successfully integrated all six items in the array. Here, the correlation would increase at larger set sizes as more information becomes available to participants. This pattern would show that observers could use all the information available to them when making their ensemble judgments (100% item integration).

an alpha of .05. The result showed that a total sample of 15 participants was required to achieve a power of 0.80 (note that this sample size justification was done with the planned comparison *t* tests in mind; see the "Results and Discussion" section). Based on this power analysis, our sample sizes throughout are sufficiently powered to detect significant effects.

The mean age of the final sample was 27.74, with 13 female, six male, and one declining to answer. Sixteen participants were right-handed, two were left-handed, one was ambidextrous, and one declined to answer. All had normal or corrected-to-normal vision, with two wearing contact lenses, 12 wearing glasses, one declining to answer, and the rest with neither glasses nor contacts.

Participant Exclusion Criteria

To ensure that participants were engaged with the task and not just randomly clicking on the scale, we compared their responses in the subset size one condition to the predicted sweetness value of those single items (derived from the ratings of a separate group of participants in Pilot 6; see the online supplemental materials). Given the results of Pilot 6, it is reasonable to predict that participants would agree on the perceived sweetness of individual food items. To that end, we conducted a correlation between the participant ratings for the food items encountered in the subset size one condition in this experiment and the predicted sweetness ratings of those same items in Pilot 6. Participants who had a correlation below an *r* of .70 were excluded from the analysis. Using this criterion, one participant was excluded from analysis leaving us with a final sample size of 20.

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 (Peirce et al., 2019), which was the platform used to run the experiment. The experiment was coded using the Psychopy3 Experiment Builder (Peirce et al., 2019). Participants were only permitted to take part in the experiment using a desktop or laptop computer. Both Macs and Windows computers with various screen sizes were used. Due to the lack of control in online experiments, the observers' distance from the screen could not be reliably controlled. Participants were instructed to perform the experiment in a distraction-free environment, arms-distance from the screen, with their computer plugged in and the screen set to maximum brightness.

Stimuli and Procedure

Using the 150 images from Pilot 6 (see the online supplemental materials), we randomly drew six images without replacement, yielding 25 sets of images with six images per set. Each set was assigned a predicted sweetness rating, which was calculated by averaging the ratings of the six items within the set (using the data from Pilot 6). To ensure that no single item was representative of the predicted sweetness rating, images that were within 0.5 of the predicted sweetness rating were replaced with a different image. The ensembles' predicted sweetness ratings were normally distributed around a mean of 4.93. For the subset conditions, one, two, or four items were randomly drawn from the full set (with replacement).

The ensemble arrays were presented in a 3 × 2 grid in the middle of the screen on a gray background, and the location of each item was randomly determined within that grid (see Figure 3). Each stimulus was 0.30 × 0.225 times the screen's height. The images were separated horizontally and vertically by 0.05 times the screen's height, and a white fixation cross (0.04 × 0.04 times the screen's height) was presented in the middle of the screen. The rating scale was similar to the one used in Pilots 1, 2, and 6, except the granularity of the scale was set to 0.25 to allow participants to use whole values, half values, and quarter values, and the number of tick marks was increased to 21 (so each half value was now represented by a tick mark). The instructions above the scale read "On average, how sweet were those foods? Click on the rating scale to make your response. 0 = *not sweet at all*, 10 = *extremely sweet*." Participants were encouraged to use the full range of the scale.

Participants were instructed to make judgments about the average sweetness of groups of food items that could vary in size while maintaining fixation on the central cross (which was present 500 ms prior to the appearance of the ensemble and remained on the screen until the rating scale appeared). Participants were then shown either the full six-item ensemble or a subset of the ensemble (one, two, or four items) for 1 s, followed by a 500 ms delay in which only the fixation cross was visible. Each ensemble display was followed by a rating scale that participants used to indicate the average sweetness of the preceding ensemble by clicking on the appropriate value on the scale. The scale was present until a response was made. Each participant responded to all 25 ensembles in each possible set size condition, in random order, for a total of 100 trials.

Data Analysis

Analysis scripts and data are available online via Open Science Framework (Gillies et al., 2022). This study was not preregistered.

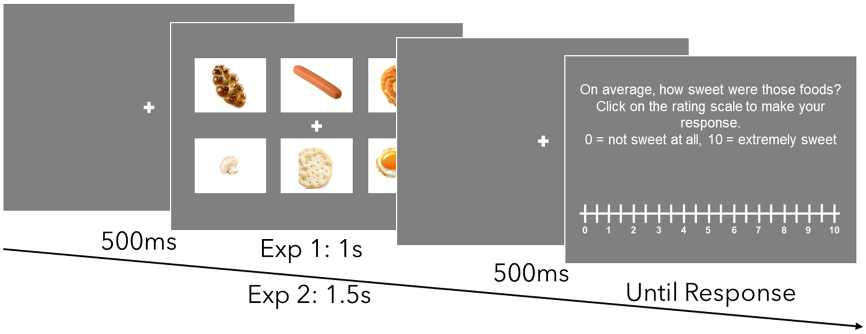
To examine item integration, we calculated the correlation between the participants' ratings and the predicted ratings for each subset condition. Individual Pearson correlations were converted to Fisher *z*-scores (normalizing the distribution of Pearson coefficients) and were then averaged across participants. We then examined the relationship between the magnitude of the correlation (Fisher *z*-scores) and the number of items shown to individuals using a linear regression.

Though a significant linear regression would suggest that observers are incorporating more information as it is made available to them (i.e., getting closer to the true average rating as more information is added), the linear regression was followed up with a series of three planned comparison paired sample *t* tests. To account for inflations to Type 1 error due to multiple comparisons, we used a Bonferroni corrected alpha value of .016 (i.e., for three comparisons). This series of comparisons allowed us to examine differences in the magnitude of the correlation between the different subset conditions (subset sizes one to two, two to four, and four to six), meaning that we could identify if there was a significant increase in the magnitude of the correlation (i.e., more information was incorporated) between all subset conditions, or if the magnitude of the correlation plateaued after a certain subset size, indicating that not all the available information was incorporated.

