Characterising the role of awareness in ensemble perception

Patxi Elosegi[1,2], Ning Mei[1], and David Soto[1,3]

[1]*Basque Center on Cognition, Brain and Language, San Sebastian, Spain*

[2]*University of the Basque Country- UPV/EHU, Spain*

[3]*Ikerbasque, Basque Foundation for Science, Bilbao, Spain*

January 15, 2024

Author Note

The authors have no known conflict of interests to disclose.

The authors confirm that the data and code supporting the findings of this study are available at OSF in the following link https://osf.io/whr2n/?view$_o nly = eb32844bc86a42b2b7e41ee1e48bacc6$ (Elosegi and Soto, 2023).

Correspondence to:p.elosegi@bcbl.eu, d.soto@bcbl.eu, Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, 2nd Floor 20009 San Sebastian.

Abstract

Ensemble representations are efficient codes that the brain generates effortlessly even under noisy conditions. However, the role of visual awareness for computing ensemble representations remains unclear. We present two psychophysical experiments (N=15x2) using a bias-free paradigm to investigate the contribution of conscious and unconscious processing to ensemble perception. Here we show that ensemble perception can unfold without awareness of the relevant features that define the ensemble. Computational modeling of the type-1 and type-2 drift-rates further suggest that awareness lags well behind the categorization processes that support ensemble perception. Additional evidence indicates that the dissociation between type-1 from type-2 sensitivity, was not driven by type-2 inefficiency or a systematic disadvantage in type-2 decision making. The present study demonstrates the utility of robust measures for studying the role of visual consciousness and metacognition in stimuli and tasks of increasing complexity, crucially, without underestimating the contribution of unconscious processing in an otherwise visible stimulus.

Public significance statement

Psychologists have been studying the contribution of conscious and unconscious processes in human perception, but previous research mainly looked at how we recognize simple single objects. However, our conscious experience involves a lot of different and complex things happening at the same time. We created a way to study how visual awareness affects our perception of groups of things (i.e., ensembles), without underestimating the role of the unconscious mind. We found that human observers can still understand and process the big picture of a group, even when they are not consciously aware of the critical information defining the ensemble. This discovery can help us better understand how we perceive complex things and the scope of non-conscious processes in visual perception.

Keywords: awareness, ensemble perception, unconscious processing, 2IFC, task-relevant features

Introduction

Each second $\sim 2 \cdot 10^8$ photons reach our retinae (Rodieck, 1998) posing a huge challenge for the visual system, which is further constrained by limited attentional and working memory resources (Cohen et al., 2012; Luck and Vogel, 2013). The brain compensates for these computational limitations by exploiting the fact that visual information is often correlated and redundant (Cohen et al., 2016). For instance, when we look at the mountains, we have effortless access to the landscape's hue distribution, which the visual system encodes as a summary statistic (Alvarez, 2011). Ensemble representations are efficient codes that the brain generates by pooling together many noisy spatial or temporal measurements in the form of a probabilistic representation (Whitney and Leib, 2018).

Four decades of psychophysical work have demonstrated that human observers can extract the average of different kinds of low-level visual features, namely, motion direction (Watamaniuk et al., 1989), speed (Watamaniuk and Duchon, 1992), center of mass (Alvarez and Oliva, 2008), orientation (Dakin and Watt, 1997), colour (Webster et al., 2014) and also size (Ariely, 2001; Chong and Treisman, 2005b). More recent work has further investigated the scope of ensemble perception for higher-order visual features such as the average facial identity (Bai et al., 2015), emotional expression (Haberman and Whitney, 2007), gender (Haberman and Whitney, 2009) and animacy (Leib et al., 2016). Even though the underlying neural mechanisms for averaging low-level and high-level visual information may be different (Haberman and Whitney, 2012), these studies suggest that the computation of summary statistics is a ubiquitous feature of the visual system. Such a general mechanism has been linked to a variety of visual processing functions such as texture perception (Cavanagh, 2001), outlier detection (Alvarez, 2011) or processing the gist of a scene (Oliva and Torralba, 2006). In sum, representing multiple noisy measurements as an ensemble can enhance visual cognition (Alvarez, 2011).

A key issue relates to the automaticity of ensemble coding and whether it can be computed outside the focus of attention. Different studies have demonstrated that participants can indeed extract the average feature of a group of items automatically even when they are not explicitly required to do so (see Ariely, 2001; Ji and Hayward, 2021; Neumann et al., 2013). Crowding experiments also reveal that adaptation (Harp et al., 2007) and motion perception (Allik, 1992) occur even when participants cannot individuate single items from the group. Besides, there is evidence that attention may not be necessary for ensemble perception (Alvarez and Oliva, 2008; Bronfman et al., 2014; Chong and Treisman, 2005a; Whitney and Leib, 2018), thereby also suggesting some sort of implicit tracking of the stimulus summary statistics. However, it should be noted that attention can significantly influence ensemble perception (Chong and Treisman, 2005a; de Fockert and Marchant, 2008).

For instance, Huang (2015) observed that the effect of attention, measured as the advantage of pre-cueing one of two tasks, was similar across tasks involving single-object features or ensemble statistical properties. Similarly, Jackson-Nielsen and colleagues' (2017) showed inattentional blindness in many observers performing a color diversity task, suggesting that attention is crucial for a conscious perception of the ensemble. Therefore, while some level of attention may be necessary for conscious ensemble perception, ensemble information can be extracted and influence behavior, even when attentional resources are constrained (Corbett et al., 2023).

The fact that ensembles seem to be processed automatically and with minimal attention brings into question the role of awareness in ensemble perception. Prior studies showed that awareness of the individual elements within the ensemble may not be necessary for extracting the summary statistic (Bronfman et al., 2014; Haberman and Whitney, 2009; Oriet and Corbett, 2008; Ward et al., 2016). More recently, it has been shown that masked portions of the ensemble can still influence the overall averaging decision (Choo and Franconeri, 2010; Sekimoto and Motoyoshi, 2022).

However, measuring awareness is not any trivial issue. Previous studies investigating the role of awareness in visual perception (including the ones mentioned above) have relied mostly on subjective measures, which can be contaminated by criterion biases in reporting the presence or absence of awareness (i.e. *the subjective criterion problem*, Eriksen, 1960; Michel, 2022). The use of objective measures (i.e., null sensitivity) to determine the absence of awareness could potentially resolve this issue. However, this approach typically uses masking techniques that severely reduce the signal to noise ratio in the stimuli and hence limits the ability to isolate traces of unconscious perception at the behavioral level (see Soto et al., 2019; Mei et al., 2022). We also note perceptual thresholds may vary across sessions and hence any awareness test that uses objective signal detection measures to establish the presence or absence of awareness must be conducted within the same experimental session as the one aiming to show the effects of unconscious items on behavior or the brain. This approach diverges from the conventional methodology employed in subliminal priming studies. Typically, the awareness of the priming stimulus is evaluated 'offline' either prior to or following the primary experiment, rather than being assessed on a trial-by-trial basis during the main experiment. Because of this, subliminal priming studies have been subjected to criticism (Newell and Shanks, 2014) including failures to replicate (Stein et al., 2020), thereby leading to the view that unconscious information processing is, if anything, limited in scope.

Recent developments in two-interval-forced-choice tasks (2-IFC) (Barthelmé and Mamassian, 2010) provide a potential solution to the criterion problem. Peters and Lau (2015)

adapted a 2-IFC to test unconscious perception. The key manipulation of this study was that unbeknownst to the participants only one of the intervals contained a (masked) stimulus (i.e., an oriented grating presented at different levels of signal strength) while the other interval contained no target stimulus. Despite that, the task required to make a type-1 orientation discrimination judgment for each interval, followed by a type-2 forced-choice decision regarding which of the two intervals they felt more confident in. Importantly, since this paradigm does not require participants to map their states of visibility or perceptual confidence into a continuous rating scale, minimises the effect of criterion bias on measured variables (Mamassian, 2020). As such, above-chance orientation discrimination performance accompanied by chance-level metacognitive detection of the stimulus-present interval would suggest unconscious perception.

In psychophysics, there is a long tradition of using confidence ratings as a measure of stimulus awareness (Kolb and Braun, 1995; Peirce and Jastrow, 1884). However, this practice is not free of criticism (Michel, 2023; Rosenthal, 2019). For instance, some studies have shown above chance metacognitive sensitivity in conditions of null visual awareness (Evans and Azzopardi, 2007; Jachs et al., 2015; Persaud et al., 2007; Reder and Schunn, 2014), indicating that the relationship between consciousness and metacognition is complex and both constructs might be dissociated. Nevertheless, confidence ratings have proven valuable in assessing consciousness, sometimes better than subjective visibility ratings (Morales and Lau, 2021).

