

# Journal of Experimental Psychology: Human Perception and Performance

## **The Outlier Paradox: The Role of Iterative Ensemble Coding in Discounting Outliers**

Michael L. Epstein, Jake Quilty-Dunn, Eric Mandelbaum, and Tatiana A. Emmanouil

Online First Publication, August 6, 2020. <http://dx.doi.org/10.1037/xhp0000857>

### CITATION

Epstein, M. L., Quilty-Dunn, J., Mandelbaum, E., & Emmanouil, T. A. (2020, August 6). The Outlier Paradox: The Role of Iterative Ensemble Coding in Discounting Outliers. *Journal of Experimental Psychology: Human Perception and Performance*. Advance online publication. <http://dx.doi.org/10.1037/xhp0000857>

# The Outlier Paradox: The Role of Iterative Ensemble Coding in Discounting Outliers

Michael L. Epstein  
The City University of New York

Jake Quilty-Dunn  
Oxford University and Washington University in St. Louis

Eric Mandelbaum and Tatiana A. Emmanouil  
The City University of New York

Ensemble perception—the encoding of objects by their group properties—is known to be resistant to outlier noise. However, this resistance is somewhat paradoxical: how can the visual system determine which stimuli are outliers without already having derived statistical properties of the ensemble? A simple solution would be that ensemble perception is not a simple, one-step process; instead, outliers are detected through iterative computations that identify items with high deviance from the mean and reduce their weight in the representation over time. Here we tested this hypothesis. In Experiment 1, we found evidence that outliers are discounted from mean orientation judgments, extending previous results from ensemble face perception. In Experiment 2, we tested the timing of outlier rejection by having participants perform speeded judgments of sets with or without outliers. We observed significant increases in reaction time (RT) when outliers were present, but a decrease compared to no-outlier sets of matched range suggesting that range alone did not drive RTs. In Experiment 3 we tested the timing by which outlier noise reduces over time. We presented sets for variable exposure durations and found that noise decreases linearly over time. Altogether these results suggest that ensemble representations are optimized through iterative computations aimed at reducing noise. The finding that ensemble perception is an iterative process provides a useful framework for understanding contextual effects on ensemble perception.

### **Public Significance Statement**





Ensemble perception, the perception of objects by their group properties, is a mechanism by which the visual system may compress large amounts of information in visual scenes. In this study we examined how ensemble estimates discount outliers, which poses a paradox since outliers themselves are identified based on ensemble properties. Our proposed solution is that ensemble perception is an iterative process, identifying and rejecting outliers over time. Our results provide novel insights into how ensemble perception operates—rather than giving a single snapshot summary of complex visual scenes, ensemble perception provides a continuously unfolding, self-correcting summary.

**Keywords:** ensemble perception, outlier rejection, iterative processing, RT, visual masking

**Supplemental materials:** <http://dx.doi.org/10.1037/xhp0000857.supp>

Leaves on a tree, cars on the highway—the visual environment contains many groups of similar but distinct objects. These object

groups can be efficiently and effortlessly summarized by their statistical properties, such as the mean size of the leaves or the

 Michael L. Epstein, Program in Psychology, The Graduate Center, The City University of New York;  Jake Quilty-Dunn, Faculty of Philosophy, Oxford University, and Department of Philosophy and PNP Program, Washington University in St. Louis;  Eric Mandelbaum, Department of Philosophy, Baruch College, and Program in Philosophy, The Graduate Center, The City University of New York;  Tatiana A. Emmanouil, Department of Psychology, Baruch College, and Program in Psychology, The Graduate Center, The City University of New York.

Support for this project was provided by a PSC-CUNY Award (TRADA-49-634) jointly funded by The Professional Staff Congress and The City University of New York. This project additionally received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement 681422. Raw data are available at <https://osf.io/x8wqb/>.

Correspondence concerning this article should be addressed to Michael L. Epstein, Program in Psychology, The Graduate Center, The City University of New York, 365 5th Avenue, New York, NY 10016. E-mail: [mepstein@gradcenter.cuny.edu](mailto:mepstein@gradcenter.cuny.edu)

mean speed of the cars. A large body of literature shows that the visual system can economically represent object groups by extracting their statistical properties (Alvarez, 2011; Ariely, 2001; Whitney & Yamanashi Leib, 2018). This phenomenon, termed ensemble perception or statistical processing, has been demonstrated in several feature dimensions, such as size (Ariely, 2001; Chong & Treisman, 2003), speed (Watamaniuk & Duchon, 1992) orientation (Dakin & Watt, 1997; Hochstein, Pavlovskaya, Bonneh, & Soroker, 2018), color (Maule & Franklin, 2015), location (Alvarez & Oliva, 2009), as well as for complex properties such facial expression (Haberman & Whitney, 2009), identity (De Fockert & Wolfenstein, 2009), and lifelikeness (Leib, Kosovicheva, & Whitney, 2016). Most studies show encoding of the statistical mean, although recently more studies have shown that the range and variance of sets is also accurately represented (Hochstein et al., 2018; Khayat & Hochstein, 2018; Lau & Brady, 2018). Importantly, ensemble perception seems to proceed quickly and efficiently, yielding statistical summaries of large amounts of information that would otherwise overload our limited capacity attentional resources (Ariely, 2001; Chong & Treisman, 2003; Cohen, Dennett, & Kanwisher, 2016; Leib et al., 2016).

Several studies have suggested that individual objects' properties can be aggregated into an ensemble representation without being stored individually. Ariely (2001) showed that participants could accurately report the mean size of a set of circles even though they could not accurately recognize individual objects that were presented in the set. Subsequent studies extended these findings to sets of faces, showing that participants could judge the average emotion of the faces while being unable to detect emotional changes in individual faces (Haberman & Whitney, 2011). Ensemble perception has also been shown to be unimpaired by visual and verbal working memory load, supporting that the process is not subject to working memory limitations typically found for individual objects (Bauer, 2017; Epstein & Emmanouil, 2017). These results are consistent with models of ensemble perception as a rapid process that pools information across objects and does not require individual object encoding or binding of individual object features (Alvarez, 2011; Hochstein & Ahissar, 2002; Treisman, 2006).

At the same time, evidence exists that individual objects are discounted from mean representations when their properties deviate substantially from the group mean. Haberman and Whitney (2010) showed participants groups of 12 faces that varied in facial expression, two of which were emotional deviants. Participants were asked to adjust a face probe to the average expression of the group. The distribution of participants' responses indicated that they were filtering out the emotional deviants from their mean estimates. Furthermore, Haberman and Whitney showed that emotional deviants did not pop out, suggesting that the filtering of the outliers was not based on a conscious strategy of discounting items that strike subjects as "oddballs." Therefore, although objects are not individually encoded in ensemble perception, they can be filtered from the mean representation based on their individual properties.

The finding that outliers are filtered from ensembles suggests that ensemble processing is not always faithful to the raw information in the image but is rather biased against individual items that may not be part of the group or may be noisily represented. However, the mechanism by which outliers are identified and

filtered out remains a puzzle. Since outliers are defined by their deviation from the set mean, it follows that they must be excluded only after set statistics have already been computed. On the other hand, if ensemble statistics precede outlier detection, then it would be impossible that they filter out outliers. In the current study, we test one possible solution to this apparent paradox: the idea that ensemble perception is a continuously updating process and that outliers are filtered out through iterative ensemble computations.

We reasoned that if outlier noise is removed through iterative processing, then estimating the mean of ensembles should take longer when outliers are present. Moreover, estimates of the mean should vary over time, with estimates made under short exposures showing larger bias toward the outliers than estimates under long exposures, which afford more time to "clean" the outliers from the statistical summary via iterative steps of processing. These hypotheses were tested in three experiments. In Experiment 1, we verified that outliers are filtered out when processing mean orientation, in the same way they are discounted from estimates of mean facial expression. In Experiment 2, we tested whether ensemble estimates are slower when outliers are present and whether such reaction time (RT) differences are due to the increased range that sets with outliers typically exhibit. In Experiment 3 we varied exposure duration and observed the degree to which outliers are included in ensemble estimates as a function of time.

To anticipate the results: Consistent with previous findings, we observed that outliers are filtered out of average orientation estimates, even though they are not entirely excluded. These effects are not solely due to increased range, since outlier displays produce a different pattern of responses compared to displays of matched range. Finally, outliers are progressively filtered out over time, with early estimates showing a strong outlier bias and later estimates showing almost no outlier bias. Altogether, the results support the idea of ensemble perception as an iterative process with the capacity to continuously update and filter out noise.

## Experiment 1

In Experiment 1, we aimed to confirm that outliers are rejected when perceiving the average orientation of groups of lines. The purpose here was twofold: first, to extend previous reports of outlier rejection in groups of faces (Haberman & Whitney, 2010) to a basic, low-level feature of the visual environment: orientation. This extension provides a conceptual replication of Haberman and Whitney's results while providing evidence that outlier rejection is a general property of ensemble perception across multiple feature dimensions. The second purpose was to establish the basic design of Experiments 2 and 3, which explore the timing of outlier rejection in more detail.