To examine what the upper limit of participants' performance would look like across the set sizes, we ran an ideal observer simulation. The simulation showed that the largest possible correlations for each subset size were .49 for set size one, .67 for set

Figure 3

Trial Sequence for Experiment 1 and 2


On average, how sweet were those foods? 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
Exp 1: 1s
Exp 2: 1.5s
500ms
Until Response

Note. This is an example of the full six-item ensemble condition. Participants viewed the images for 1 s for Experiment 1, and 1.5 s for Experiment 2, then made an average sweetness rating using a rating scale from 0 (*not sweet at all*) to 10 (*extremely sweet*). Food pictures are from the Food-Pics (bread, egg, hotdog, mushroom) (Blechert et al., 2019) and FreeFoodPhotos.com. The Food-Pics are adapted from "Food-Pics_Extended—An Image Database for Experimental Research on Eating and Appetite: Additional Images, Normative Ratings and an Updated Review," by J. Blechert, A. Lender, S. Polk, N. A. Busch, and K. Ohla, 2019, *Frontiers in Psychology, 10*, p. 307 (https://doi.org/10.3389/fpsyg.2019.00307). CC BY-NC. Exp 1 = Experiment 1; Exp 2 = Experiment 2. See the online article for the color version of this figure.

size two, .87 for set size four, and one for set size six. This shows that the magnitude of the correlation increases with set size if observers are able to use all the available information. To preview, this "ideal observer" pattern closely aligns with participants' performance in Experiments 3 and 4 (i.e., participants could incorporate all the available information).

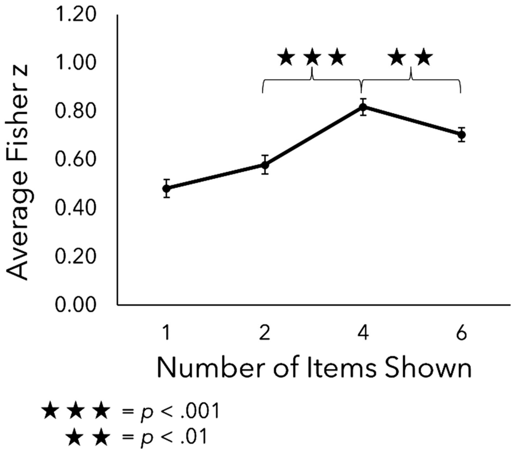
Results and Discussion

The average Fisher *z*-scores were fit by a linear regression, $r^2 = .2$, $p < .001$, showing that participants were incorporating more information as it was made available to them (see Figure 4). The planned comparison paired sample *t* tests revealed that there was no significant difference in average Fisher *z* between subset size one and two, $t(19) = 1.76$, $p = .09$, $d = 0.39$, 95% confidence interval (CI) [0.07, 0.84]. However, there was a significant increase from subset size two to four, $t(19) = 4.24$, $p < .001$, $d = 0.95$, 95% CI [0.41, 1.47]. This difference shows that observers were not merely sampling from a single item in the array. There was also a significant decrease in average Fisher *z* from subset size four to six, $t(19) = -2.87$, $p = .01$, $d = 0.64$, 95% CI [0.15, 1.12].

Overall, this pattern of results suggests that participants were unable to integrate information from more than four items in the display, meaning that, to some degree, participants were subsampling from the array when making their cross-modal ensemble judgments, rather than integrating information from all six items, as has been demonstrated previously (Yamanashi Leib et al., 2016, 2020). In other instances, subsampling strategies can be sufficient to produce a summary statistic that is still precise (Ji et al., 2018; Maule & Franklin, 2016). However, it is unclear if subsampling strategies occur because ensemble coding abilities are subject to some form of capacity limitation, or if this strategy is used when there is no utility in incorporating all the available information (e.g., subsampling can produce a relatively accurate summary statistic). The results of Experiment 1 suggest that it is the former, as the ensembles were designed in such a way that subsampling could not produce an accurate summary statistic. Although other studies looking at summary statistical representations of high-level or abstract features did not find that observers were engaging in subsampling (Haberman & Whitney, 2007, 2009; Yamanashi Leib et al., 2016, 2020), the number of items that can be integrated for a given statistical moment is dependent on a variety of factors. For example, the stimuli used or individual differences across participants can drive changes in item integration (Haberman et al., 2015; Whitney & Yamanashi

Figure 4

Results for Experiment 1



Note. The magnitude of the correlations increased with set-size but peaked at a maximum of four items. This provides evidence that observers were not merely sampling from a single item in the array, despite the limitation to the number of items they could integrate into their estimates of average sweetness. Error bars represent Morey's *SEM* (Morey, 2008). *SEM* = standard error of the mean.

Leib, 2018). While the results of this experiment demonstrate that participants can indeed form cross-modal ensemble percepts, there appears to be a limitation in the number of items that can be integrated into these percepts, which is not observed in other studies of high-level ensemble perception. In the next experiment, we investigate a potential explanation for this limitation.

Experiment 2

Experiment 2 was conducted to examine if additional viewing time would remove the capacity limit observed in Experiment 1. It is possible that computing average sweetness, which uses visual information to access and integrate multisensory representations of taste stored in LTM, is a complex process that requires additional processing time to complete. Indeed, a recent electroencephalogram study demonstrated that high-level ensemble processing (i.e., computing average facial identity) benefits from later-stage processing (Roberts et al., 2019). Alternatively, there may be a distinct limit as to how much information can be retrieved and integrated cross-modally from LTM at a given moment. If the former is true, we should observe a significant increase in average Fisher *z*-scores from subset size four to six. If, however, the latter is true, we should see average Fisher *z*-scores plateau at subset size four, as was observed in Experiment 1.

Method

Participants

Participants were recruited via Prolific (Prolific, 2021) using the same prescreening criteria and payment details as described in Experiment 1. Participants who completed the other experiments (including Pilot experiments) were not permitted to participate in this experiment.

A total of 27 participants were recruited, and seven were excluded (see the “Participant Exclusion Criteria in Experiment 1” section), leaving a final sample size of 20. The mean age of the final sample was 25.60, with 11 female and nine male. Eighteen participants were right-handed, one was left-handed, and one was ambidextrous. All participants had normal or corrected-to-normal vision, with 15 wearing glasses, two wearing contact lenses, and three with neither.