Using the 2-IFC task Peters and Lau (2015) did not observe any gradient of above chance discrimination performance when the type-2 decisions failed to discriminate the informative interval. In fact, they observed that as soon as participants were able to discriminate the grating's orientation, they were also able to detect the informative interval, thereby showing no evidence of unconscious perception. Interestingly, the same pattern of results was observed in subsequent experiments using different masking techniques (see Peters et al., 2017; Knotts et al., 2018). These results indicate that under the strong assumptions of the 2-IFC task participants do not display a behavioral pattern consistent with unconscious perception.

However, a stimulus like a grating is a multidimensional percept composed of multiple features like luminance, color, orientation, etc. It is therefore important to note that the ability to detect the presence of a stimulus in one of the intervals of the 2-IFC does not necessarily indicate that observers were aware of the orientation information guiding their type-1 discrimination responses. Indeed, equating awareness of the task-relevant features to the awareness of something at all can result in the underestimation of unconscious perception (i.e. *the criterion content fallacy*, Kahneman and Miller, 1986; Michel, 2022). Hence, for

type-2 responses to be informative, they should reflect the level of awareness of the task-relevant features, that is, of the very same features underlying type-1 performance (Michel, 2022). To avoid underestimating unconscious perception using the 2-IFC, visual ensembles rather than single stimulus can be used. This approach involves assessing the presence versus absence of task-relevant features of stimuli across two intervals, as opposed to requiring observers to detect the presence or absence of a stimulus as in previous studies (Peters and Lau, 2015; Peters et al., 2017; Knotts et al., 2018.

The study by Peters and Lau (2015) and also Knotts et al. (2018) did not consider the criterion content fallacy. This factor can explain the failure to dissociate type-1 from type-2 performance in their studies. We propose that by using a non-informative stimulus instead of an absent stimulus in the non-informative interval type-1 and type-2 tasks may be better matched in terms of the task-relevant features, hence addressing the criterion content fallacy (Kahneman and Miller, 1986; Michel, 2022). This is critical for setting the experimental conditions to dissociate type-1 and type-2 performance (i.e. here ensemble perception and visual awareness). However, it is important to note that even after accounting for the criterion content fallacy in studies of single object perception, additional challenges remain, in particular, the use of strong image degradation techniques (e.g., masking, low luminance, brief presentation times). If the stimulus strength is very small, it may be extremely difficult to observe above chance discrimination performance when the stimulus can not even be detected (Lau, 2022). Using visual ensembles in this context does not impose severe constraints in visibility. By simply manipulating the ratio of the two classes of objects within the ensemble, high levels of visual uncertainty can be generated, without the need for degradation of the physical properties of the stimulus. In other words, unmasked visual ensembles may allow for a higher signal to noise ratio in the informative interval compared to what would be achieved by visual masking.

Accordingly, the present study aimed to characterize the role of awareness in ensemble perception in an unbiased way and without underestimating unconscious perception. By doing so, we also wanted to address the long-standing, theoretical question of whether it is possible to isolate traces of unconscious perception while simultaneously controlling for the subjective criterion problem and the criterion content fallacy. To achieve this, we adapted the Peters and Lau (2015) approach by presenting two intervals of temporal ensembles, with only one interval containing an informative sequence of objects (i.e. a sequence with a predominant animacy class), while the other interval presented the same number of living and non-living objects (an information-absent interval). We hypothesized that if ensemble perception can unfold without awareness of the task relevant features, then participants would perform above chance discriminating the predominant animacy class for the informative

interval, while their type-2 discrimination performance will be no different from chance .

Experiment 1: Awareness in ensemble perception

Methods

Transparency and openness

The experimental data and scripts are available at OSF in the following link https://osf.io/whr2n/?vi*eb*32844*bc*86*a*42*b*2*b*7*e*41*ee*1*e*48*bacc*6 (Elosegi and Soto, 2023). The experiments were not pre-registered.

Participants

Fifteen right-handed healthy subjects (12 women, $\overline{X}_{\text{Age}} = 25$ years, $\text{SD}_{\text{Age}} = 4.44$ years) were recruited. We conducted a power analysis using G*power (Faul et al., 2009) to determine the minimum sample size required to achieve above-chance type-1 discrimination performance across experimental conditions with $\alpha = .001$ and an expected power (1 - β) = .99. This analysis was based on a prior study investigating ensemble discrimination which demonstrated significant predominant animacy discrimination performance, with a Cohen's d effect size of 2.35 and a sample size of 10 participants (Tiurina and Markov, 2022). The power analysis results showed that 11 participants would be needed in this case. However, given that our aim was also to provide evidence of chance-level type-2 discrimination performance, we elected to fix the sample size to 15 and use Bayesian statistics to estimate the evidence in favour of the null finding (i.e. chance-level discrimination). We also conducted a sensitivity analysis which confirmed that a sample size of 15 participants is sufficient to detect an effect size of 1.70 with $\alpha = .001$ and power (1 - β) = .99 in terms of type-1 performance. All of them had normal or corrected to normal vision, gave informed consent before the experiment and were reimbursed with 8 euros per hour of experiment. The protocol was approved by the BCBL's Ethics Review Board and complied with the guidelines of the Helsinki Declaration (2008). Data were acquired during the spring of 2022.

Apparatus and stimuli

The experiment was programmed in Python using OpenSesame (Mathôt et al., 2012) and it was displayed on a Viewsonic G90fB computer monitor with a resolution of 1024×768 and a refresh rate of 60Hz. To ensure that participants maintained a viewing distance of 70cm, they were instructed to use a chin-rest. The stimuli set was composed of 96 images selected from an original set of 360 high-quality ecological color images (Moreno-Martínez and

Montoro, 2012). Half of the images contained living objects and the other half contained non-living objects. The living category could be further divided into mammals, insects, birds and marine creatures. The non-living category could be further divided into furniture, kitchen items, musical instruments, tools and vehicles. Stimuli from both classes were balanced in terms of different visual and linguistic variables (See supplementary table 1).

Procedure

On each trial of the 2IFC experiment, participants observed two consecutive sequences of forty items containing varying proportions of living and nonliving objects presented in a rapid serial visual presentation fashion (RSVP) (Figure 1A). More specifically, each item was presented for 3 frames ($\sim 50ms$) which prior research has shown to be within the range of the temporal integration limits of ensemble perception (Leib et al., 2016; Whitney and Leib, 2018). After watching each interval, participants were asked to complete the type-1 task which required them to estimate the more frequent object category (either living or nonliving) in the string. At the end of the trial, participants had to complete the type-2 task which was a confidence forced choice between the first and second intervals. To do so, participants were instructed to introspect on their internal confidence states and discriminate the interval they felt more confident in. Critically, unbeknownst to the participants, only one of the intervals displayed an informative ratio of living and nonliving objects (the informative interval) while the other interval showed the same proportion of items from each category (non-informative interval) (Figure 1B). Thus, only one of the sequences, the informative sequence, actually contained information to complete the type-1 task.

Under these conditions, the type-2 response was based on the comparison between the confidence signal generated by an ensemble without a predominant class from an ensemble with a predominant class. In other words, it involved distinguishing the absence from the presence of an objective signal that could be further associated with a correct response. Hence, we anchored the confidence state regarding the stimulus present interval to that generated by the non-informative interval, thereby mitigating confounds from the placement of subjective criteria (see Peters and Lau, 2015). It's worth noting however that the 2IFC task is not entirely free of subjective biases (Yeshurun et al., 2008), albeit of a different nature. For instance, participants might have preferences for selecting either the first or the second interval, but this is unrelated to biases in confidence criteria, which can occur in tasks involving the detection or discrimination of single stimuli.

In order to characterize the role of awareness of task relevant features in ensemble perception, we systematically manipulated the ensemble ratio of the majority class in the informative interval across eleven possible ratios of living and nonliving objects. These con-