As this was mostly a conceptual replication of Haberman and Whitney (2010), our methods were largely adapted from their study, with minor adjustments for our orientation stimuli. Participants viewed groups of lines with varied orientations and were instructed to indicate the mean orientation using an adjustable probe. For trials with outliers, this allowed us to compare participants' error to both the local mean (mean of only nonoutlier items) and the global mean (mean of all items, including outliers). If participants adjust their responses closer to local means, error will be lower for the local relative to the global measures, suggesting that they are reducing the weight of outliers in their

calculation of the average orientation. On the other hand, if error is found to be lower for the global relative to the local mean, it would suggest that participants are weighing the outliers evenly in their statistical summaries. Thus, by comparing the difference in orientation of the probe to the local and global mean, we can directly test the extent to which outliers are filtered during ensemble perception. We additionally included a condition where no outliers were present, allowing the comparison of error between this control condition and the outlier condition. This was used to test if error could be introduced by outliers, even if they were partially rejected.

This experiment was carried out in two groups using identical methods aside from a minor adjustment to how outliers were calculated for the second half of participants (see stimuli below). Repeated-measures ANOVAs with group as a between-subjects variable revealed no differences in performance between groups and no interactions for any of our comparisons (all  $ps > .142$ ), and so data were collapsed across groups for the purposes of this report.

## Method

**Participants.** A total of 26 participants (13 female, 2 left-handed, average age 21.15, range 18 to 26) were recruited from Baruch College's student subject pool for this experiment. All were tested and passed a vision test for normal or corrected-to-normal vision and provided informed consent. All participants received course credit for participation.

A priori power analysis was conducted in G\*Power on data collected from 5 pilot subjects (Faul, Erdfelder, Lang, & Buchner, 2007). Comparison between Von Mises fits to local and global error distributions (see below for more details on this procedure) showed an effect size, measured with Cohen's  $d$ , of .95. This effect size, with an alpha of .05, indicated 11 subjects would be sufficient to achieve .80 power. Due to extra participants recruited to account for no-shows, we slightly exceeded this target in each group of 13 participants, which are presented together in this experiment.

**Stimuli.** All stimuli were generated using PsychoPy (Peirce, 2007) and displayed on a Dell CRT monitor with a 75-Hz refresh rate. The screen was set by default to be gray (#A1A1A1) with a small white (#FFFFFF) fixation cross displayed in the center. The fixation cross was removed while the adjustable probe was present on the screen. Sets consisted of 12 white oriented lines ( $1.6^\circ$  in length, 5 pixels wide) displayed in a  $3 \times 4$  grid with the center of each item  $3^\circ$  apart. Item locations were jittered a small amount ( $<0.475^\circ$ ) each trial to ensure stimuli were not shown in the exact same position every trial.

Orientations for the 12 items within the set were randomly selected from four values evenly spread across a range of  $21.6^\circ$ . In outlier trials, two items within the set were chosen randomly and set to be  $72^\circ$  from the mean of the other items. For the first half of participants, outlier values were calculated based on the mean of the full set, before two items were removed to be replaced by outliers. This resulted in the outliers ranging from  $63.36^\circ$  to  $79.92^\circ$  from the mean of the other items. For the second half of participants, outlier values were calculated to always be exactly  $72^\circ$  from the mean of the nonoutlier items. The tilt direction for outliers was randomly selected on each trial to be to the left or to the right of the set mean. Our stimulus range was chosen to match the methods

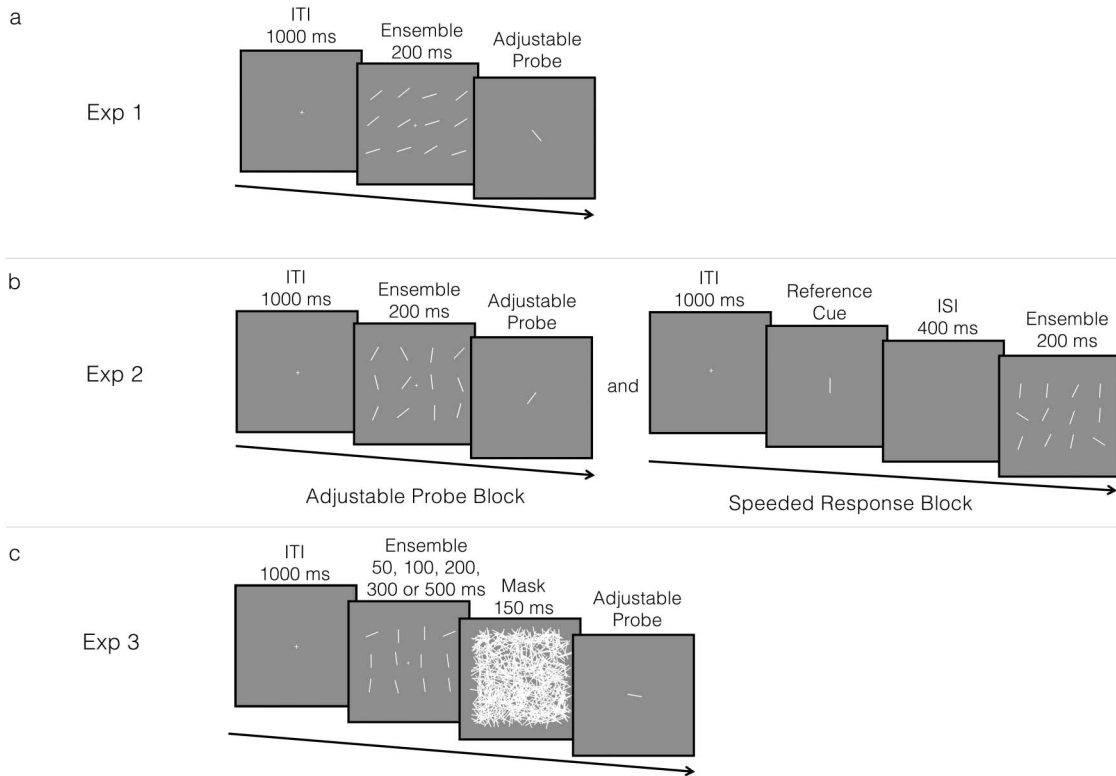
reported by Haberman and Whitney (2010). The face stimuli in their study comprised 150 possible units, item sets displayed contained faces within 18 units of the mean, and outliers were 60 units from the mean. We converted these values to 180 units for orientation by multiplying all values by 1.2, resulting in a range of  $21.6^\circ$  for our sets, and outlier values of  $72^\circ$ .

Finally, to prevent identical presentation across trials, the orientation of all lines within sets were adjusted an equal and randomly selected amount each trial. This range was controlled so that no item could tilt within  $10^\circ$  of horizontal. This ensured that the mean orientation was unique in every trial while maintaining set distribution and range. Adjustable probes were presented with a random initial orientation on every trial, to avoid introducing bias to response.

**Procedure.** Figure 1a shows an example of the task performed in Experiment 1. Participants were presented with a set of 12 oriented lines for 200 ms. After presentation of set items, participants were presented with an adjustable probe, consisting of a single line that could be tilted clockwise and counterclockwise using the arrow keys on the keyboard. Participants were instructed to use the probe to indicate what they perceived as the average tilt of all the lines within the group. The beginning orientation of the probe was random. Participants were given as much time as required to adjust the probe and pressed the space bar to confirm their response. After their response, a screen containing only the fixation cross was shown for a 1000-ms intertrial interval. On half the trials, stimuli orientations were adjusted so that two objects would exhibit distinct-enough tilts within the sets as to constitute outliers (outlier condition), while on the other half there were no outliers in the set (no-outlier condition). Participants were presented with 100 trials of each condition in random order. Every 50 trials, participants were given the opportunity to take a break.

**Analysis.** Error in the no-outlier trials was measured simply as the difference in degrees between the value participants indicated with the adjustable probe, and the mean of the items within the set. For outlier trials, global error was measured as the difference between participant response and the mean of all items within the set, and local error as the difference between participant response and the mean of only nonoutlier items within the set. Because orientation data is circular (i.e.,  $180^\circ = 0^\circ$ ), error was recorded as the smallest difference between response and target considering both possible directions. For example, if the target mean was  $30^\circ$  and a participant responded  $170^\circ$ , error was calculated as  $-40^\circ$  rather than  $140^\circ$ . These calculations produced distributions of error for each condition to which we fit Von Mises distributions after converting degrees to radians and converting to a  $2\pi$  radian space (Bays, Catalao, & Husain, 2009; Berens, 2009; Haberman & Whitney, 2010). The Von Mises distribution was appropriate to use in this case, as orientation is a circular set. First we extracted  $\kappa$ , a measure of precision, from the Von Mises distributions fit to the recorded error for each condition. These  $\kappa$  values were then converted to circular standard deviation, measured in radians, for the statistics and figures reported here. We performed these error analyses using all trials.