Apparatus

The apparatus was identical to that used in Experiment 1.

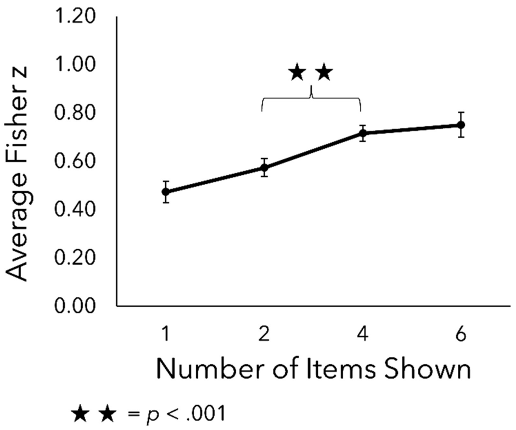
Stimuli and Procedure

The stimuli were identical to that of Experiment 1. The procedure was identical to that of Experiment 1, but the stimulus duration was 1.5 s rather than 1 s (see Figure 3). The logic for the change was simple. If people could integrate four items in 1 s, perhaps they could integrate six items with 1.5 times the viewing duration.

Results and Discussion

Using the same analysis as Experiment 1, we found that the average Fisher *z*-scores across subset size conditions were fit by a linear regression, $r^2 = .23$, $p < .001$, again showing that there was a positive relationship between the number of items shown to participants and how much information they were incorporating into their cross-modal ensemble percepts of average sweetness (see Figure 5). As in Experiment 1, there was no difference between subset size one and two, $t(19) = 1.79$, $p = .09$, $d = 0.40$, 95% CI [0.06, 0.85], and there was a significant increase in average Fisher *z* from subset size two to four, $t(19) = 3.08$, $p < .01$, $d = 0.70$, [0.20, 1.17]. Interestingly, the difference between subset size four and six was not significant, $t(19) = 0.60$, $p = .58$, $d = 0.13$, 95% CI [0.32, 0.56], showing that the increased viewing time did not enable observers to incorporate all six images into their cross-modal ensemble percepts, and replicating the four-item capacity limit observed in Experiment 1. This is unlike other forms of abstract ensemble coding where observers were able to integrate information from six items to extract summary statistics under brief viewing durations (e.g., 250 ms for object lifelikeness, 1 s for object value; Yamanashi Leib et al., 2016, 2020).

Figure 5
Results for Experiment 2



Note. The magnitude of the correlations increased with set-size but again peaked at a maximum of four items. Thus, even with the additional viewing time, observers were still limited by the number of items they could incorporate into their cross-modal ensemble percepts of average sweetness. Error bars represent Morey’s *SEM* (Morey, 2008). *SEM* = standard error of the mean.

Experiment 3

Given the consistent capacity limit of four items observed in Experiments 1 and 2, we examined if this ensemble coding limit can be explained by the nature of the information transformation required to compute average sweetness. Previous work has shown that observers can utilize semantic information during ensemble coding (Yamanashi Leib et al., 2020), demonstrating that ensemble processing can operate based on interactions between different cognitive systems (e.g., visual perception and semantic memory). However, previous demonstrations of this (where a capacity limit was not observed) were either unisensory in nature (i.e., using visual features to retrieve stored knowledge of lifelikeness; Yamanashi Leib et al., 2016), or used visual features to retrieve semantic information that is not tied to a particular sense (i.e., economic value is not strongly related to the visual features of an object; Yamanashi Leib et al., 2020).

The question of whether computing average taste is too computationally complex can be addressed by using the same type of stimuli but asking observers to compute a summary statistic for information that does not require the cross-modal retrieval of information from LTM (e.g., the size of an object on a computer screen). To investigate this, we had participants rate the average screen size (i.e., retinal-image size, not real-world size) of groups of food pictures using the same methods as Experiments 1 and 2.

Method

Participants

Participants were recruited via Prolific (Prolific, 2021) using the same prescreening criteria and payment details as described in previous experiments. Participants who completed the previous experiments were not permitted to participate in this experiment.

A total of 21 participants were recruited, and one was excluded (see the "Participant Exclusion Criteria in Experiment 1), leaving a final sample size of 20. The average age of the sample was 26.75 years. There were nine male and 11 female. Seventeen participants were right-handed, two were left-handed, and one was ambidextrous. All had normal or corrected-to-normal vision, with 13 wearing glasses, and the rest not wearing glasses or contacts.

Apparatus

The apparatus was identical to that used in previous experiments.

Stimuli and Procedure

This experiment used the 150 images (see Figure 6) generated from Pilot 7 (see the online supplemental materials). The ensembles were generated using the same method described in Experiment 1. The ensembles' predicted screen-size ratings were normally distributed around a mean of 4.92. For the subset conditions, one, two, or four items were randomly drawn from the full set (with replacement).

Three additional ensembles were created to use in practice trials. Eighteen images were taken from the Food-Pics database (Blechert et al., 2019) and one from the FoodCast research image database (Foroni et al., 2013). Of the Food-Pics images, two were edited in Photoshop to ensure only a single food item was present, and two were edited in Photoshop to change their size.

The rating scale was similar to the one used in previous experiments and the instructions above the scale read "What is the average screen-size of those foods? Click on the rating scale to make your response. 0 = *small*, 10 = *large*."

The procedure was similar to Experiments 1 and 2. However, participants were instructed to make judgments about the average screen size of groups of food pictures (see Figure 7A). Specifically, participants were explicitly instructed to judge the size of the food pictures relative to the white box they appeared in. In addition, participants were given an optional 1-min break every 50 trials. At the end of each block, participants were reminded of how to use the rating scale (see Figure 7B).

To ensure participants could use the scale, each participant performed a block of 12 practice trials prior to starting the experimental trials. Using the three practice ensembles, the practice trials proceeded the same way as the experimental trials. The data from the practice trials were not analyzed.