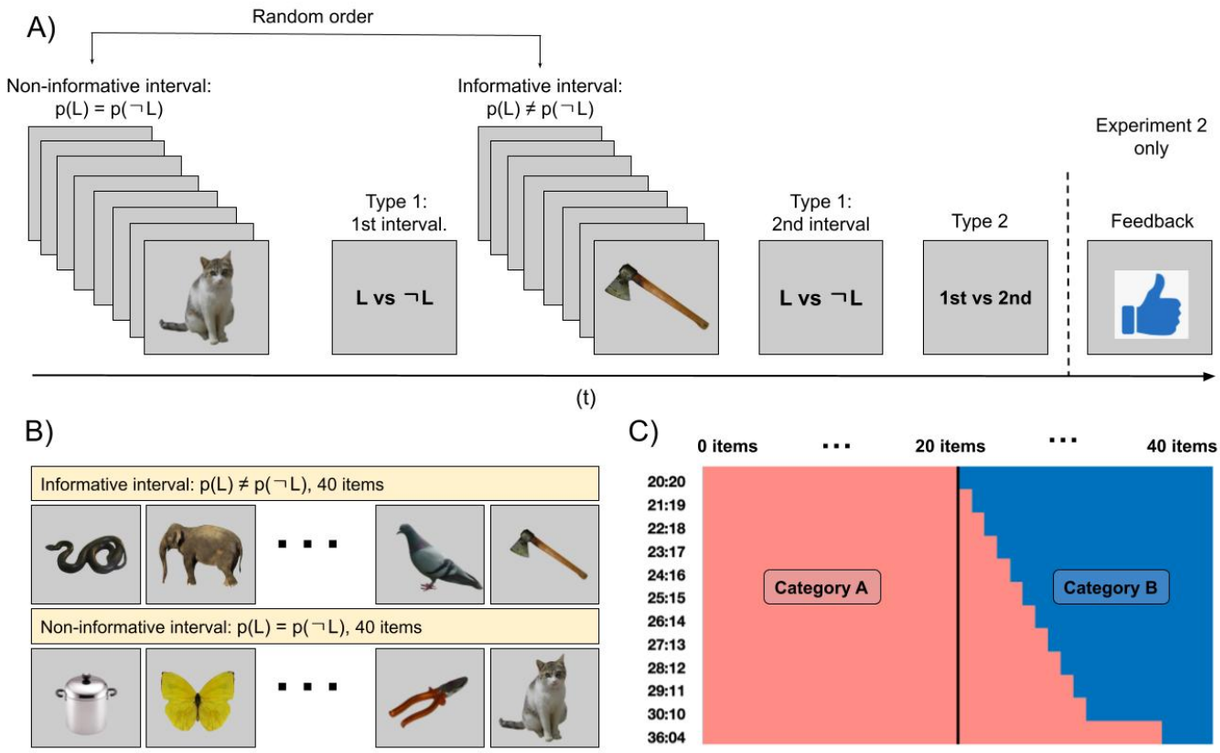

Figure 1: ***The 2IFC design used in Experiments 1 and 2.*** *A) Trial protocol. Each trial involved the RSVP of two sequences composed of forty images of living and nonliving objects. Despite that only one of the sequences had an informative ratio of items, the type-1 task required to estimate the most frequent object category for both intervals. At the end of the trial, participants had to report on which of the two intervals they felt more confident. The presentation order of the informative sequence was pseudo-randomized across trials. Only Experiment 2 provided feedback conditional upon type-2 responses on the informative interval. B) Example of the object sequences in the informative and non-informative intervals. The non-informative sequence presented the same number of fully-colored living and nonliving items. While in the informative sequence one category always predominated over the other. C) Visual depiction of all possible conditions of ensemble ratio in the temporal ensembles. The x axis represents the 40 items in the list. The y axis represents all the ensemble ratios for each category (blue and red). The first row depicts the 20:20 ratio always presented in the non-informative interval, while the rest of the 11 conditions (21:19-36:04) were presented in the informative sequence. For explanatory purposes all the images from each category are represented together, however in the experiment the order of presentation of the items from each category was interleaved.*

ditions ranged from the minimum possible difference between categories represented by the 21:19 ratio to the highest ratio of 36:4 (Figure 1C).The inclusion of the latter "easy condition" served a dual purpose. Firstly, it allowed us to evaluate participants' engagement in the experiment, as they were expected to perform at a near-ceiling level in the 36:4 ratio. Secondly, we expected that the introduction of such an easy condition could mitigate potential frustration or uncertainty experienced by participants in near-threshold trials, thereby enhancing overall motivation throughout the experiment. We note that the 36:4 ratio was only included after the first five participants, which were initially recruited as pilots. Consequently, the analyses for the 36:4 condition were conducted with a sample size of ten participants, instead of the fifteen participants involved in the analysis of the other ratios. Besides, to prevent any processing facilitation due to repetition, each item could only appear once on each interval of a given trial.

Every participant completed a total of 1000 trials in two separate 60-minute sessions. On each trial, the informative ratio conditions were randomly selected, leading to a slightly varied total number of trials for each condition among participants. However, this variance posed no significant issue, as our analyses were conducted at the group level, and the variability in the number of trials among participants was minimal, as demonstrated in Supplementary Table 2. Additionally, on average all participants completed at least 96 trials of each condition.

Analysis

For each participant and ensemble ratio we calculated: (i) The probability of being correct in the type-1 and type-2 tasks, (ii) The probability of being correct in the type-1 task conditional on type-2 accuracy (i.e., when participants made a correct vs incorrect type-2 response), (iii) Signal detection theory (SDT) measures including type-1 sensitivity regarding ensemble discrimination sensitivity) and the type-2 sensitivity regarding how well the confidence forced choice discriminates the task relevant information.

Signal detection theory measures (SDT). In assessing perceptual sensitivity, namely type-1 ensemble discrimination sensitivity, we calculated all SDT trial types based on the informative interval. In this context, a 'hit' referred to trials where participants correctly responded *living* when a *predominantly living ensemble* stimulus was presented. Conversely, instances where participants provided *nonliving* responses after the presentation of a *predominantly living ensemble* were categorized as 'misses.' 'False alarms' (FA) were registered when participants gave *living* responses when the predominant target class was *nonliving*. 'Correct rejections' (CR) were noted when participants correctly identified *nonliving* responses in predominantly *nonliving ensembles*. We adopted Hautus' log-linear correction method 1995 to calculate the hit rate (HR) and false alarm rate (FAR), considering extreme proportion values:

In assessing ensemble discrimination sensitivity, namely type-1 sensitivity, we calculated all SDT trial types based on the informative interval. In this context, a 'hit' referred to trials where participants correctly responded 'living 'when a predominantly living ensemble stimulus was presented. Conversely, instances where participants provided 'nonliving' responses after the presentation of a predominantly living ensemble were categorised as 'misses.' 'False alarms' (FA) were registered when participants gave living responses when the predominant target class was nonliving. 'Correct rejections' (CR) were noted when participants correctly identified nonliving responses in predominantly nonliving ensembles. We adopted Hautus' log-linear correction method 1995 to calculate the hit rate (HR) and false alarm rate (FAR), considering extreme proportion values:

$$\mathrm{HR} = \frac{X + \mathrm{Hits}}{2X + \mathrm{Hits} + \mathrm{Misses}} \tag{1}$$

$$\mathrm{FAR} = \frac{Y + \mathrm{FAs}}{2Y + \mathrm{FAs} + \mathrm{CRs}} \tag{2}$$

Where X is the proportion of signal trials (Hits + Misses /All trials) and Y is the proportion of noise trials (FAs + CRs /All trials). Monte Carlo simulations demonstrated that the log-linear correction yields less biased estimates of sensitivity than the most-commonly used 1/(2N) rule (Hautus, 1995). Then we calculated the non-parametric sensitivity index A ' (Macmillan and Creelman, 2004).

$$\mathrm{A'} = 0.5 + 0.25 \times \frac{(\mathrm{HR} - \mathrm{FAR}) \times (1 + \mathrm{HR} - \mathrm{FAR})}{\mathrm{HR} \times (1 - \mathrm{FAR})} \tag{3}$$

For each participant we also computed the type-2 HR and FAR. According to SDT, type-2 hits refer to trials where participants correctly identified the informative interval following a correct type-1 response. In other words, they were accurate in discerning the majority class within the ensemble in the informative interval. Conversely, type-2 FA were recorded when participants correctly identified the informative interval but provided an incorrect type-1 response.

Further, we calculated the extent to which the type-2 responses discriminated the position of the informative from the non-informative interval using the non-parametric sensitivity index A'. Accordingly, hits were defined as type-2 responses associated with the first interval when the target also appeared in the first interval (here type-2 responses associated with the second interval were defined as misses). FAs were trials in which type-2 responses were associated with the first interval but the target appeared in the second interval. Finally, CRs occurred when participants type-2 responses and the target belonged to the second interval.

Group-level statistical analyses. First, we assessed whether type-1 and type-2 processes responded differently to ensemble ratio by running a repeated measures ANOVA with task type and ensemble ratio as within subjects factors. Then, we repeated the same analyses separately for conditions with a high ratio of the predominant class (26:14 - 30:10), a low ratio of the predominant class (21:19 - 25:15) and all ratios (21:19 - 30:10). In all ANOVAs, sphericity criterion violations were corrected by means of the Greenhouse-Geisser correction. Additionally, we asked whether participants could successfully extract the most prominent animacy category in the ensemble while lacking awareness of the task relevant features, particularly when the relevant stimuli ratio was low but about 0.5. To study this, we run one-sample non-parametric bootstrapping t-tests to calculate the p-value of the difference

between participants' performance and an empirically estimated chance-level performance. The bootstrapping t-test involved different steps; first of all, participants' mean performance distribution was centered to chance-level (0.5). Then, we sampled 100% of the chance-centered-data with replacement and estimated the distance of the sampled data from chance. We repeated the same process 10,000 times to simulate an empirical null distribution of the difference. Finally, we calculated the p-value as the probability of encountering a value from the null distribution greater or equal to the distance of participants' actual performance from chance. In addition to the p-value, in order to have a better understanding of the evidence in favor of the null hypothesis, we also calculated the Bayes Factor (BF) of the difference of participants' performance from chance. We used the default Cauchy distribution prior (0.707) for the BF analysis, and robustness checks (Lee and James Press, 1998) confirmed that our results were not influenced by the choice of prior. All bootstrapping t-tests and BF analyses were run independently for every task and ensemble ratio combinations, always excluding the control condition 36:04 from the analysis. We used JASP (Goss-Sampson, 2019) for ANOVA and BF analyses.