Probe adjustment time was measured from the time of stimulus onset. To avoid bias from extreme responses, trials with response times with a z-value greater than 2.5 were excluded for the probe adjustment time analyses (on average 2.35% of trials for the no-outlier condition and 2.5% of trials for the outlier condition).



*Figure 1.* Task examples. Paradigms for Experiments 1, 2 and 3 are shown here. Orientations and relative sizes are direct examples from each experiment, but overall size has been exaggerated for clarity. a) In Experiment 1, participants viewed ensembles where outliers were present or absent, and indicated their judgment of the average tilt using an adjustable probe. Here stimuli with no outlier present are shown. b) In Experiment 2, participants performed two blocks—an adjustable probe task, and a speeded response task. Each task contained an outlier, a no-outlier and a range control condition. Participants performed these two blocks in a counterbalanced order. Here range control stimuli are shown for the adjustable probe block, and outlier stimuli for the speeded response block. The outlier stimuli conflict with the direction of correct response in this example. c) In Experiment 3, participants performed a task similar to Experiment 1 with the addition of masking 50, 100, 200, 300, and 500 ms after stimulus onset. In this figure outlier stimuli are shown.

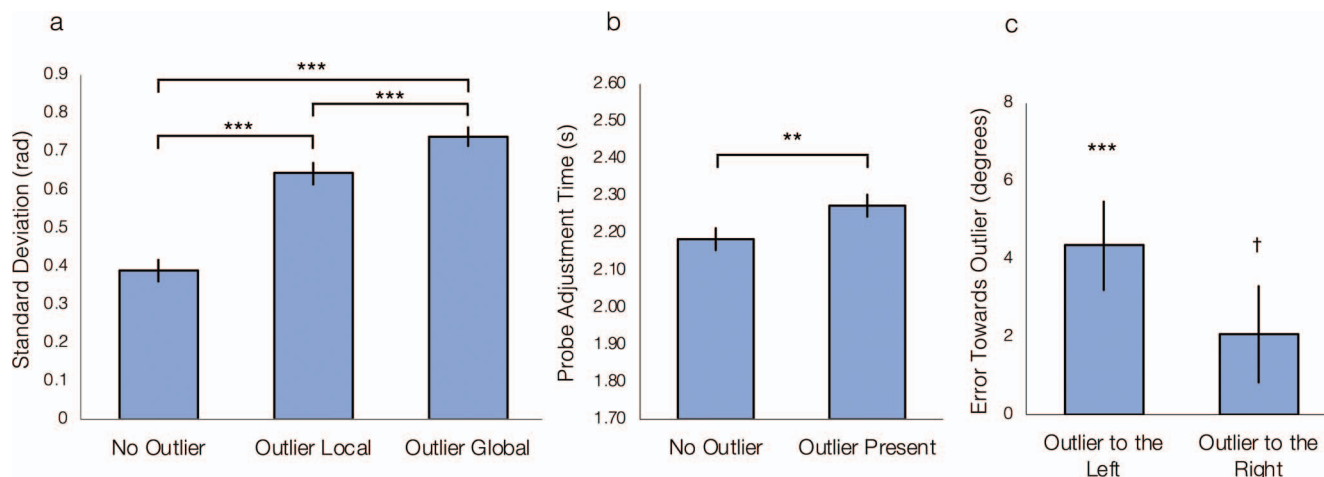
An additional analysis was carried out measuring local error with trials split for cases where the outlier was to the left or right of the mean. If error was unrelated to the direction of the outlier, (i.e., they erred in either direction), it would average out to zero. If, however, error is biased toward the direction of the outlier, we would see positive (when biased to the right) or negative (when biased to the left) values. Note that in figures we present the absolute error biased in the direction of the outlier for ease of visualization.

Comparisons between conditions were made using two-tailed paired  $t$  tests. Comparisons of the direction of the error were made using one-tailed one-sample  $t$  tests compared against a test value of 0. Effect sizes for all  $t$  tests are reported with Cohen's  $d$ . Where applicable, error bars shown have been corrected for within-subject designs using the Cousineau-Morey method, which involves normalizing the data to remove between-subjects differences and multiplying by a correction factor based on number of within-subject conditions (Cousineau, 2005; Cousineau & O'Brien, 2014; Morey, 2008). We additionally report descriptive statistics and tests of normality for our data in Table 1

in the online supplemental materials, as well as results of nonparametric tests in Supplemental Table 2, confirming results for any conditions that did not pass tests of normality.

## Results

See Figure 2 for the results of all conditions. Lower standard deviation was measured for local as compared to global error distributions ( $t(25) = -3.995, p < .001, d = -0.78$ ) indicating that participants reduced the weight of outliers when calculating average orientations. However, standard deviation was significantly lower for the no-outlier condition compared both to the local ( $t(25) = -9.22, p < .001, d = -1.81$ ) and global ( $t(25) = -14.55, p < .001, d = -2.85$ ) outlier conditions. Probe adjustment time was also lower in the no-outlier condition as compared to the outlier conditions,  $t(25) = -3.06, p = .005, d = -0.6$ . In the trial split tests, error to the local mean was significantly biased in the direction of the outlier for outliers to the left ( $t(25) = -3.82, p < .001, d = -.75$ ) and trending significance



*Figure 2.* Results for Experiment 1. a) Circular standard deviation for Von Mises distributions fit to participant error when no outliers were present, and to error distributions to the local and global mean in the outlier condition. Values are in radians. b) Probe adjustment time measured in seconds for conditions with and without outliers. c) Error biased toward outliers measured in degrees. Absolute values of the means are shown here. For a) and b), error bars represent standard error using the Cousineau-Morey method. For c), error bars represent between-subjects standard error. † Trending. \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ . See the online article for the color version of this figure.

to outliers to the right ( $t(25) = 1.67, p = .054, d = .33$ ) of the mean.

## Discussion

These results support the hypothesis that participants reduce the weight of outliers when estimating the average orientation of a group of lines. However, the improved accuracy in the no-outlier condition indicates that outlier presence does increase error overall, suggesting that outlier rejection is partial, rather than all-or-nothing. Interestingly, even though responses were not speeded, probe adjustment time was longer for outlier conditions. This finding could suggest that outliers filtering incurs a processing cost. Note, however, that the outlier displays also had an increased range compared to the no-outlier displays, which could have contributed to noise in the ensemble calculation and uncertainty in the response. In Experiment 2, we sought to replicate and extend the results of Experiment 1, using a more sensitive measure of RT, as well as a control condition to measure the effect that the increased stimulus range in the outlier displays may have on the results.

## Experiment 2

Following our observation of increased error and probe adjustment to outliers in Experiment 1, we sought in Experiment 2 to replicate and more directly measure these effects. Participants performed two tasks designed to measure the effects of outliers on accuracy and RT. In the adjustable probe task, they judged average tilt using an adjustable probe as in Experiment 1, with an emphasis on accuracy. In a separate speeded response task they judged the direction of the set tilt with respect to vertical as quickly and as accurately as possible. For each of these tasks, in addition to the no-outlier and outlier conditions, we added a third control condi-

tion, in which no outlier was present but the total range of the orientations within the group matched that of the outlier condition. As range has been shown to influence ensemble coding accuracy (Ji, Rossi, & Pourtois, 2018; Maule & Franklin, 2015; Utochkin & Tiurina, 2014), it is possible that displays with outliers are judged with greater error because of their increased range. However, if outliers are filtered out, this would effectively decrease set range, allowing outlier sets to be judged with reduced error compared to no-outlier displays with a similar range. Thus, by comparing error and RT between outlier and range conditions, we can determine if a unique process underlies ensemble coding when outliers are present. Altogether, this experiment will allow us to more rigorously test RT when outliers are present, as well as confirm that any effects of outliers on response error or RT are not solely due to the increased range in the set (Utochkin & Tiurina, 2014).

## Method

**Subjects.** Seventeen participants (11 female, 1 left-handed, average age 20.18, range 18–25) were recruited from Baruch's student subject pool for this experiment. All participants passed a vision test for normal or corrected-to-normal vision and provided informed consent. One participant exhibited low accuracy in the speeded response task (average accuracy = 36.11%) and was thus excluded from the analysis of both tasks. A primary effect of interest in this experiment was again the differences between fits to local and global error distributions, we utilized the same power analysis described for Experiment 1, indicating that 11 participants would be required to achieve .80 power. Extra participants recruited to account for potential no-shows resulted in a final sample size of 16.