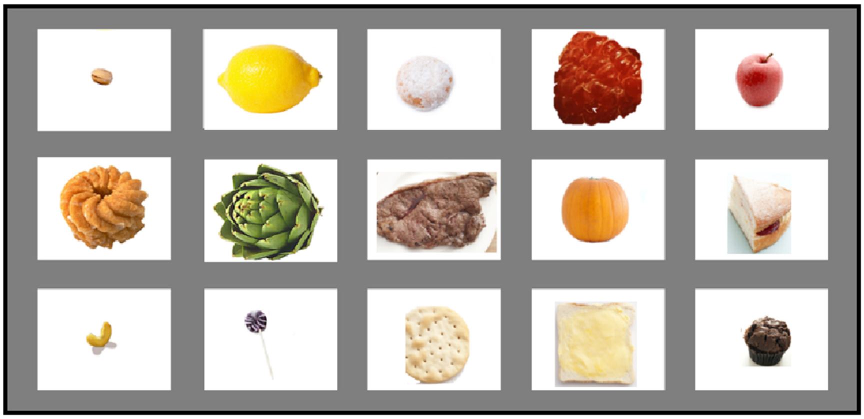
Results and Discussion

Using the same analyses described previously, we found that the average Fisher z-scores across subset size conditions were well fit by a linear regression, $r^2 = .47$, $p < .001$, showing that the correlation between participant ratings and the predicted ratings increased with set size (see Figure 8).

There was no significant difference in average Fisher z between subset size one and two, $t(19) = 1.54$, $p = .14$, $d = 0.34$, 95% CI

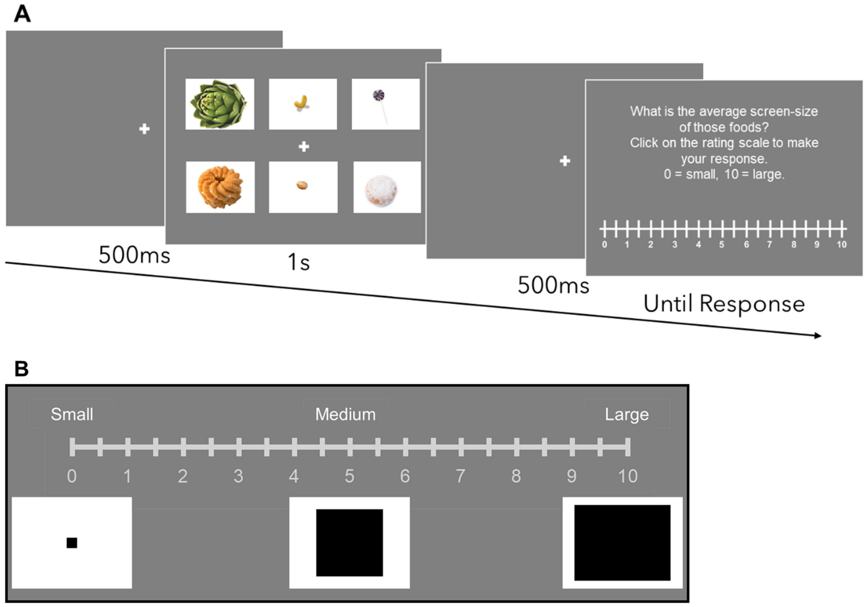
Figure 6

Sample Stimuli Used in the Average Screen-Size Experiment



Note. An example of the food pictures used in Experiment 3. Pictures are from the Food-Pics (Blechert et al., 2019) database (pistachio, lemon, donut, artichoke, cashew, lollipop). The rest of the food pictures are freely available illustrative examples of the stimuli (https://www.freefoodphotos.com) but are not images used in the actual study. The Food-Pics are adapted from "Food-Pics_Extended—An Image Database for Experimental Research on Eating and Appetite: Additional Images, Normative Ratings and an Updated Review," by J. Blechert, A. Lender, S. Polk, N. A. Busch, and K. Ohla, 2019, *Frontiers in Psychology*, *10*, p. 307 (https://doi.org/10.3389/fpsyg.2019.00307). CC BY-NC. See the online article for the color version of this figure.

Figure 7
Trial Sequence for Experiment 3


A
+
+
+
What is the average screen-size of those foods? Click on the rating scale to make your response. 0 = small, 10 = large.
0 1 2 3 4 5 6 7 8 9 10
500ms
1s
500ms
Until Response
B
Small
Medium
Large
0 1 2 3 4 5 6 7 8 9 10

Note. (A) This is an example of the full six-item ensemble condition. Participants viewed the images for 1 s and then made an average screen-size rating using a rating scale from 0 (*small*) to 10 (*large*). Food pictures are from the Food-Pics (pistachio, donut, artichoke, cashew, lollipop) (Blechert et al., 2019) and FreeFoodPhotos.com. (B) The infographic was shown to participants to instruct them how to use the scale. The Food-Pics are adapted from "Food-Pics_Extended—An Image Database for Experimental Research on Eating and Appetite: Additional Images, Normative Ratings and an Updated Review," by J. Blechert, A. Lender, S. Polk, N. A. Busch, and K. Ohla, 2019, *Frontiers in Psychology*, *10*, p. 307 (https://doi.org/10.3389/fpsyg.2019.00307). CC BY-NC. See the online article for the color version of this figure.

[0.11, 0.79]. In contrast, there was a significant increase in average Fisher z from subset size two to four, $t(19) = 4.11$, $p < .001$, $d = 0.91$, 95% CI [0.38, 1.44], and importantly, from subset size four to six, $t(19) = 3.87$, $p = .001$, $d = 0.87$, [0.34, 1.37]. Unlike the results of Experiments 1 and 2, which also used simultaneous presentation of stimuli, we found that observers could incorporate all the available information from the display to make their ratings of average screen size and were thus not subject to the capacity limitation observed in previous experiments. This suggests that the capacity limit we observed in Experiments 1 and 2 was likely driven by the requirement to access cross-modal representations stored in LTM when computing average sweetness.

In addition, to the best of our knowledge, this is the first experiment of its kind that shows that observers can generate a summary statistic for screen size even when the images used are of different shapes, highlighting the robustness and flexibility of ensemble coding.