Data availability

All code and data are openly available at the following link https://osf.io/whr2n/?view$_o nly =$ $eb32844bc86a42b2b7e41ee1e48bacc6$ (Elosegi and Soto, 2023).

Results

First, as expected, we found that ensemble ratio had a significant impact on both type-1 ($F(10,139)=12.842$, $p < .001$, $\eta_2=.463$) and type-2 discrimination performances ($F(10, 139)=14.023$, $p < .001$, $\eta_2=.485$) (see Figure 2A), showing that both increase as the target ensemble ratio increases. Additionally, Figure 2D shows that while HR significantly increased with ensemble ratio ($F(1, 12.54) = 45.172$, $p < .001$) the FAR tended to reduce although this was not significant ($F(1, 10.36) = 1.165$, $p = .305$). In line with this observation, we found that participants were more likely to discriminate the predominant category in trials in which the type-2 response was associated with the informative interval compared to the non-informative interval ($F(1, 14) = 42.24$, $p < .001$, $\eta_2=.149$). This indicates that awareness of the informative interval was associated with better discrimination performance of the ensemble (see Figure 2G). Taken together, these results suggest that observers were correctly following task instructions.

The most relevant question of Experiment 1 was whether ensemble perception can be dissociated from awareness of the task relevant features. To study this question we used a combination of non-parametric bootstrapping t-test and Bayes Factor (BF) analysis to

assess for each level of ensemble ratio whether participants' type-1 accuracy and type-2 performance were better than chance. Specifically, if ensemble perception and awareness of the task relevant features are dissociable we should expect above-chance type-1 performance with chance-level type-2 performance. We found exactly this pattern of results for conditions 21:19 and 22:18, in which type-1 accuracy was significantly above chance (**type-1$_{\mathbf{21:19}}$**: $\overline{X}_{\text{Acc}}$ = .55, p < .001, BF_{10} = 30.27; **type-1$_{\mathbf{22:18}}$**: $\overline{X}_{\text{Acc}}$ = .59, p < .001, BF_{10} > 100) while type-2 was at chance (**type-2$_{\mathbf{21:19}}$**: $\overline{X}_{\text{Acc}}$ = .49, p = .821, BF_{10} = .15; **type-2$_{\mathbf{22:18}}$**:$\overline{X}_{\text{Acc}}$ = .51, p = .250; BF_{10} = .469). Importantly, these observations were not due to a biased response criterion since A' analyses, yielded exactly the same results for the same ensemble ratio conditions (**type-1$_{\mathbf{21:19}}$**: $\overline{X}_{\text{A'}}$ = 0.59, p < .001, BF_{10} > 30; **type-1$_{\mathbf{22:18}}$**: $\overline{X}_{\text{A'}}$ = 0.66, p < .001; BF_{10} > 100; **type-2$_{\mathbf{21:19}}$**: $\overline{X}_{\text{A'}}$ = 0.47, p = .933, BF_{10} = 0.12; **type-2$_{\mathbf{22:18}}$**: $\overline{X}_{\text{A'}}$ = 0.50, p = .451, BF_{10} = 0.30) (see Figure 3A). Despite the observation that type-1 and type-2 performance responses were sensitive to the ensemble ratio, we still wanted to know whether both measures responded to this ratio in a similar way. Repeated measures ANOVA run for all levels of ensemble ratio revealed a task x ratio interaction (F(9, 261) = 11.77, p < .001, η_2=.024). Further analyses splitting by the ratio (low vs high), showed that this effect was observed exclusively in conditions with a low ratio of the predominant class (F(3.05, 88.69) = 12.928, p < .001, η_2=.046). In contrast, conditions characterized by a high ratio of the predominant class did not exhibit this effect (F(4, 116) = 0.254, p = .9, η_2=5.415e-4). This difference is represented in Figure 3A by a linear increase in type-1 performance accompanied by a type-2 performance plateau between conditions 22:18 to 25:15.

Discussion

In Experiment 1 we characterized the role of awareness of the task relevant features in ensemble perception. Even when participants' type-2 responses discriminated the informative interval at chance levels, the identification of the predominant semantic category of the ensemble was significantly above chance (i.e. with a 59% of accuracy). These results confirm our hypothesis that ensemble perception can occur without awareness of the task-relevant features (i.e. the ensemble ratio that defines the ensemble). Interestingly, there was an interaction between task type and ensemble ratio for the lower ratios. This observation suggests that, at least when the task relevant signal is small, the function linking type-1 accuracy to type-2 discrimination of the informative interval is clearly nonlinear. As an example of this, figure 3A reveals that across conditions 23:17, 24:16 and 25:15, awareness of the informative interval remained stationary while type-1 accuracy continued increasing linearly. This indicates that here participants were introspectively blind to changes in type-1 accuracy when the level of performance was lower. Thus, it could be argued that perceptual

awareness of the ensemble for conditions 23:17, 24:16 and 25:15 was indistinguishable.

Here, ensemble perception was assessed using a two-alternative forced-choice task (2AFC) involving the discrimination of the predominant class in a temporal sequence of living and non-living items. Despite deviating from the traditional averaging tasks, this approach aligns with the operational definition of ensemble perception provided by Whitney and Leib (2018). It was impossible to use an averaging task in the context of the two-interval confidence forced-choice task used here (see also Peters and Lau, 2015), because it requires a target absent interval and in averaging tasks a signal is continuously present as long as an image is displayed. However, our two-interval paradigm required a signal-absent interval to assess detection performance. Consequently, we decided to focus on the predominant class discrimination which allows to effectively generate a signal-absent interval by equating the number of items from both classes in a 1:1 ratio.

Experiment 2: Assessing the role of feedback

In Experiment 1 we find a dissociation between ensemble perception and awareness (i.e. of the task relevant features). One possible limitation of the Experiment 1 is that participants were not explicitly informed about the presence of both an informative and a non-informative interval that contained no task relevant signals (Peters and Lau, 2015). Experiment 2 used the same 2IFC paradigm as in Experiment 1 but this time participants were given feedback on whether the type-2 response was associated with the informative interval with the aim of directing their attention to the relevant information. Hence, the aim of Experiment 2 was twofold: (i) assess the replicability of the results from Experiment 1 and (ii) test if the observed dissociation between awareness and ensemble perception can be modulated by directing participants' attention to the task relevant information.

Methods

Transparency and openness

The experimental data and scripts are available at OSF in the following link https://osf.io/whr2n/?vi*eb*32844*bc*86*a*42*b*2*b*7*e*41*ee*1*e*48*bacc*6 (Elosegi and Soto, 2023). The experiments were not pre-registered.

Participants

For the Experiment 2, an independent sample of fifteen participants (12 women, $\overline{X}_{\text{Age}} = 23$, $\text{SD}_{\text{Age}} = 5.3$ years) was recruited from the BCBL´s online platform in return of monetary reward (8 €/hr). Considering that the same statistical power considerations apply