**Stimuli.** Stimuli were identical in size, shape, and location as those described in Experiment 1. For the adjustable probe task, set orientation values were generated using similar procedures as to those described in the methods for Experiment 1, with orientations

randomly selected from four values, and two items randomly selected to be  $72^\circ$  from the mean of the other items for the outlier condition. Outlier direction relative to the mean was again randomized on each trial (while maintaining set range and item distribution), and set orientations adjusted an equal and randomly selected amount on each trial, with a restricted range so that no items fell within  $10^\circ$  of horizontal. Unlike Experiment 1, however, set properties such as range and relative distribution of items within sets were calculated prior to the experiment and reused for each participant. This was done to ensure that set ranges in the range control trials could match exactly the set ranges present in the outlier trials. Values for range control stimuli were determined by calculating the range of each outlier trial and creating a matching range set with 12 orientations evenly distributed across that range.

For the speeded response task, stimuli were generated using similar methods to that described for the adjustable probe task, but with the mean orientation for each set controlled to be  $15^\circ$  to the left or right of vertical. To ensure that the values were not identical to the adjustable probe task, new set values were calculated for each participant for the speeded response task. Range control sets were set to match the range of outlier sets, with items set to four evenly distributed values. All other set properties were identical to those shown in the adjustable probe task. Vertical cue stimuli used in the speeded response task were identical to set items in size and shape and were positioned centrally.

**Procedure.** Figure 1b shows an example of the adjustable probe and speeded response tasks used in Experiment 2. Participants completed each task in separate blocks, with three blocks per task. The order in which participants performed the tasks was counterbalanced.

In the adjustable probe task, methods were identical to those described for Experiment 1, except for the addition of a range control condition where no outliers were present but the overall range of orientations displayed by the stimuli matched that of the outlier condition. Participants saw a set of oriented lines for 200 ms and then provided their estimate of mean orientation using the adjustable probe, taking as much time as they needed. The initial orientation of the adjustable probe was randomly determined. Overall participants completed 180 trials (60 for each condition).

In the speeded response task, participants judged the orientation of a set with respect to vertical as quickly and as accurately as possible. Trials began with a vertical cue (400 ms) to prepare participants for the onset of stimuli and remind them of the vertical reference point. After a 400-ms blank screen, the ensemble displays were shown for 200 ms. Participants were instructed to indicate the direction the set was tilted on average using the left and right arrow keys on the keyboard. Importantly, the direction of the outlier (to the left or right of the average) was set to align with the direction of the average (left or right of vertical) 50% of the time so that responding to the outlier could not be used as a strategy. Participants were given unlimited time to respond but were instructed to respond as quickly and accurately as possible. After response, a screen containing only the fixation cross was shown for a 1000-ms intertrial interval. As in the adjustable probe task, participants completed 180 trials (60 per condition).

**Analysis.** Performance in the adjustable probe task was measured in the same way as in Experiment 1, by fitting Von Mises distributions to error for each condition and calculating circular

*SD*. All trials were included when fitting Von Mises distributions. For measurement of response time, probe adjustment times higher than 2.5 z-value were excluded for the adjustable probe task (no-outlier: 1.98% of trials on average; outlier: 1.67% of trials on average; range control: 2.08% of trials on average). Response times higher than 2.5 z-value were similarly excluded for the speeded response task (no-outlier: 2.19% of trials on average; outlier: 3.13% of trials on average; range control: 3.13% of trials on average). Accuracy within the speeded response task was recorded simply as the percentage of responses indicating the correct mean direction on the non-excluded trials. Response time for both tasks was recorded from stimulus onset. Reaction time measurements within the speeded response task are reported here with incorrect responses excluded, but statistics run with the incorrect responses included resulted in similar results for all tests.

## Results

**Adjustable probe task.** See Figure 3 for adjustable probe results. Error recorded in this experiment was similar to that of Experiment 1, with lower *SD* for the local mean as compared to the global mean,  $t(15) = -7.56, p < .001, d = -1.89$ . Again, *SD* was significantly lower for the no-outlier condition as compared to both local,  $t(15) = -5.21, p < .001, d = -1.30$ , and global,  $t(15) = -8.74, p < .001, d = -2.18$ , outlier errors. Additionally, *SD* was higher in the range condition compared to the no-outlier condition,  $t(15) = 11.62, p < .001, d = 2.90$ , as well as the local  $t(15) = 5.37, p < .001, d = 1.34$ , and global error distributions,  $t(15) = 2.28, p = .04, d = .57$ . These results replicate Experiment 1. Importantly, they also provide evidence that the error observed in outlier conditions cannot be explained solely by differences in range. On the contrary, the significantly higher error in the range condition compared to outlier conditions suggests that in outlier conditions, even if increased range is a factor on error, the influence of deviant stimuli can be reduced, effectively reducing the range and boosting accuracy.

Error toward the outlier was measured again with trials split for outliers in each direction. Bias toward outliers displayed to the right of sets trended significance,  $t(15) = 1.65, p = .06, d = .41$ , and did not reach significance for outliers to the left of sets,  $t(15) = -1.06, p = .15, d = -.26$ , presumably due to lower power as a result of fewer trials as compared to Experiment 1. Probe adjustment time within the adjustable probe task was similar for all conditions; no tests reached significance.

**Speeded response task.** See Figure 4 for speeded response task results. Accuracy in the speeded response task was highest for the no-outlier condition compared to both the outlier,  $t(15) = 5.19, p < .001, d = 1.30$ , and range control,  $t(15) = 10.69, p < .001, d = 2.67$ , conditions. Notably, the range condition was by far the least accurate (outlier vs. range:  $t(15) = 7.26, p < .001, d = 1.82$ ), again supporting the hypothesis that error in outlier conditions is not due simply to increased range in the stimuli. Consistent with the accuracy results, RT was fastest in the no-outlier condition, with significant differences measured both compared to the outlier,  $t(15) = -4.22, p < .001, d = -1.05$ , and range condition,  $t(15) = -3, p = .009, d = -.75$ . Responses to the range condition were also significantly slower than the outlier condition,  $t(15) = 2.26, p = .039, d = .57$ .

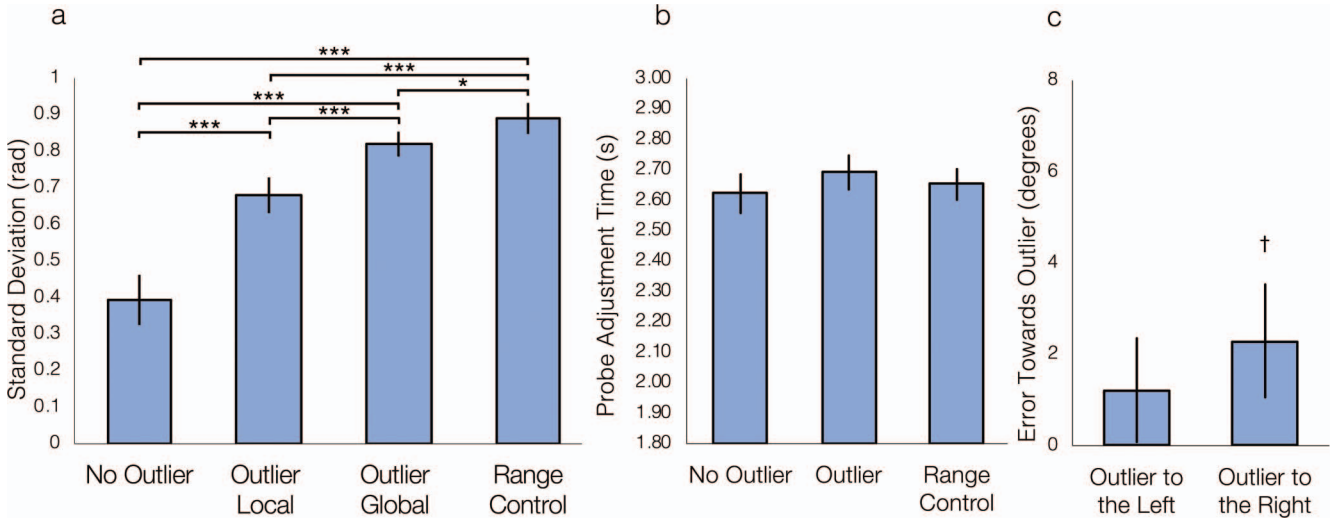


Figure 3. Adjustable probe task results. a) Circular standard deviations measured for Von Mises distributions fit to participant error for all conditions. Results displayed in radians. b) Probe adjustment time results in seconds. c) Error biased in the direction of the outlier when the outlier was to the left or to the right of the local set mean. Absolute values of the averages are shown here. For a) and b), error bars represent standard error using the Cousineau-Morey method. For c), error bars represent between-subjects standard error. † Trending. \*  $p > .05$ . \*\*  $p > .01$ . \*\*\*  $p < .001$ . See the online article for the color version of this figure.