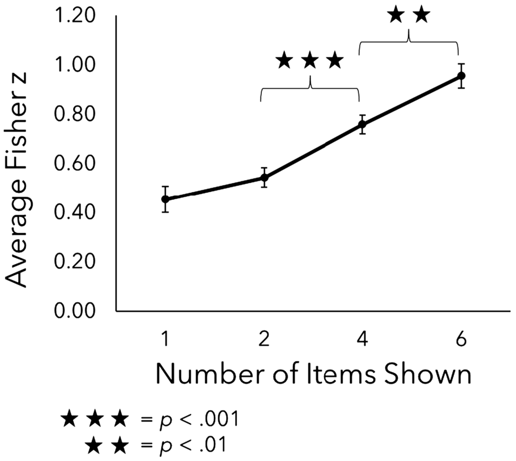
Experiment 4

The results of Experiment 3 suggest that the capacity limit observed in Experiments 1 and 2 may be driven by the requirement to retrieve information cross-modally from LTM. We posit that, due to the complexity of this computation, spatial attention may be a limiting factor when attempting to ascertain the average taste of visually presented information. This limitation may emerge in LTM or in VWM, or perhaps earlier. That is, the limitation may emerge as a failure of parallel visual processing of multiple features across many food items, as there are no visual features that entirely betray sweetness. As such, parallel processing and/or spatial attentional mechanisms encounter difficulty pooling sensory signals across the display, and observers would be forced to use a subsampling strategy. In this case, it may be that spatial attention and/or parallel processing constrains the number of functional visual cues an observer can integrate at a time to trigger the cross-modal retrieval required to compute an average taste, preventing observers from using all available stimuli under simultaneous viewing conditions. One way to circumvent this issue is to use a sequential display.

Some research has demonstrated that the integration of multiple objects tends to be more efficient when individual items are displayed in sequence rather than simultaneously (Florey et al., 2017; Gorea et al., 2014). One possible explanation is that there are limits to the distribution of spatial attention when viewing multiple items at the same time (Chong & Treisman, 2005; Florey et al., 2017). When viewing those same items in sequence, limits to distributing spatial attention no longer apply. If spatial attention (or parallel processing) is the limiting factor preventing people from incorporating all the available information into their cross-modal ensemble percepts under simultaneous viewing conditions, particularly for a computationally complex ensemble metric such as average sweetness, then using a sequential presentation approach should circumvent this issue.

Figure 8

Results for Experiment 3



Note. Observers were able to incorporate all the available information into their ensemble percepts for average screen size (i.e., they were not limited by item capacity). Error bars represent Morey's *SEM* (Morey, 2008). *SEM* = standard error of the mean.

Method

Participants

Participants were recruited via Prolific (Prolific, 2021) using the same prescreening criteria and payment details as described in Experiment 1. Participants who completed the other experiments were not permitted to participate in this experiment.

A total of 21 participants were recruited, and one was excluded from further analysis (see the "Participant Exclusion Criteria in Experiment 1" section), leaving a final sample size of 20. The mean age of the final sample was 25.55, with 11 female and 10 male. Fifteen participants were right-handed, and five were left-handed. All participants had normal or corrected-to-normal vision, with seven wearing glasses, six wearing contacts, and the rest with neither.

Apparatus

The apparatus was identical to the previous experiments.

Stimuli and Procedure

The stimuli were identical to that of the previous experiments, with the exception that stimuli were presented one-at-a-time in the middle of the screen (with a horizontal and vertical jitter up to 0.25 the screen's height), instead of simultaneously in a 3 × 2 grid.

The procedure was similar to the previous experiments, but participants were instructed to make judgments about the average sweetness of different foods presented one-at-a-time in sequence (see Figure 9). Participants were instructed to keep their eyes on a central fixation cross, which was present for 500 ms at the beginning of each trial, and then disappeared when the food stimuli appeared. Participants were then presented with the same ensembles from Experiments 1 and 2, sequentially viewing either the full six-item ensemble or a subset of that ensemble (one, two, or four items). In the full six-item condition, each item was displayed for 250 ms (1,500 ms of total viewing time). In the subset conditions, participants viewed each item for a longer duration to equalize the total stimulus duration (375 ms for subset size four, 750 ms for subset size two, and 1,500 ms for subset size one). In all conditions, the interstimulus interval was 100 ms. Following a 500 ms delay after the last item was presented, the rating scale appeared. The rating scale was present until a response was made. Each participant responded to all 25 ensembles at each possible set-size condition, in random order, for a total of 100 trials.

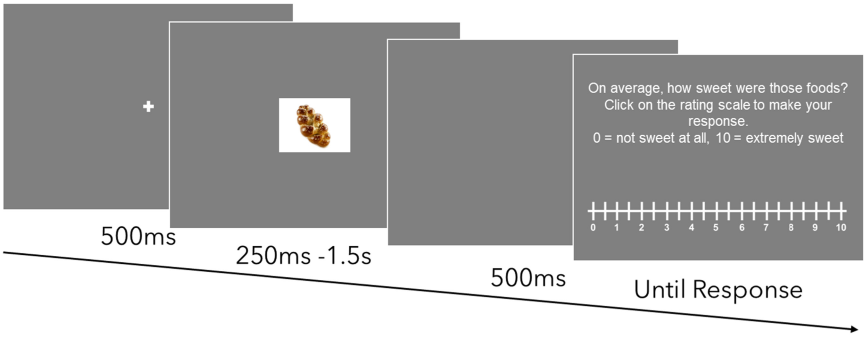
Results and Discussion

Using the same set of analyses described previously, we found that the positive relationship between the number of items shown to observers and the average Fisher z was stronger than in Experiments 1 and 2 (see Figure 10), $r^2 = .45$, $p < .001$. Indeed, there were significant increases in the average Fisher z between all the subset conditions, $t(19) = 3.50$, $p < .01$, $d = 0.77$, 95% CI [0.26, 1.27] for subset size one to two, $t(19) = 3.46$, $p < .01$, $d = 0.77$, [0.26, 1.27] for subset size two to four; and $t(19) = 2.70$, $p = .01$, $d = 0.60$, [0.12, 1.08] for subset size four to six. When equating for the viewing time from Experiment 2 but moving to a sequential display, observers were able to include all available information into their cross-modal ensemble percepts. In other words, the item capacity bottleneck seen in Experiment 2 was bypassed by utilizing a sequential display. Furthermore, the lack of an item capacity limit in this experiment suggests that the capacity limit observed in Experiments 1 and 2 might be explained by limitations in the deployment of spatial attention or parallel processing due to the complexity of computing average sweetness (e.g., observers were constrained by spatial attention when integrating useful visual cues to trigger cross-modal retrieval of taste information from LTM retrieval).