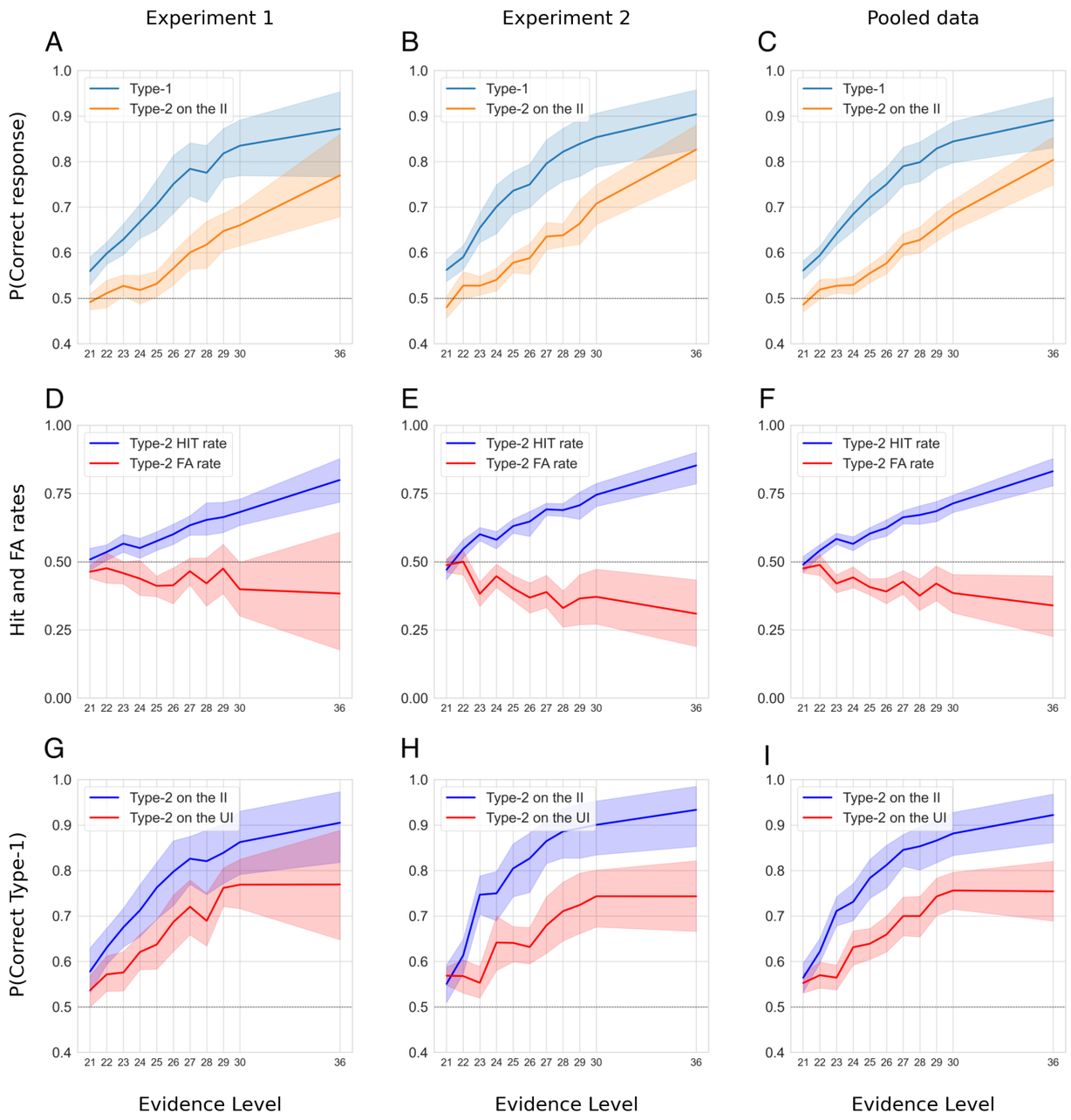

Figure 2: ***Results from Experiment 1 and Experiment 2****. Columns one and two present the results from experiments 1 and 2 respectively and the third represents results after pooling the data from both experiments together. The first row (panels A, B, C) shows type-1 and type-2 accuracy for all ensemble ratios. In both experiments type-2 performance was at chance for the levels of minimal evidence. Note that type-2 performance here represents participants' accuracy detecting the informative interval. The second row (panels D, E, F) depicts the type-2 hit and false alarm rates (i.e., HR and FAR) for all ensemble ratios. The third row (panels G, H, I) shows participants type-1 accuracy conditional upon type-2 responses regarding the informative interval. The separation between both curves can be taken as a proxy of participants' metacognitive performance (the ability to discriminate correct from incorrect responses) which clearly improved as the ensemble ratio increased. The ensemble ratio is represented by the* antecedent *term of the ratios (e.g., A in A:B) which are the number of items in favor of the predominant category within the 40 item sequence. The* consequent *term of the ratio can be easily calculated subtracting the antecedent term to 40 (e.g, the consequent term of 23 is 17, 23:17). Error shades are the 95% CIs.*

to Experiment 1 and 2 and given the minimal difference in the experimental procedure, we kept the sample size identical to Experiment 1. All of them were right-handed, had normal or corrected to normal vision and gave informed consent to take part in the experiment. The protocol was approved by the BCBL Ethics Review Board and complied with the guidelines

of the Helsinki Declaration (2008). Data were acquired during the spring of 2022.

Apparatus and stimuli

All materials and apparatus were the same as in Experiment 1.

Procedure

For Experiment 2 everything was maintained the same as in Experiment 1 except that participants received feedback conditional upon type-2 responses on the informative interval (Figure 1A). Specifically, positive feedback was given when participants were more confident on the informative interval whereas negative feedback was given when they were more confident in the information absent interval. In order to be able to complete the task, participants were explained that only one of the intervals contained information for making the task. Also as before, participants completed 1000 trials in two separate sessions.

Analysis

To assess the replicability of the results from Experiment 1, we repeated the same analyses described above with the data from Experiment 2. Additionally, to study the effect of feedback, we run repeated measures ANOVA taking feedback presence as between-subjects factor and the ensemble ratio as within-subjects factor to compare participants' type-1 and type-2 accuracy and sensibility between both experiments. We run these analyses separately for conditions with a low ratio of the predominant class (21:19 – 25:15), a high ratio of the predominant class (26:14 – 30:10) and all ratios (21:19 – 30:10).

Data availability

All code and data are openly available at the following link https://osf.io/whr2n/?view$_o nly$ = $eb32844bc86a42b2b7e41ee1e48bacc6$ (Elosegi and Soto, 2023).

Results

Participants correctly followed task instructions. Both HR and FAR significantly correlated with ensemble ratio but in opposite directions (HR: $F(1, 27.28) = 248.692$, $p < .001$; FAR: $F(1, 12.28) = 10.469$, $p = .007$) and type-1 accuracy was higher when participants type-2 responses belonged to the informative interval compared to the non-informative interval ($F(1, 14) = 137.88$, $p < .001$, η_2=.260) (see Figure 2E). Interestingly, we found that feedback administration maximized the difference in type-1 performance between correct and incorrect type-2 responses ($F(1, 28) = 5.62$, $p = .025$, η_2=.005) (i.e., trials in which the type-2 response was placed on the informative vs non-informative interval) which can

be appreciated in the bottom row of Figure 2G-I as a smaller overlap between lines in the Experiment 2 plot compared to Experiment 1. This feedback effect also manifested as an interaction between the ensemble ratio and type-2 correctness on participants' type-1 performance ($F(5.83, 81.72) = 9.025$, $p < .001$, $\eta_2 = .063$).

In Experiment 2 we replicated the dissociation between ensemble perception and awareness observed in the first experiment for condition 21:19 (**type-1$_{\mathbf{21:19}}$**: $\overline{X}_{\text{Acc}} = .56$, $p < .001$, $BF_{10} > 100$; **type-2$_{\mathbf{21:19}}$**: $\overline{X}_{\text{Acc}} = .48$, $p = .930$, $BF_{10} = .12$) but not for condition 22:18 or others (**type-1$_{\mathbf{22:18}}$**: $\overline{X}_{\text{Acc}} = .59$, $p < .001$; $BF_{10} > 100$; **type-2$_{\mathbf{22:18}}$**: $\overline{X}_{\text{Acc}} = .52$, $p = .03$; $BF_{10} = 1.610$). Sensitivity analyses nonetheless, showed this dissociation for both conditions (**type-1$_{\mathbf{21:19}}$**: $\overline{X}_{\text{A'}} = 0.61$, $p < .001$, $BF_{10} > 100$; **type-2$_{\mathbf{21:19}}$**: $\overline{X}_{\text{A'}} = 0.43$, $p = .980$, $BF_{10} = .103$; **type-1$_{\mathbf{22:18}}$**: $\overline{X}_{\text{A'}} = 0.64$, $p < .001$; $BF_{10} > 100$; **type-2$_{\mathbf{22:18}}$**: $\overline{X}_{\text{A'}} = 0.54$, $p = .06$; $BF_{10} = 1.181$) (see Figure 3B). Similarly, we also found the same interaction between task type (ie, type-1 vs type-2) and ensemble ratio (ie, 21:18-30:10) ($F(9, 261) = 11.77$, $p < .001$, $\eta_2 = .024$). As in Experiment 1 this effect occurred when conditions with a low ratio of the predominant class were considered ($F(3.05, 88.69) = 12.92$, $p < .001$, $\eta_2 = .046$) but not for conditions with a high predominant class ratio ($F(4,116) = 0.254$, $p = .907$, $\eta_2 = 5.415e\text{-}4$). There were no significant differences in type-1 or type-2 accuracy or sensitivity between Experiments 1 and 2 (**type-1$_{\mathbf{Acc}}$**: $F(3.21, 89.88) = 0.69$, $p = .56$, $\eta_2 = .003$; **type-2$_{\mathbf{Acc}}$**: $F(5.12, 143.75) = 0.83$, $p = .53$, $\eta_2 = .009$; **type-1$_{\mathbf{A'}}$**: $F(3.46, 97.10) = 0.52$, $p = .691$, $\eta_2 = .003$; **type-2$_{\mathbf{A'}}$**: $F(5.39, 151.08) = 1.22$, $p = .3$, $\eta_2 = .015$).