**Discussion**

In Experiment 2, we controlled for range between the outlier and no-outlier displays in order to examine whether any effects of the outlier condition could be due to increased range. Indeed, range has been previously shown to influence ensemble perception accuracy independent of outliers (Ji et al., 2018; Maule & Franklin, 2015; Utochkin & Tiurina, 2014). We found evidence that the outlier condition produced a different pattern of RT and accuracy results compared to the matched range no-outlier condition. Specifically, the outlier condition produced decreased errors and faster RTs than the

matched range control. One natural way to interpret this finding is that discounting outliers effectively reduces the range, which is then based on remaining items, therefore decreasing the cost of range on RT and accuracy. What remains unclear is whether outlier filtering was uniquely initiated in the outlier condition compared to the matched range control, and which specific factors prompted the outlier filtering process. First, it is possible that outlier filtering was attempted in both conditions based on increased range, but only succeeded in the outlier condition, contributing to faster RTs. However, it is also possible that outlier filtering only took place in the outlier condition, in which

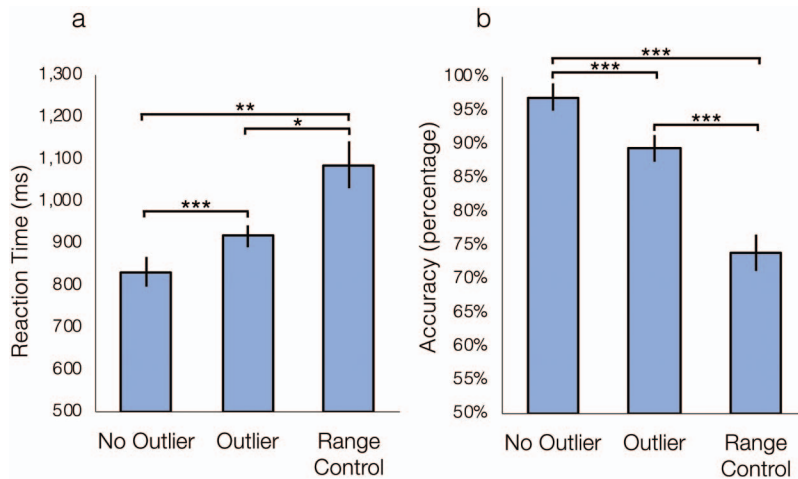


Figure 4. Results for the speeded response task. a) Reaction time for each condition as measured from stimulus onset. b) Percentage of correct responses for each condition. Error bars represent standard error using the Cousineau-Morey method. \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ . See the online article for the color version of this figure.



outliers were tagged by virtue of their separation from the rest of the items in the population code. We will return to this possibility in the General Discussion.

Experiment 2 therefore replicates and extends the results of Experiment 1, suggesting that the differences observed in Experiment 1 were not solely due to range differences. The results also support the hypothesis that outlier rejection is automatic, as suggested by Haberman and Whitney (2010), or at least intuitively preferable, as participants worked to reject outliers even when pressured to respond as fast as possible, despite not being explicitly instructed to do so.

### Experiment 3

The results of Experiments 1 and 2 provide evidence that outliers can be successfully, if perhaps only partially, excluded from statistical summaries, and that this process influences RT. While this suggests that outlier rejection requires additional steps of processing, a critical remaining question is the precise timing by which outliers are filtered during ensemble perception. Here we test this using visual masking at 50, 100, 200, 300, and 500 ms after stimulus onset, to test how ensemble perception in the presence of outliers changes over time. We predict that ensemble perception is an ongoing process, cleaning representations over time through iterative processing. This model makes a clear prediction for how error would adjust over time. A first pass, with an even distribution of attention, would provide an unbiased statistical average of all stimuli, with error biased to include outliers. This raw signal would then be used to redistribute attention to down-weight outliers, resulting in the responses becoming more clustered around the local mean over time. Importantly, using this method we will be able to gain some insight into the timing by which this occurs. If it is a discrete stepwise process, for example, with one pass to detect outliers and a second to provide a clear signal, we would predict a stepped or curved pattern in error over time. Conversely, if it is a continuous ongoing correction, we would predict a linear effect.

### Method

**Participants.** Twenty-seven participants (15 female, 4 left-handed, average age 20.7, range 18–27) were recruited from Baruch’s student subject pool for this experiment. Participants were compensated with course credit. To calculate power for the main effect of local versus global error in Experiment 3, we used PANGEA, an open source application designed specifically to conduct power analyses for ANOVA designs (Westfall, 2016). We anticipated potentially increased noise in this experiment as compared to Experiments 1 and 2, due to the short masking latencies and fewer trials per condition. For this reason, we carried out a power analysis using a more conservative effect size of  $d = .475$  (one half of our pilot effect size). The power analysis found that 26 participants would be sufficient for .80 power. We recruited 27 to account for potential no-shows. One participant was excluded due to self-expressed eyestrain, which caused difficulty completing the task, resulting in  $n = 26$  for data reported here.

**Stimuli.** Due to the timing required for this experiment, we used an alternate KDS monitor that could be set to a 60-hz refresh rate. This was required to display stimuli for exactly 50 ms.

Stimulus properties were identical to those described for the adjustable probe block in Experiment 2, including randomizing outlier direction and adjusting set orientations an equal and randomly selected amount on each trial, while maintaining set range and distribution. As in Experiment 2, set properties were defined before the experiment, so that sets with identical ranges and distributions could be repeated for each masking latency condition. This was done to ensure that any recorded differences in accuracy and RT could not be due to differences in set distribution or range.

Visual masks were structural masks constructed with 750 variably oriented lines identical in size and shape as those used in the experiments, with their centers randomly positioned in a  $10 \times 10^\circ$  square in the center of the screen. Twenty of these masks were generated prior to this experiment and saved as lossless images, with a randomly selected mask used for each trial. This ensured that participants would largely experience the same masks, but that masks would not be identical over the course of trials.

**Task.** The task was again identical to that described earlier for accuracy conditions in previous experiments, but with the addition of visual masking at stimulus onset asynchronies (SOAs) of 50, 100, 200, 300, and 500 ms. Stimuli were kept on the screen for the full SOA before masks appeared. Masks were displayed for 150 ms immediately followed by an adjustable probe. Participants were given as much time as required to respond. See Figure 1c for an example of the paradigm used for this experiment.

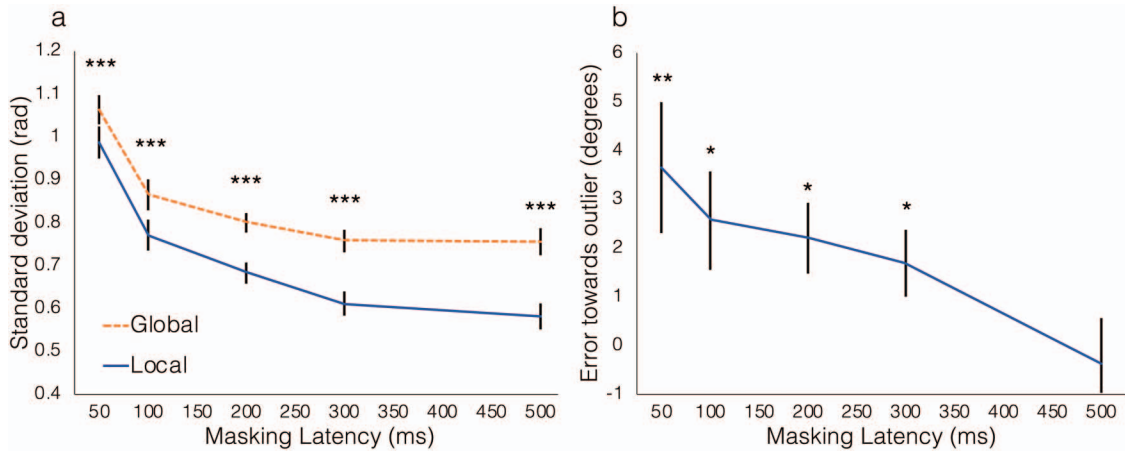
As we were primarily interested in the timing of outlier rejection, and in order to keep the experiment at a manageable length, all trials in Experiment 3 included outliers. Participants were presented with 40 trials at each masking latency, resulting in 200 trials overall.

**Analysis.** As described earlier, Von Mises distributions were fit to local and global error to allow measurements of circular  $SD$ . Again, no trials were excluded in calculation of local and global error. A  $2 \times 5$  repeated-measures ANOVA (local and global  $SD \times 5$  masking latencies) was used to test for main effects and interactions. To ensure the accuracy of planned contrasts, we designed the ANOVA to use a polynomial contrast with a metric specifically accounting for the unequal spacing of the masking latencies. Effect sizes for all tests are reported as partial eta squared. Follow-up paired sample  $t$  tests for each individual masking latency were additionally run between local and global  $SD$  to determine at which time points specifically differences could be observed.