General Discussion

Across three experiments, we observed that participants had a limited ability to form cross-modal ensemble percepts. Namely, under simultaneous viewing conditions, observers were limited in the number of items they could incorporate into their percepts of average sweetness (i.e., a capacity limit of four food items). This four-item capacity limit persisted when viewing time was increased from 1 s (Experiment 1) to 1.5 s (Experiment 2). The results of Experiment 3 showed that changing the ensemble coding requirement to reflect a lower-level feature of the food stimuli, while maintaining simultaneous viewing, enabled participants to use all the available information to report average size. This finding suggests that it is the requirement to transform visual information into taste information that drives the capacity limit observed in Experiments 1 and 2. The item capacity limit for taste ensembles was removed by utilizing a sequential viewing condition (Experiment 4), suggesting that spatial attention is a limiting factor when needing to compute the average taste of visually presented information but is not when computing lower-level ensemble statistics (i.e., average size). In summary, under simultaneous viewing conditions, observers were unable to use all available visual information to cue knowledge of taste stored in LTM to form cross-modal ensemble percepts of

Figure 9
Trial Sequence for Experiment 4


+
On average, how sweet were those foods?
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
250ms -1.5s
500ms
Until Response

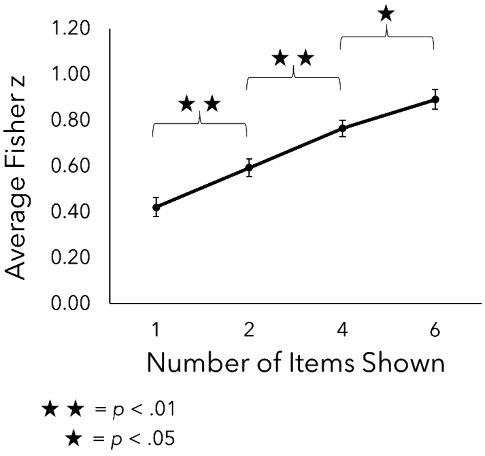
Note. This is an example of the subset size one condition (where the image would be viewed for 1.5 s). For subset sizes 2, 4, and 6, there was a 100-ms interstimulus interval. Participants viewed the images one-at-a-time in sequence at varying presentation times and then made an average sweetness rating using a rating scale from 0 (*not sweet at all*) to 10 (*extremely sweet*). Food pictures are from the Food-Pics (Blechert et al., 2019). The Food-Pics are adapted from "Food-Pics_Extended—An Image Database for Experimental Research on Eating and Appetite: Additional Images, Normative Ratings and an Updated Review," by J. Blechert, A. Lender, S. Polk, N. A. Busch, and K. Ohla, 2019, *Frontiers in Psychology*, *10*, p. 307 (https://doi.org/10.3389/fpsyg.2019.00307). CC BY-NC. See the online article for the color version of this figure.

average sweetness. Below, we explore the possible factors that may underlie this capacity limitation.

Is the Capacity Limitation Due to Limited Encoding Time?

Increasing the viewing time in Experiment 2 to 1.5 s did not remove the capacity limit. One possibility for this persistent capacity limit is that observers did not have enough time to sufficiently encode all the available information. We argue that based on prior literature where observers needed to extract summary statistics from complex objects (Yamanashi Leib et al., 2016, 2020), even 1 s is a sufficient amount of encoding time to reveal the integration of six items into a high-level or abstract ensemble representation. However, using a sequential viewing paradigm with 250 ms of viewing time per item in the set size six condition (equivalent to 1.5 s of simultaneous viewing time) did remove the capacity limit in Experiment 4. When viewing time was matched between Experiment 2 (simultaneous display) and Experiment 4 (sequential display), the capacity limit was removed in Experiment 4. Given this, we do not believe that insufficient encoding time explains the capacity limits observed in Experiments 1 and 2.

Figure 10
Results for Experiment 4



Note. Using a sequential display enabled observers to incorporate all available information into their cross-modal ensemble percepts of average sweetness. Error bars represent Morey's *SEM* (Morey, 2008). *SEM* = standard error of the mean.

Are Observers Using VWM?

The first impulse when faced with an item capacity limitation is to consider the role of VWM, whereby observers are using item-level information to inform their ensemble judgments. Indeed, the item capacity limits observed in Experiments 1 and 2 could just be reflective of a VWM capacity limit, which are also limited to about four items (Luck & Vogel, 1997). Experiments 1 and 2 used 1 and 1.5 s of viewing time, respectively, which is enough time to make several saccades (Cutsuridis, 2009; Wolfe, 2021). As such, participants could have serially fixated on four items contained within the displays and integrated their perceived taste values into a judgment of average sweetness. In other words, ratings of average taste could have been formed based on the combination of four (or less) specific items from the ensemble display held in VWM.

VWM could also potentially explain the removal of the four-item capacity limit in Experiment 4, where items were presented sequentially Although maintaining six images concurrently is outside the traditional capacity limit for VWM, observers could have accomplished the task by continuously updating their representation of "average sweetness" with each presentation of a new image. Indeed, the shortest image presentation time in Experiment 4 was 250 ms, which is enough time to encode the item into VWM. If observers encoded each item into VWM sequentially, they could calculate average sweetness "on the fly," updating their average rating with the presentation of each piece of new information, and then "dropping" the image from the contents of VWM once it has been used. If this were the case, then observers would likely not have explicit memory for all the items they nonetheless used to generate an estimate of average sweetness, even though they utilized their VWM to accomplish the task. If this is the case, it suggests that spatial attention is not a limiting factor in the formation of cross-modal ensemble percepts, per se. However, this "continual VWM updating" strategy cannot explain the results of other studies that used sequential presentation (e.g., Haberman & Whitney, 2009; Yamanashi Leib et al., 2016; Whitney & Yamanashi Leib, 2018), as images in these studies were presented too rapidly to enable individual object identification, despite presentation times being adequate to generate accurate summary statistics (Whitney & Yamanashi Leib, 2018). An open question with our results is whether there is an item capacity limit for average sweetness computations when using a sequential display. We do not think it is likely that VWM can be updated in perpetuity, but exactly how many items can be integrated into sequence is currently unknown and warrants further investigation.