Discussion

In Experiment 2 we found that the dissociation between ensemble perception and conscious awareness (i.e. of the task relevant features of the stimulus sequence) found in Experiment 1 can be replicated even when participants are explicitly oriented towards the informative interval and they are given feedback on their type-2 performance. Additionally, we also replicated the same interaction between task type (type 1 vs type 2) and ensemble ratio from Experiment 1, occurring for low- but not for high predominant class ratios, thereby reinforcing the observation of a non-linear relationship between ensemble perception and awareness of the informative interval. We found that feedback administration affected the adequacy of participant's type-2 discrimination responses. When participants selected the informative interval, type-1 accuracy was better than when they selected the non-informative interval, and this effect was enhanced by feedback. However, feedback did not modulate type-2 performance (i.e. whether participants' type-2 responses discriminated between the informative interval vs the non-informative interval) (Haddara and Rahnev, 2022; Peters and Lau, 2015). It could be that feedback amplified the task relevant information contained

in the informative interval, rather than merely facilitating the detection of the informative interval.

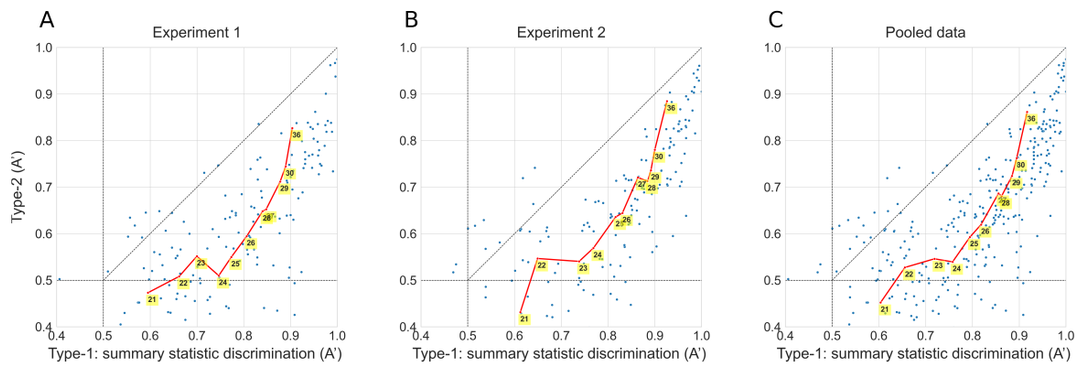

Figure 3: ***Type-1 sensitivity at discriminating the predominant category of the ensemble across ensemble ratios vs. discrimination of the informative interval based on type-2 responses (i.e. type-2 A')****. Each dot represents a participant's sensitivity discriminating the predominant object category in the ensemble (i.e., type-1 task) vs discriminating the informative interval (i.e., type-2 task). The red line represents the group-mean for each ensemble ratio. Ensemble ratios are represented by the* antecedent *term of the ratios (e.g., x in x:y). Panel A, B and C depict the results from experiments 1, 2 and the pooled data respectively. In all cases, participants sensitivity discriminating the informative interval was at chance for the minimal ensemble ratio conditions indicating that introspectively for participants both the informative and non-informative intervals conveyed the same amount of information and thus they were not aware of the task relevant information.*

Experiments 1 and 2: Joint drift-diffusion and type-1 response classification analyses

Given that the results from Experiment 1 and Experiment 2 were virtually the same, we decided to pool together the data from both experiments, to better understand the dissociation between type-1 and type-2 performance. More specifically, we were interested in three questions.

It remains unclear how many items can be efficiently integrated during ensemble perception (Whitney and Leib, 2018), with studies suggesting two items (Allik et al., 2014; Maule and Franklin, 2016) and 4-8 items (Haberman and Whitney, 2009; Leib et al., 2016; Sweeny et al., 2015). Studies of temporal ensemble perception reported the temporal weighting of stimulus sequences in the form of primacy and recency effects, such that the first- or last-seen objects biased the estimated ensemble property (Hubert-Wallander and Boynton, 2015). Here we assessed sampling strategies across participants, by using multivariate pattern analysis (MVPA) models to find which items in the sequence contributed more to participant's perceptual decisions.

Second, considering that our type-1 perceptual task deviates somewhat from traditional ensemble perception tasks (See Experiment 1 - Discussion), we also used multivariate pattern analysis to assess the degree to which participants' responses were affected by the

lifelikeness of the items within the ensemble, and thereby, validate our task as an ensemble perception task.

Third, we employed drift diffusion models (DDM) (Ratcliff and McKoon, 2008) to complement the SDT analysis by assessing the efficiency of type-1 and type-2 processes and how they relate. DDMs characterize decision-making as a two-choice task where each choice is represented by an upper- or lower-threshold. A drift process accumulates evidence over time until it crosses one of the two thresholds and commits to a response. For this study, we used a Pythonic implementation of the Hierarchical Drift Diffusion Model (HDDM) which uses Bayesian parameter estimation methods for enhancing the statistical power of the model (Wiecki et al., 2013). Compared to non-hierarchical methods, the HDDM requires fewer data per subject/condition, allows for the simultaneous individual and group-level parameter estimation, and provides a measure of uncertainty through parameters' posterior distributions (Wiecki et al., 2013). We used this information to characterize the type-1 and type-2 evidence accumulation process for each ensemble ratio.

Methods

Classification analysis 1. Previous research has shown limited consensus on the efficiency and temporal weighing of information integration in ensemble perception (Whitney and Leib, 2018). To assess participants' sampling in our experiments, we trained L2 penalty logistic regression classifiers (LRC) to predict type-1 responses (i.e., living or nonliving) based on the animacy of the items in the informative interval. Stimulus sequences were binarised according to the objects' animacy class (i.e., living = 1, non-living = 0). Classifiers were trained using 100-fold Stratified-Shuffle-Split cross-validation, preserving the original dataset's class distribution. In each iteration, they learned from 80% of the data and tested on the remaining 20%, yielding one classification ROC-AUC score per iteration.

Considering the possibility of idiosyncratic response strategies across participants and ratios, the classification analyses were conducted within-subjects and within each ensemble ratio conditions. To assess the statistical significance of the classification scores, for each participants we averaged the classifiers' performance to obtain a group-level classification distribution that was compared to an empirically derived chance-level distribution. This distribution was estimated by randomly shuffling the target vector of the training set during each cross-validation iteration.

Since logistic regression was employed, we extracted odds ratios as a measure of each predictors' importance on each iteration. Odds ratios offer direct interpretability by quantifying the change in odds of an event occurring for a one-unit change in the predictor variable and thus, can be used to reveal the position of the most relevant items in the

sequence regarding participants' responses. Consequently, if participants responded based on a primacy/recency effect, we should find higher odds ratios for the first/last items of the sequence. Alternatively, if participants were sampling items rhythmically, we should find a subset of distributed peaks of odds ratios along the sequence (see Figure 4A).

Classification analysis 2. In the present study ensemble performance was assessed using a two-alternative-forced-choice task 2AFC involving the discrimination of the most frequent predominant class in the sequence. While this approach differs from conventional ensemble perception tasks (see Experiment 1 - Discussion), we contend that it relies on the same averaging mechanism characteristic of ensemble perception.

To investigate this hypothesis, we gathered subjective lifelikeness ratings for all 96 single-object stimuli used in Experiments 1 and 2 following a similar procedure as Leib et al. (2016). Twenty independent volunteers participated in this rating process (see Supplementary materials - Classification analysis 2), where each object was displayed for one second, and participants rated its lifelikeness on a scale from one to ten. We calculated the lifelikeness score for each item as the average rating across all twenty participants (see Supplementary Figure 5). Subsequently, we trained logistic regression classifiers (LRC) using the lifelikeness ratings of the forty items within each sequence to predict participants' type-1 responses.

Notably, this analysis was conducted across ten different conditions based on the *number of living items* (ranging from 15 to 25 living items), as opposed to the ratio conditions in Classification Analysis 1. This was done because at the ratio level, half of the trials would have a predominance of living things and the other half a predominance of non-living things, and thus, would not allow to disentangle the effect of the number of living items from the pure lifelikeness of the sequence.

Given that the primary aim of this analysis was not to uncover individual participant strategies, classifiers were trained for each condition including all participants' data in a leave-one-participant-out cross-validation protocol, whereby on each iteration one participant is excluded from the training set and the classifier is tested on the data of the unseen participant. We also note that we had an average of 50 trials on each number of living conditions participant which is not sufficient for within-participant within-condition classification analysis due to the small number of observations relative to the number of features - an issue commonly referred as the curse of dimensionality (Tibshirani et al., 2017).

Classification analyses 1 and 2 were implemented in Python using the Scikit-Learn package (Pedregosa et al., 2012).