Error data were again split by outlier direction for each masking latency to determine how much local error was biased toward the outlier for each processing duration. For this experiment, error was collapsed across bias direction due to fewer trials per condition. As described above, a  $2 \times 5$  repeated-measures ANOVA was used to test for main effects, again using a polynomial contrast accounting for unequal spacing of the masking latencies. Where violations of sphericity were detected, Greenhouse–Geisser corrections were applied. Follow-up one-sample one-tailed  $t$  tests were used to test bias toward the outlier at each actual masking latency.

### Results

See Figure 5 for results from Experiment 3. The repeated-measures ANOVA showed clear main effects of local versus global  $SD$ ,  $F(1, 25) = 53.77$ ,  $p < .001$ ,  $\eta_p^2 = .68$ , supporting that



**Figure 5.** Results for Experiment 3. a) Circular standard deviations of Von Mises distributions fit to participant error to local (blue [dark gray] solid) and global (orange [light gray] dashed) mean values. Asterisks mark results of paired  $t$  tests between measures. b) Error biased in the direction of the outlier, combined for outliers to the left and right of set means. Asterisks mark results of one-sample one-tailed  $t$  tests against 0. Error bars represent standard error using the Cousineau-Morey method. \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ . See the online article for the color version of this figure.

error was lower for local estimates. The effect of masking latency was also significant,  $F(3.05, 76.33) = 25.08, p < .001, \eta_p^2 = .50$ , suggesting that error decreased overall with increased exposure. Follow-up paired  $t$  tests showed that local  $SD$  was significantly lower than global  $SD$  for each individual masking latency (all  $t(25) < -3.99$ , all  $p < .001$ , all  $d > -0.78$ ). An interaction between local/global  $SD$  and masking latency was also found,  $F(3.24, 81.09) = 8.25, p < .001, \eta_p^2 = .25$ , indicating that local error reduced more compared to global error with longer stimulus durations. Tests of within-subjects contrasts revealed a linear,  $F(1, 25) = 43.37, p < .001, \eta_p^2 = .63$ , as well as a quadratic,  $F(1, 25) = 34.24, p < .001, \eta_p^2 = .58$ , trend for masking latency. Higher-order contrasts also reached significance but are not discussed further here, as they were not planned. The interaction also showed a significant linear trend,  $F(1, 25) = 22.72, p < .001, \eta_p^2 = .48$ .

Critically, the following test of bias toward outliers indicated that bias decreased over time,  $F(3.03, 75.76) = 2.99, p = .036, \eta_p^2 = .11$ , and showed a linear trend,  $F(1, 25) = 7.70, p = .01, \eta_p^2 = .24$ . Follow-up  $t$  tests revealed these differences to be significant at 50, 100, 200, and 300 ms (all  $t(25) > 1.85$ , all  $p < .038$ , all  $d > .36$ ), but not at 500 ms,  $t(25) = -.34, p = .63, d = -.06$ . Probe adjustment time was similar at all masking latencies,  $F(1, 25) = 2.55, p = .12, \eta_p^2 = .093$ .

## Discussion

Again, our results suggest that outliers are quickly and easily excluded from ensemble perception, with lower  $SD$  even at a 50-ms masking latency. However, these results further show that outliers were increasingly discounted from average calculations over time, with this process leveling off between 300 and 500 ms. The observed linear and quadratic trends provide evidence that this cleaning is carried out iteratively through ongoing ensemble calculations, potentially reflecting the use of derived statistics from earlier summaries to inform cleaning used in later statistical sum-

maries. Together, these results indicate that outlier rejection is an ongoing process, beginning surprisingly early in ensemble perception and continuing on through multiple iterations of statistical coding.

## General Discussion

The aim of our study was to explore the temporal dynamics of outlier rejection in ensemble perception. In Experiment 1, we extended results of prior research (Haberman & Whitney, 2010), showing that outliers are excluded when reporting the average tilt of groups with varying orientation, although we found that they increase overall error relative to when no outliers are present. Additionally, we found that the presence of outliers increased RT even while responses were given with an adjustable probe, and were thus unsped. In Experiment 2, we replicated the error results of Experiment 1 and also examined the possible effects of range on our results. The results showed faster RT and improved accuracy in the outlier condition compared to a matched range control, suggesting that outlier filtering effectively reduces range. Finally, in Experiment 3, we used variable latency visual masking to map the precise timing of outlier rejection. We found that while outliers were partially discounted even with 50-ms presentations, they were increasingly removed from average estimates over time, and were fully discounted by 500-ms post stimulus onset.

Altogether, these results suggest a solution to the apparent paradox of how ensemble representations filter outlier noise when outliers themselves must be defined by statistical ensemble properties. Outlier filtering appears to occur through an iterative process wherein perceived averages with deviant stimuli are refined over time. The idea of ensemble perception as a continuously updating process is consistent with current theories that postulate that ensemble perception occurs rapidly but also may adjust over time as attention is redistributed to objects of interest in the scene (Hochstein & Ahissar, 2002; Treisman, 2006). Here we provide

evidence in support of these theories, showing that initial statistical judgments are based more heavily on raw features, and only with sufficient processing time are they adjusted to provide an updated, clean signal.

The implications of outlier rejection for the mechanism of ensemble perception merit further discussion. Outlier rejection is technically an error since it produces estimates that are biased against the raw information in the image. It suggests that ensemble perception applies inferences in order to derive the true mean of natural groups. This could be adaptive in many cases in which natural groups appear in close proximity with other objects and must be distinguished (Utochkin, 2015). It could also make ensemble representations more robust against noise introduced by items registered with low fidelity in the system (Alvarez, 2011). However, it could also be maladaptive in cases in which the deviant carries important information or is truly a part of the group. For example, one yellow leaf on a tree may signify the very beginning of fall and must not be ignored. It is additionally important to note that participants performed these corrections despite receiving clear instructions to calculate the overall mean of the set, suggesting that this process is beyond conscious control. Further corroborating the unconscious nature of outlier filtering, the longer exposure in Experiment 3 resulted in participants discounting outliers more effectively as opposed to consciously correcting for the bias against them. Future research could productively contribute to understanding the biases built into ensemble perception by varying the degree by which outliers deviate, the frequency of outlier presence, and how these factors may influence the automaticity with which inferences are applied. These manipulations could help delineate the criteria that the visual system uses to define outliers and the flexibility by which these criteria are applied in different contexts.

The current study provides insight into the timing of outlier filtering. To our knowledge, there are no prior studies that have specifically measured outlier filtering over time. However, a handful of studies have explored the speed with which statistics are initially computed, and how the accuracy of the statistics changes over time. Using unmasked displays, Chong and Treisman (2003) found that mean size estimates were strikingly accurate with exposure durations as short as 50 or 100 ms, and that accuracy improved further with presentations of 1000 ms. These findings were later supported by Haberman, Harp, and Whitney (2009), who found that accuracy of ensemble coding of emotional expression in sequentially presented faces increased with overall exposure duration to the set. These results were partly corroborated by Li et al. (2016), who observed an improvement in mean facial emotion discrimination between 50 and 500 ms, but also, using a diffusion model analysis found an inverse quadratic relationship between exposure time and drift rate. Elaborating on these studies, we found that error does not decrease evenly over time, but rather that the weighting of individual items within the average is actively adjusted to downplay set deviants, and that this process appeared to level off between 300 and 500 ms after stimulus onset.

It is important to note that our results are consistent with previous studies in finding decrements in ensemble processing at short exposure durations. Whiting and Oriet (2011) found that participants were impaired in distinguishing the trial mean size with masked displays shorter than 200 ms. In our Experiment 3, performance was also least accurate at short exposure durations.

One possibility is that at short exposure durations, due to the noise in the representation, participants rely on information that they would downplay or altogether discard with more time. In the case of Whiting and Oriet (2011), participants based their responses on the experiment-wise mean. In our experiment, participants seem to derive noisy estimates from the set as a whole, including outliers that would be subsequently discounted. Note also that the specific parameters of the experiment, including the type of feature tested, could play an important role in how much information can be extracted at such short exposure durations. Our study examined average orientation compared to previous studies using short exposure durations with average size (Chong & Treisman, 2003; Whiting & Oriet, 2011) and therefore may have engaged partly different mechanisms, (Haberman, Brady, & Alvarez, 2015).

Taken together, findings that ensemble representations improve with increased exposure duration and our results that outliers are gradually filtered over time can be explained within the framework of ensemble perception proposed by Alvarez (2011). Alvarez proposed that ensemble representations rely on noisy individual item representations but nevertheless achieve high accuracy as they aggregate information across numerous items. The model also postulates that individual objects are unevenly weighed in ensemble estimates depending on the variance in their representation and their task relevance. It is possible that with increased exposure duration, individual item representations become less noisy perhaps due to repeated measurements. If this were the case, then mean estimates would become more accurate with time, and outliers would gradually become distinct from other items and progressively discounted from the group.