Importantly, there is still a question as to whether the presence of subsampling strategies or capacity limits in ensemble coding directly indicates the involvement of VWM. In some instances, subsampling strategies can be used to accurately compute summary statistics (Ji et al., 2018; Maule & Franklin, 2016). For example, Maule and Franklin (2016) observed that a subsampling strategy can be used to accurately compute the average color of groups of circles. Specifically, they found that a model that randomly subsampled two out of 16 available circles produced results of equivalent accuracy to that of the observers. However, studies that have found capacity limitations or that participants were engaging in subsampling strategies did not investigate the possible involvement of VWM. As such, based on these studies it is difficult to ascertain the nature of the relationship between VWM processing and ensemble perception. To do so, one could utilize a member-identification paradigm (e.g., Ariely, 2001; Yamanashi Leib et al., 2016, 2020) to investigate if observers have explicit memory for the individual items within an ensemble.

Importantly, we do not want to make the claim that the involvement of VWM implies that observers are not engaging in ensemble coding at all. Recent evidence suggests that VWM and ensemble coding can interact, in that the contents of visual features held in VWM have a persistent influence on subsequent perceptual averaging tasks (Williams et al., 2021). Clearly, the degree to which the contents of VWM are used to generate ensemble percepts is an intriguing question, and future research should investigate the boundary conditions under which these cognitive processes interact, or whether they operate independently in most scenarios. In summary, we have evidence to suggest that observers have a limited ability to transform visual information rapidly and automatically into taste information when engaging in ensemble coding. Specifically, under simultaneous viewing conditions, observers are limited in the number of items they can incorporate into their ensemble percepts. This could be because, for this type of cross-modal ensemble computation, item-specific information held in VWM is necessary to generate the ensemble percept and thus this process is subject to standard capacity limitations inherent in VWM, which may or may not interact with processes controlling the deployment of spatial attention. Under sequential viewing conditions, VWM could still be at play whereby observers continually update their VWM.

Future studies could further explore capacity limitations in ensemble processing by using an inefficient observer model. For example, a model where performance is limited by sample size, early noise associated with estimating feature values (e.g., sweetness) from single items, and late noise associated with the process of integrating single-item estimates into an average value could give a quantitative estimate of the number of items individuals use to make reliable average sweetness judgments.

Is This Limit Specific to Cross-Modality?

Very little work has been conducted on exactly how much information can be retrieved in a single moment (e.g., Fukuda & Woodman, 2017), and much of LTM research involves successful item-level encoding, which does not need to occur for ensemble coding (Whitney & Yamanashi Leib, 2018). Some researchers have argued that memories cannot be retrieved in parallel, at least for episodic memories (Orscheschek et al., 2019). The results of previous work on summary statistical representations of economic value (Yamanashi Leib et al., 2020) provide some evidence to the contrary, as participants were able to perform tasks that required the rapid retrieval of some form of information about multiple items from LTM.

It is unclear why the transformation of visual information into taste information cannot be done rapidly and automatically. Other studies have shown that rapid transformation of information (requiring some sort of retrieval from LTM) is possible (Yamanashi Leib et al., 2020). One possibility is the cross-modal nature of the memory retrieval required to generate a summary statistic for taste. If cross-modal memory retrieval is the limiting factor, then this should extend to other cross-modal ensemble tasks. One intriguing possibility is examining interactions between vision and touch, which has yet to be explored in the context of ensemble coding. For example, can observers extract average weight (tactile information) from visually presented objects? To do this, observers will have to use visual information to cue knowledge of weight stored in LTM. If capacity limitations persist, this would suggest that it is the specific requirement to retrieve information cross-modally from LTM that is driving the capacity limit observed in Experiments 1 and 2. Conversely, if no capacity limit is observed when observers report the average weight of visually presented objects, this would suggest that there may be something specific to the cross-modal retrieval of taste information that drives the capacity limit.

As discussed previously, another possibility is that the bottleneck occurs before items are encoded into VWM or retrieved from LTM. No single visual feature (e.g., color) can be used to reliably give an estimate of average sweetness from a group of different food items. This raises the possibility that the four-item capacity limitation we observed in Experiments 1 and 2 is explained by a failure of parallel visual processing of multiple features across multiple food items, which could also explain why using a sequential display in Experiment 4 removed this capacity limit.

Conclusion

Our results indicate that observers have a limited ability to perceive the average sweetness of visually presented groups of food. Specifically, under simultaneous viewing conditions, observers were limited in the number of items they could incorporate into their cross-modal summary statistics. However, we found that changing the ensemble-coding requirement to reflect a lower-level feature enabled participants to use all the available information under simultaneous viewing conditions. These findings suggest that it is the requirement to transform visual information into taste information that drives the capacity limit observed under simultaneous viewing conditions. This capacity limit was removed when utilizing a sequential display, which suggests that spatial attention (or parallel processing), may be a limiting factor in the formation of cross-modal ensemble percepts. Specifically, spatial attention may constrain the number of visual cues an observer can integrate in a given moment to trigger cross-modal retrieval (something that is not necessary when engaging in a lower-level ensemble coding task). Moreover, the degree to which VWM is involved in this process is an intriguing avenue for future research. Taken together, our results thus demonstrate that there is a limit to the flexibility of ensemble coding, particularly when multiple cognitive systems must interact (e.g., visual perception, taste perception, and LTM) to compress sensory input into an ensemble representation.