HDDM analysis. We fitted two separate HDDM models with trial-by-trial type-1 and type-2 accuracy vectors and response times (See Supplementary Table 3 for a complete

description of RTs). The HDDM uses Markov Chain Monte Carlo (MCMC) as an inference algorithm to estimate the posterior distributions of the model's parameters, which is a computationally costly process. So, to speed up Markov chain convergence we set up the starting point of the MCMC to the maximum a-posterior value, which was estimated through gradient ascent optimization. Then, we drew 2000 samples to estimate the posterior distributions of the parameters and we discarded the first 500 instances as burn-in samples. These initial samples represent the 'heat-up steps' that the MCMC algorithm requires to reach a stationary sampled distribution within an acceptable error, which is known as convergence. Prior research has shown that, when estimating HDDM parameters, getting rid of the initial 20-1000 samples is usually enough for the Markov chains to converge (Wiecki et al., 2013). Nonetheless, we statistically tested for convergence running 6 independent Markov chain simulations and calculating the Gelman-Rubin $\hat{R}$ statistic for all sampled parameters (Gelman and Rubin, 1992).

The drift-rate is the main parameter of interest extracted from a DDM model and represents the rate of evidence accumulation which is determined by the quality of the information extracted from the stimulus (Ratcliff and McKoon, 2008). Consequently, we used the trained HDDM to extract the posterior distribution of the drift-rate for each participant and ensemble ratio. For statistical analysis however, we only took the mean of the posterior of each participant to generate a group-level distribution of drift-rates for each ensemble ratio. In the context of the present study, we expected that the ensemble ratio would positively correlate with drift-rate values (ie, the greater the ratio for a given stimulus class, the greater the drift-rate). Additionally, we also expected that on average drift-rates would be greater for the type-1 task than for type-2 task. To assess these predictions, we run repeated measures ANOVA with task type and ensemble ratio as within subject factors.

Results and discussion

Pooling data from Experiments 1 and 2 enabled us to assess with greater statistical power the two main behavioural effects from these experiments, namely, the isolation of ensemble perception from awareness of task relevant features and the interaction between type-1 and type-2 tasks and ensemble ratio. Regarding the former, we only found the dissociation for condition 21:19 (**type-1$_{\mathbf{21:19}}$**: $\overline{X}_{\text{Acc}} = .56$, $p < .001$, $BF_{10} > 100$; **type-2$_{\mathbf{21:19}}$**: $\overline{X}_{\text{Acc}} = .48$, $p = .965$, $BF_{10} = .07$) because for condition 22:18 type-2 accuracy was already slightly above chance but with only anecdotal evidence in favour of the alternative hypothesis (**type-1$_{\mathbf{22:18}}$**: $\overline{X}_{\text{Acc}} = .59$, $p < .001$, $BF_{10} > 100$; **type-2$_{\mathbf{22:18}}$**: $\overline{X}_{\text{Acc}} = .51$, $p = .037$, $BF_{10} = 1.321$). Running the same analyses based on A' which is a non-biased sensitivity measure, revealed that the dissociation was present for both ensemble ratios (**type-1$_{\mathbf{21:19}}$**:

$\overline{X}_{A'}$ = 0.60, p < .001, BF_{10} > 100; **type-2$_{21:19}$**: $\overline{X}_{A'}$ = 0.45, p = .992, BF_{10} = .063; **type-1$_{22:18}$**: $\overline{X}_{A'}$ = 0.65, p < .001; BF_{10} > 100; **type-2$_{22:18}$**: $\overline{X}_{A'}$ = 0.52, p = .125; BF_{10} = .567). Considering all data together, we observed that the interaction between task and ensemble ratios only emerged for conditions with a low ratio of the predominant class (F(3.59, 88.69) = 12.92, p < .001, η_2=.046) but not for high predominant class ratios (F(4, 116) = 0.254, p < .907, η_2=5.415e-4). Apart from the main results, we found that FA and HIT rate significantly correlated with ensemble ratio (HR: F(1, 28.53) = 170.191, p < .001; FAR: F(1, 28.75) = 10.682, p = .003) demonstrating that in general participants correctly followed task instructions by shifting the response criterion according to the signal intensity defined by the ensemble ratio.

These results suggest that ensemble perception can unfold without awareness of the task-relevant features. However, as shown in Figure 3C we also observe that type-2 performance was systematically lower than type-1 performance across all the ratios. This brings into question whether the observed dissociation is actually reflecting unconscious perception or if alternatively, it is produced by an inherent inefficiency of type-2 compared to type-1 performance (see Michel, 2022 for a contextualization of the problem). To address this issue, we fitted a Gaussian Process regression model (GPR) on the pooled data to predict participants' type-2 sensitivity based on type-1 sensitivity (see Supplementary materials - Gaussian process regression analysis 1)). This model was exclusively trained on the ratios in which type-1 performance exceeded 0.70 (see Figure 3C), allowing to learn the relationship between type-1 and type-2 performance when participants were consciously aware of the ensembles (Michel, 2022). Then, we use this model to predict the level of type-2 sensitivity in ratios 21:19 and 22:18. Importantly, these predictions reflected the expected type-2 sensitivity if observers were conscious of all the sensory information used to perform the type-1 task. If participants' type-2 sensitivity was lower than predicted by the model, then it would suggest that the dissociation between type-1 and type-2 sensitivity is not explicable by decision-making or type-2 inefficiencies alone, and thereby, it would argue in favour of unconscious perception.

One sample t-tests revealed that in condition 21:19 participants sensitivity was significantly lower than predicted by the model (t(29) = -3.42, p <.01, d = -0.62) while for condition 22:18 this difference did not reach significance (t(29) = -0.97, p = .34, d = -0.18) (Supplementary Figure 1). Taking the mean and standard deviation of the posterior distributions provided by the GPR model we simulated 1000 data points assuming a normal distribution for each ratio to assess if the models' predictions were significantly above chance. The model predicted that type-2 sensitivity would be above chance in both ratios (**predicted type-2$_{21:19}$**: $\overline{X}_{A'}$ = 0.52, t = 11.05, p < .001, BF > 100; **predicted type-2$_{22:18}$**: $\overline{X}_{A'}$ =

0.55, t = 38.5, p < .001, BF > 100), while the real data showed that on both cases the type-2 sensitivity was at chance level. Taken together, these results indicate that the dissociation between type-1 and type-2 cannot be explained in terms of type-2 inefficiency in decision making (Michel, 2022). Instead, these results support the view that performance in the critical 21:19 ratio was driven by unconscious perception.

We also note here that in the 21:19 condition, the type-1 task may have relied on a difference of two items between the predominant and non-predominant classes, while the type-2 task may involve the comparison between two sequences that only differed in one item regarding the predominant class (i.e., 21 vs 20) (see Allik et al., 2014 for a formalisation of this issue). Accordingly, the type-1 task would rely on a ratio of 0.526 while the type-2 task would rely on a ratio of 0.512. This would also apply to other ratios. For instance, in the 22:18 condition, the type-1 task would have four items of difference (0.55) while the type-2 would only have two (0.523). The above reasoning, however, does not take into account that participants do not know in advance which class is likely to be predominant and whether it appears on the first or second intervals. Hence, participants are likely to monitor both classes (not just one) across the two different intervals in order to gain information relevant for the type-2 judgments. This would equalize type-1 and type-2 judgments in terms of the amount of information available for decision-making.

In any case, to rule out the possibility that the dissociation between type-1 and type-2 performance could be driven by a systematic disadvantage of the type-2 relative to the type-1 task, we ran another GPR analysis to interpolate the expected A' values for both type-1 and type-2 tasks across the ratio conditions (see Supplementary materials - Gaussian process regression analysis 2)). Importantly, we found that all interpolated points for type-1 performance were significantly above chance, thereby providing confirmatory evidence to our initial observations. However, as revealed by one sample t-tests, in the crucial ratios 0.51 and 0.52, type-2 task performance was not significantly different from chance (**interpolated type-2$_{0.51}$**: $\overline{X}_{A'}$ = 0.46, t = -35.3, p = .99, BF < .001; **interpolated type-2$_{0.52}$**: $\overline{X}_{A'}$ = 0.49, t = -10.9, p = .99, BF < .001, further supporting our conclusions (Supplementary Figure 2).