A study by Hochstein et al. (2018) provides corroborating evidence that outliers are detected based on their overlap with other items in the set and suggests that outliers are tagged by virtue of their separate peak in the population code. In this study, participants either compared the mean orientation of two sets of oriented lines or detected the presence of an orientation outlier in one of the sets. Performance in the outlier detection task was best predicted by the distance of the outlier to the edge of the set range, whereas mean discrimination was independent of range. The authors suggested that mean and range estimates are computed through the same population coding mechanism, with the former contributing to set perception and the latter allowing individual objects to be separated from the set. Importantly, Hochstein et al. (2018) found that outlier detection incurred longer RTs compared to set mean calculations. These results corroborate the idea that outliers emerge from the population code later than mean properties, which could also partly explain why they are discounted at longer exposure durations. The current study is consistent with the results of Hochstein et al. and also suggests that population coding may change over time, such that outliers become more easily separable from the set and receive reduced weight in mean estimates.

An interesting question for further research is whether outlier filtering depends on categorical differences between the outlier and the mean of the set. In the Haberman and Whitney (2010) study, which presented faces in a continuum of emotions from happy to sad, the outlier was always categorically different (e.g., it was happy when the rest of the set was on average sad). Similarly, in our study, the outlier was almost always on a different side of vertical than the rest of the set. Only in the speeded response task

of Experiment 2, in which participants made a categorical judgment (whether the mean was to the left or right of vertical), was the category of the outlier balanced so that it was orthogonal to the correct response. Here, when splitting data by the direction of the outlier relative to the correct response, we found that accuracy was lower,  $t(15) = 4.79$ ,  $p < .001$ ,  $d = 1.20$ , and RT slower,  $t(15) = -2.75$ ,  $p = .015$ ,  $d = -.69$ , when the outlier was in the opposite direction compared to the same direction as the local mean. These results align better with the partial filtering account discussed previously as compared to a categorical rejection account. We note, however, that the categorical responses in this experiment do not provide a precise comparison of participants' mean estimates to the global and local means. Therefore, an experiment that manipulates outlier category and measures mean estimates using an adjustable probe is needed to better examine this question.

The idea of ensemble perception as an iterative process leads to the question of how, if ensemble percepts continuously update, a response threshold is met. Studies using RT measures can shed light on this question. Previously Robitaille and Harris (2011) showed that RT decreases as the number of items within a set increases, suggesting that a response threshold is reached earlier with larger sets. The results of Experiment 2 also show that participants took additional time to filter outliers and to resolve an ensemble of increased range. Taken together, these results suggest that ensemble estimates converge at different times depending on set size, range, and, in our case, presence of deviant stimuli. Importantly, a response is made only when confidence in the ensemble estimate has been achieved. The criteria by which participants evaluate ensemble noise at different points in time remain unknown. Relatedly, whether knowledge about the fidelity of ensemble representations is explicit is an important question for further research.

One question that remains in the current study is whether deviant stimuli are represented as individual objects, separately from the ensemble, or are altogether discounted. Haberman and Whitney (2010) found that participants were unable to localize or identify emotional deviants, suggesting that outlier rejection did not depend on a conscious suppression of the deviant. However, the face stimuli they used were more complex than our orientations and have been previously shown to be resistant to pop-out effects (Brown, Huey, & Findlay, 1997; VanRullen, 2006; but see Hershler & Hochstein, 2005, 2006). Hochstein et al. (2018) found that outliers can pop out with sufficient distance from the ensemble range edge. However, they used a task that instructed participants to explicitly detect the outliers, which could have influenced the results. Similarly, De Fockert and Marchant (2008) asked participants to attend to the most extreme items in a display of varying sizes and found an increased contribution of these items to mean estimates. In contrast, in the current, study participants were never told of the presence of outliers and it is therefore unclear whether these outliers attracted attention. Future studies need to test the role of attention, for example by asking participants to explicitly identify the outliers or detect a target in the spatial location of the outliers. It is possible that the discounting of outliers and contribution to the mean are modulated by task relevance and attention. Nevertheless, the encoding of the outlier is an orthogonal issue to the paradoxical nature of its filtering from ensemble representations.

Our results also contribute to a growing literature exploring the effects of set range on ensemble perception. In Experiment 2, we were specifically interested in testing whether increased RTs to the outlier displays could simply be due to increased range in the display. Multiple previous studies have shown that increased range significantly impairs ensemble perception (Ji & Pourtois, 2018; Maule & Franklin, 2015; Utochkin & Tiurina, 2014), presumably due to increased noise in the signal. Importantly, Utochkin and Tiurina (2014) showed that when range was held constant, accuracy remained consistent across set sizes and distributions, with the clear exception of cases where items were presented with a bimodal distribution. In these situations, accuracy dropped markedly, likely due to sets being automatically split into separate groups. Our results contribute to this debate by replicating the finding that increased range can have a sharp effect on accuracy and RT, even while the set size and overall distribution within the set remains even. However, the contribution of outliers to observed range effects is clearly distinct in that while outliers may impair accuracy initially, possibly due to an increased set range, outliers are easily segregated and rejected from the set. This allows new iterations of calculation on a set with a decreased range (and thus less noise) that can be performed more accurately. Future studies will have to carefully disentangle the possibly separable effects of set size, range, and distribution on ensemble perception, taking stimulus presentation time into careful consideration.

Overall, the results help to resolve the paradox of outlier rejection by showing how ensemble representation serves to both identify and filter outliers. We propose that ensemble perception takes place via iterative steps of statistical coding, wherein initial ensemble percepts provide raw, unbiased representations of the world that are then corrected over time with the goal of providing more accurate and stable views. The proposed mechanism has broader implications. We postulate that adjustments to ensemble representations are made not only away from deviant items (Haberman & Whitney, 2010) but also toward salient items (De Fockert & Marchant, 2008), or in favor of perceived as opposed to physical individual object properties (Dodgson & Raymond, 2020; Im & Chong, 2009; Tiurina & Utochkin, 2019). Thus, the idea of iterative ensemble coding bridges findings that statistical summaries are rapidly available with those that suggest that they can be systematically biased and that they factor in context parameters that require more extensive processing.

Our results support the idea that iterative ensemble coding occurs while the displays are in view, since outlier filtering improved with increased exposure duration in Experiment 3. Iterations may have also continued after display offset on iconic memory representations when displays were unmasked (Experiment 2), incurring an RT cost for the outlier compared to the no-outlier condition (Rensink, 2014). Although it is possible that iterative ensemble coding continues beyond visual and iconic memory representations, this possibility cannot be evaluated using our experimental design. Future studies comparing iterative ensemble coding for trials of fixed duration (display to probe onset) but varying visual stimulation time are needed to shed light on the possibility that iterative ensemble coding continues after visual stimulation.

Iterative computations are prevalent in the visual system and are thought to play a role in several different processes, including the refining of individual object representations (Lamme & Roelfs-

sema, 2000) and the emergence of objects into awareness (Dehaene & Naccache, 2001). For example, metacontrast (Breitmeyer & Ogmen, 2000) and object substitution masking (Enns & Di Lollo, 1997) are thought to rely on iterative or reentrant processing that discards the noisy representation of the masked object in favor of the more robust representation of the mask that appears in the same location. There are noted similarities between this re-entrant processing and the iterative processing proposed for ensemble processing in the current study. However, re-entrant processing seems to operate on a faster time scale (Enns & Di Lollo, 1997), which suggests that it may engage networks at a more local level. Therefore, it seems more likely that the two processes are separate, although they may interplay in interesting ways. Future studies should examine this potential interplay, for example by testing how ensemble representations take into account masked outliers.

The idea of ensemble perception as an iterative process needs to be rigorously evaluated in further experiments. One major step in this direction would be to include exposure duration manipulations in experiments that show biases or corrections in ensemble estimates and measure how these biases develop over time (e.g., De Fockert & Marchant, 2008). In addition, we propose that RT measures could be useful in understanding the time requirements to resolve ensemble properties under different task and stimulus conditions. Nevertheless, the idea of iterative ensemble perception is consistent with existing theories of ensemble perception as a quickly updating mechanism that serves to guide attention toward more detailed processing (Hochstein & Ahissar, 2002) and to achieve continuity across time in an ever-changing complex visual world (Manassi, Liberman, Chaney, & Whitney, 2017).