References

Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2013). An almost general theory of mean size perception. *Vision Research*, *83*, 25–39. https://doi.org/10.1016/j.visres.2013.02.018

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131. https://doi.org/10.1016/j.tics.2011.01.003

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*(2), 157–162. https://doi.org/10.1111/1467-9280.00327

Baek, J., & Chong, S. C. (2020). Distributed attention model of perceptual averaging. *Attention, Perception, and Psychophysics*, *82*(1), 63–79. https://doi.org/10.3758/s13414-019-01827-z

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

Cant, J. S., Sun, S. Z., & Xu, Y. (2015). Distinct cognitive mechanisms involved in the processing of single objects and object ensembles. *Journal of Vision*, *15*(4), Article 12. https://doi.org/10.1167/15.4.12

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

Chong, S. C., & Treisman, A. (2005). Attentional spread in the statistical processing of visual displays. *Perception and Psychophysics*, *67*(1), 1–13. https://doi.org/10.3758/BF03195009

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the bandwidth of perceptual experience? *Trends in Cognitive Sciences*, *20*(5), 324–335. https://doi.org/10.1016/j.tics.2016.03.006

Cutsuridis, V. (2009). A cognitive model of saliency, attention, and picture scanning. *Cognitive Computation*, *1*(4), 292–299. https://doi.org/10.1007/s12559-009-9024-9

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Florey, J., Dakin, S. C., & Mareschal, I. (2017). Comparing averaging limits for social cues over space and time. *Journal of Vision*, *17*(9), Article 17. https://doi.org/10.1167/17.9.17

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

Fukuda, K., & Woodman, G. F. (2017). Visual working memory buffers information retrieved from visual long-term memory. *Proceedings of the National Academy of Sciences*, *114*(20), 5306–5311. https://doi.org/10.1073/pnas.1617874114

Gillies, G., Fukuda, K., & Cant, J. (2022, October 19). *Cross-modal ensemble data and analysis*. https://doi.org/10.17605/OSF.IO/GTMDB

Gorea, A., Belkoura, S., & Solomon, J. A. (2014). Summary statistics for size over space and time. *Journal of Vision*, *14*(9), Article 22. https://doi.org/10.1167/14.9.22

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General*, *144*(2), 432–446. https://doi.org/10.1037/xge0000053

Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision*, *9*(11), Article 1. https://doi.org/10.1167/9.11.1

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), R751–R753. https://doi.org/10.1016/j.cub.2007.06.039

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 718–734. https://doi.org/10.1037/a0013899

Ji, L., Chen, W., Loeys, T., & Pourtois, G. (2018). Ensemble representation for multiple facial expressions: Evidence for a capacity limited perceptual process. *Journal of Vision*, *18*(3), Article 17. https://doi.org/10.1167/18.3.17

Khayat, N., & Hochstein, S. (2018). Perceiving set mean and range: Automaticity and precision. *Journal of Vision*, *18*(9), Article 23. https://doi.org/10.1167/18.9.23

Khayat, N., & Hochstein, S. (2019). Relating categorization to set summary statistics perception. *Attention, Perception, and Psychophysics*, *81*(8), 2850–2872. https://doi.org/10.3758/s13414-019-01792-7

Kuriki, I. (2004). Testing the possibility of average-color perception from multi-colored patterns. *Optical Review*, *11*(4), 249–257. https://doi.org/10.1007/s10043-004-0249-2

Li, H., Ji, L., Tong, K., Ren, N., Chen, W., Liu, C. H., & Fu, X. (2016). Processing of individual items during ensemble coding of facial expressions. *Frontiers in Psychology*, *7*, Article 1332. https://doi.org/10.3389/fpsyg.2016.01332

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision*, *15*(4), Article 6. https://doi.org/10.1167/15.4.6

Maule, J., & Franklin, A. (2016). Accurate rapid averaging of multihue ensembles is due to a limited capacity subsampling mechanism. *Journal of the Optical Society of America A*, *33*(3), A22–A29. https://doi.org/10.1364/JOSAA.33.000A22

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*(2), 61–64. https://doi.org/10.20982/tqmp.04.2.p061

Orscheschek, F., Strobach, T., Schubert, T., & Rickard, T. (2019). Two retrievals from a single cue: A bottleneck persists across episodic and semantic memory. *Quarterly Journal of Experimental Psychology*, *72*(5), 1005–1028. https://doi.org/10.1177/1747021818776818

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*. https://www.prolific.co

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

Roberts, T., Cant, J. S., & Nestor, A. (2019). Elucidating the neural representation and the processing dynamics of face ensembles. *The Journal of Neuroscience*, *39*(39), 7737–7747. https://doi.org/10.1523/JNEUROSCI.0471-19.2019

Robitaille, N., & Harris, I. M. (2011). When more is less: Extraction of summary statistics benefits from larger sets. *Journal of Vision*, *11*(12), Article 18. https://doi.org/10.1167/11.12.18

Solomon, J. A. (2010). Visual discrimination of orientation statistics in crowded and uncrowded arrays. *Journal of Vision*, *10*(14), Article 19. https://doi.org/10.1167/10.14.19

Spence, C. (2015). On the psychological impact of food colour. *Flavour*, *4*(1), Article 21. 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

Webster, J., Kay, P., & Webster, M. A. (2014). Perceiving the average hue of color arrays. *Journal of the Optical Society of America A*, *31*(4), A283–A292. https://doi.org/10.1364/JOSAA.31.00A283

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology*, *69*(1), 105–129. https://doi.org/10.1146/annurev-psych-010416-044232

Williams, R. S., Pratt, J., Ferber, S., & Cant, J. S. (2021). Tuning the ensemble: Incidental skewing of the perceptual average through memory-driven selection. *Journal of Experimental Psychology: Human Perception and Performance*, *47*(5), 648–661. https://doi.org/10.1037/xhp0000907

Wolfe, J. M. (2021). Guided Search 6.0: An updated model of visual search. *Psychonomic Bulletin and Review*, *28*(4), 1060–1092. https://doi.org/10.3758/s13423-020-01859-9

Yamanashi Leib, A., Chang, K., Xia, Y., Peng, A., & Whitney, D. (2020). Fleeting impressions of economic value via summary statistical representations. *Journal of Experimental Psychology: General*, *149*(10), 1811–1822. https://doi.org/10.1037/xge0000745

Yamanashi Leib, A., Kosovicheva, A., & Whitney, D. (2016). Fast ensemble representations for abstract visual impressions. *Nature Communications*, *7*(1), Article 13186. https://doi.org/10.1038/ncomms13186

Received October 19, 2022
Revision received May 1, 2023
Accepted July 6, 2023 ■