## References

- Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*, 122–131. <http://dx.doi.org/10.1016/j.tics.2011.01.003>
- Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 7345–7350. <http://dx.doi.org/10.1073/pnas.0808981106>
- Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*, 157–162. <http://dx.doi.org/10.1111/1467-9280.00327>
- Bauer, B. (2017). Perceptual averaging of line length: Effects of concurrent digit memory load. *Attention, Perception, & Psychophysics*, *79*, 2510–2522. <http://dx.doi.org/10.3758/s13414-017-1388-4>
- Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 7. <http://dx.doi.org/10.1167/9.10.7>
- Berens, P. (2009). CircStat: A MATLAB toolbox for circular statistics. *Journal of Statistical Software*, *31*, 293–295. <http://dx.doi.org/10.18637/jss.v031.i10>
- Breitmeyer, B. G., & Ogmen, H. (2000). Recent models and findings in visual backward masking: A comparison, review, and update. *Perception & Psychophysics*, *62*, 1572–1595. <http://dx.doi.org/10.3758/BF03212157>
- Brown, V., Huey, D., & Findlay, J. M. (1997). Face detection in peripheral vision: Do faces pop out? *Perception*, *26*, 1555–1570. <http://dx.doi.org/10.1068/p261555>
- Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*, 393–404. [http://dx.doi.org/10.1016/S0042-6989\(02\)00596-5](http://dx.doi.org/10.1016/S0042-6989(02)00596-5)
- Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the bandwidth of perceptual experience? *Trends in Cognitive Sciences*, *20*, 324–335. <http://dx.doi.org/10.1016/j.tics.2016.03.006>
- Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*, 42–45. <http://dx.doi.org/10.20982/tqmp.01.1.p042>
- Cousineau, D., & O'Brien, F. (2014). Error bars in within-subject designs: A comment on Baguley (2012). *Behavior Research Methods*, *46*, 1149–1151. <http://dx.doi.org/10.3758/s13428-013-0441-z>
- Dakin, S. C., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research*, *37*, 3181–3192. [http://dx.doi.org/10.1016/S0042-6989\(97\)00133-8](http://dx.doi.org/10.1016/S0042-6989(97)00133-8)
- De Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception & Psychophysics*, *70*, 789–794. <http://dx.doi.org/10.3758/PP.70.5.789>
- De Fockert, J. W., & Wolfenstein, C. (2009). Short article: Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology*, *62*, 1716–1722. <http://dx.doi.org/10.1080/17470210902811249>
- Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, *79*, 1–37. [http://dx.doi.org/10.1016/S0010-0277\(00\)00123-2](http://dx.doi.org/10.1016/S0010-0277(00)00123-2)
- Dodgson, D. B., & Raymond, J. E. (2020). Value associations bias ensemble perception. *Attention, Perception, & Psychophysics*, *82*, 109–117. <http://dx.doi.org/10.3758/s13414-019-01744-1>
- Enns, J. T., & Di Lollo, V. (1997). Object substitution: A new form of masking in unattended visual locations. *Psychological Science*, *8*, 135–139. <http://dx.doi.org/10.1111/j.1467-9280.1997.tb00696.x>
- Epstein, M. L., & Emmanouil, T. A. (2017). Ensemble coding remains accurate under object and spatial visual working memory load. *Attention, Perception, & Psychophysics*, *79*, 2088–2097. <http://dx.doi.org/10.3758/s13414-017-1353-2>
- 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*, 175–191. <http://dx.doi.org/10.3758/BF03193146>
- 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*, 432–446. <http://dx.doi.org/10.1037/xge0000053>
- Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision*, *9*(11), 1. <http://dx.doi.org/10.1167/9.11.1>
- Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 718–734. <http://dx.doi.org/10.1037/a0013899>
- Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics*, *72*, 1825–1838. <http://dx.doi.org/10.3758/APP.72.7.1825>
- Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review*, *18*, 855–859. <http://dx.doi.org/10.3758/s13423-011-0125-6>
- Hershler, O., & Hochstein, S. (2005). At first sight: A high-level pop out effect for faces. *Vision Research*, *45*, 1707–1724. <http://dx.doi.org/10.1016/j.visres.2004.12.021>
- Hershler, O., & Hochstein, S. (2006). With a careful look: Still no low-level confound to face pop-out. *Vision Research*, *46*, 3028–3035. <http://dx.doi.org/10.1016/j.visres.2006.03.023>
- Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, *36*, 791–804. [http://dx.doi.org/10.1016/S0896-6273\(02\)01091-7](http://dx.doi.org/10.1016/S0896-6273(02)01091-7)

- Hochstein, S., Pavlovskaya, M., Bonnef, Y. S., & Soroker, N. (2018). Comparing set summary statistics and outlier pop out in vision. *Journal of Vision, 18*(13), 12. <http://dx.doi.org/10.1167/18.13.12>
- Im, H. Y., & Chong, S. C. (2009). Computation of mean size is based on perceived size. *Attention, Perception, & Psychophysics, 71*, 375–384. <http://dx.doi.org/10.3758/APP.71.2.375>
- Ji, L., & Pourtois, G. (2018). Capacity limitations to extract the mean emotion from multiple facial expressions depend on emotion variance. *Vision Research, 145*, 39–48. <http://dx.doi.org/10.1016/j.visres.2018.03.007>
- Ji, L., Rossi, V., & Pourtois, G. (2018). Mean emotion from multiple facial expressions can be extracted with limited attention: Evidence from visual ERPs. *Neuropsychologia, 111*, 92–102. <http://dx.doi.org/10.1016/j.neuropsychologia.2018.01.022>
- Khayat, N., & Hochstein, S. (2018). Perceiving set mean and range: Automaticity and precision. *Journal of Vision, 18*(9), 23. <http://dx.doi.org/10.1167/18.9.23>
- Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences, 23*, 571–579. [http://dx.doi.org/10.1016/S0166-2236\(00\)01657-X](http://dx.doi.org/10.1016/S0166-2236(00)01657-X)
- Lau, J. S. H., & Brady, T. F. (2018). Ensemble statistics accessed through proxies: Range heuristic and dependence on low-level properties in variability discrimination. *Journal of Vision, 20*(5), 1. <http://dx.doi.org/10.1167/18.9.3>
- Leib, A. Y., Kosovicheva, A., & Whitney, D. (2016). Fast ensemble representations for abstract visual impressions. *Nature Communications, 7*, 13186. <http://dx.doi.org/10.1038/ncomms13186>
- 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*, 1332. <http://dx.doi.org/10.3389/fpsyg.2016.01332>
- Manassi, M., Liberman, A., Chaney, W., & Whitney, D. (2017). The perceived stability of scenes: Serial dependence in ensemble representations. *Scientific Reports, 7*, 1971. <http://dx.doi.org/10.1038/s41598-017-02201-5>
- Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision, 15*(4), 6. <http://dx.doi.org/10.1167/15.4.6>
- Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*, 61–64. <http://dx.doi.org/10.20982/tqmp.04.2.p061>
- Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods, 162*, 8–13. <http://dx.doi.org/10.1016/j.jneumeth.2006.11.017>
- Rensink, R. A. (2014). Limits to the usability of iconic memory. *Frontiers in Psychology, 5*, 971. <http://dx.doi.org/10.3389/fpsyg.2014.00971>
- Robitaille, N., & Harris, I. M. (2011). When more is less: Extraction of summary statistics benefits from larger sets. *Journal of Vision, 11*(12), 18. <http://dx.doi.org/10.1167/11.12.18>
- Tiurina, N. A., & Utochkin, I. S. (2019). Ensemble perception in depth: Correct size-distance rescaling of multiple objects before averaging. *Journal of Experimental Psychology: General, 148*, 728–738. <http://dx.doi.org/10.1037/xge0000485>
- Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition, 14*, 411–443. <http://dx.doi.org/10.1080/13506280500195250>
- Utochkin, I. S. (2015). Ensemble summary statistics as a basis for rapid visual categorization. *Journal of Vision, 15*(4), 8. <http://dx.doi.org/10.1167/15.4.8>
- Utochkin, I. S., & Tiurina, N. A. (2014). Parallel averaging of size is possible but range-limited: A reply to Marchant, Simons, and De Fockert. *Acta Psychologica, 146*, 7–18. <http://dx.doi.org/10.1016/j.actpsy.2013.11.012>
- VanRullen, R. (2006). On second glance: Still no high-level pop-out effect for faces. *Vision Research, 46*, 3017–3027. <http://dx.doi.org/10.1016/j.visres.2005.07.009>
- Watamaniuk, S. N. J., & Duchon, A. (1992). The human visual system averages speed information. *Vision Research, 32*, 931–941. [http://dx.doi.org/10.1016/0042-6989\(92\)90036-I](http://dx.doi.org/10.1016/0042-6989(92)90036-I)
- Westfall, J. (2016). *Power ANalysis for GEneral Anova designs* [Computer software]. Retrieved from <http://jakewestfall.org/pangea/>
- Whiting, B. F., & Oriet, C. (2011). Rapid averaging? Not so fast! *Psychonomic Bulletin & Review, 18*, 484–489. <http://dx.doi.org/10.3758/s13423-011-0071-3>
- Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology, 69*, 105–129. <http://dx.doi.org/10.1146/annurev-psych-010416-044232>

Received October 16, 2019

Revision received June 11, 2020

Accepted June 18, 2020 ■