Classification analysis 1. The objective of this analysis was to estimate the number and location of items that were primarily contributing to participants' type-1 responses. Figure 4B shows that we could classify better than chance type-1 responses in all conditions, even in the 21:19 condition ($\overline{X}_{\text{ROC-AUC}}$ = .53, t(58) = -2.18, p = .02, d = -0.56), although the ROC-AUC was barely above chance. Panels C-D from Figure 4 illustrate the odds ratios associated with each of the 40 items in the sequence for conditions 21:19 and 22:18 (see Supplementary Figure 3 and 4 for the rest of the ratios). The results indicate that the last

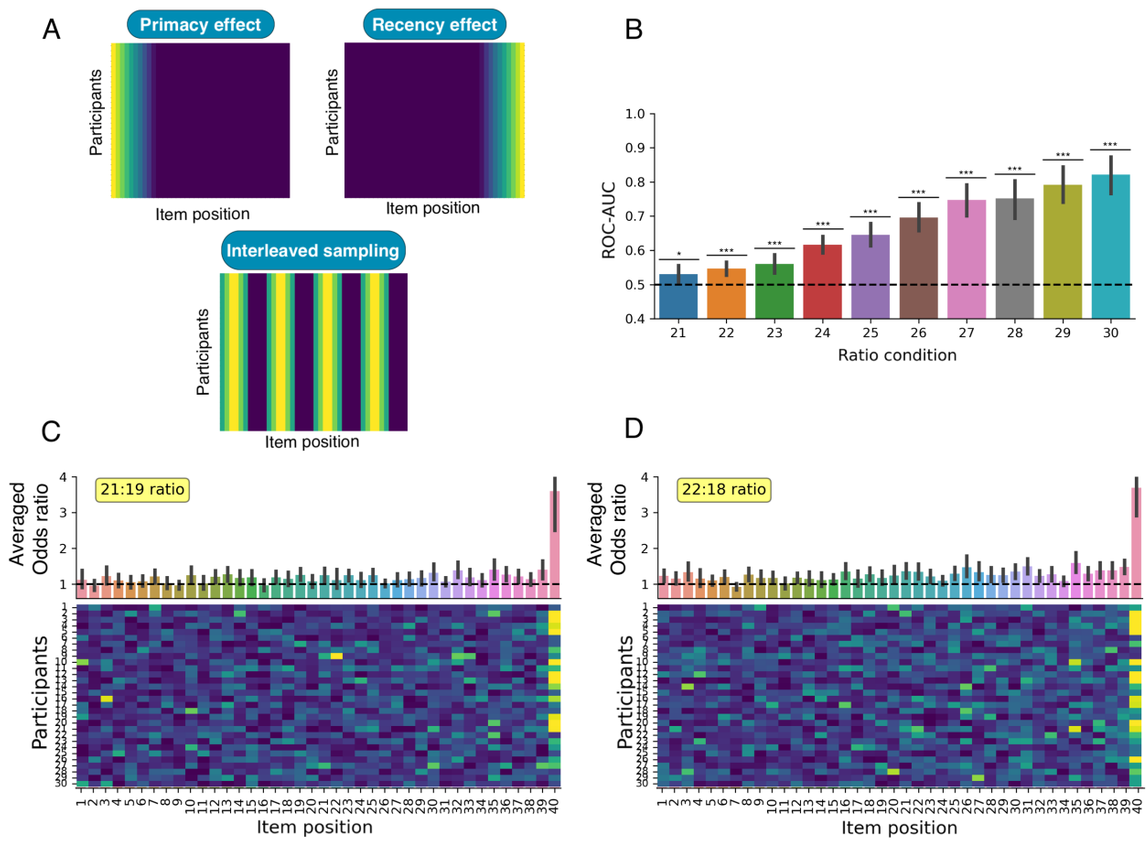

Figure 4: ***Type-1 response classification performance across ratios based on the binarised sequences***. *A) Theoretical examples of how recency/primacy or interleaved sampling would be reflected in the odds ratio patterns. B) Classification performance of participants' responses (i.e., living, nonliving), type-1 accuracy (i.e., correct, incorrect) and trial ground truth (i.e., living, nonliving) based on the binarised sequences of living and non-living items. C) Odds ratio results for condition 21:19. The heat-map matrix shows the odds ratio of the 40 items for each participant. Specifically, the* x *axes represent the position of the items in the sequence and the* y *axes represent the feature importance for each of the 30 participants. The barplot on the top represents the odds ratio results averaged across participants. D) Odds ratio results for condition 22:18. The odds ratios for the rest of the conditions can be found on Supplementary Figure 3 and 4. Error bars are the 95% CIs.* $* : p < .05, ** : p < .01, *** : p < .001$.

item of the sequence was the most relevant feature predicting type-1 responses for most participants across all ratios.

We conducted a post-hoc analysis to investigate whether the level of type-1 performance across different ratios could be attributed solely to a recency effect, where participants' decisions were influenced primarily by the animacy class of the last item in the sequence. To assess this, we computed the probability that the last item's class matched the trial ground truth for each participant and condition (see Figure 5). Crucially, our findings reveal that, across all ratios, and notably in the critical 21:19 ratio condition (t(58) = -1.73, p = 0.03, d = -0.448), type-1 accuracy significantly exceeded the probability that the last item's animacy corresponded to the trial ground truth. This suggests that while a recency effect may explain participants perceptual choices to some extent, additional information from earlier sections of the sequence also influenced their decisions. Notably, the recency effect that we observed is likely driven by the nature of our temporal ensemble task, in which late-stage

information is more important to determine the ground truth for a given trial, thus inflating the importance of the last item in the logistic regression analysis.

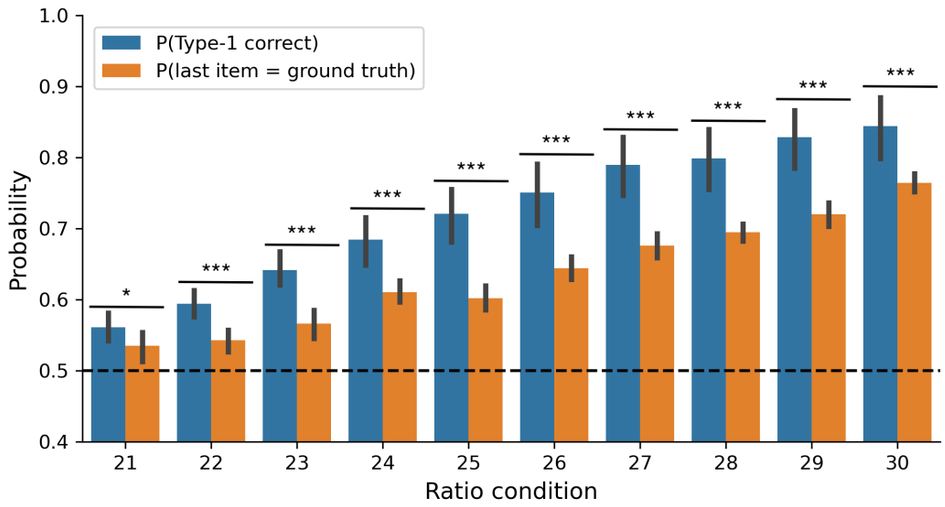

Figure 5: ***Probability of correct type-1 response compared to the probability that the sequences' last item matched the trial ground truth.*** *For each ratio, we calculated the probability that the last item presented was equal to the predominant class in the sequence and compared it against participant's performance. Error bars are the 95% CIs.* $* : p < .05, ** : p < .01, *** : p < .001$.

.

Our results are consistent with previous findings of recency effects in ensemble tasks related to average size, motion, and face perception (Gorea et al., 2014; Hubert-Wallander and Boynton, 2015). However, the study by Leib et al. (2016), using a comparable display, did not observe subsampling effects in an animacy average estimation task. Several factors may explain this discrepancy. First, Leib et al. (2016) used sequences with 12 items, while our sequences had 40 items. Ensemble perception, despite its efficiency, has capacity limitations (Attarha and Moore, 2015; Florey et al., 2017; Luo and Zhao, 2018). Thus, dealing with a larger amount of information can provoke memory leaks leading to recency biases (Gorea et al., 2014). Also, it is reasonable to presume that when continuously averaging information, the most recent items have a more prominent role in predicting and perceiving incoming information (Cheadle et al., 2014). Second, our study employed a two-alternative forced-choice task based on the discrimination of the predominant class, whereas Leib et al. (2016) used a continuous averaging task. Task differences may affect participants' response strategies and information integration. Third, variations in the analysis methods used to estimate sampling strategies could also contribute to differing results.

Classification analysis 2. The aim of the second classification analysis was to investigate whether the two-alternative-forced-choice task used in Experiments 1 and 2 (i.e., discriminating the most frequent predominant class within a sequence), relies on a similar averaging mechanism that is characteristic of ensemble perception. We trained logistic regression classifiers for ten different number of living conditions to predict participants type-1

responses based on the lifelikeness of the items in the sequence. Importantly, the examples were matched regarding the number of living items but could still differ in terms of the lifelikeness of the individual objects. We observed that participants responses were predicted significantly better than chance across all conditions (see Supplementary Figure 6). This pattern of results indicates that participant responses were indeed influenced by the lifelikeness of the items in the sequence, hence suggesting that participants integrated the lifelikeness information throughout the sequence to guide their choices. However, we acknowledge that our task could also rely to some extent on other processes such as subitizing (Wender and Rothkegel, 2000).

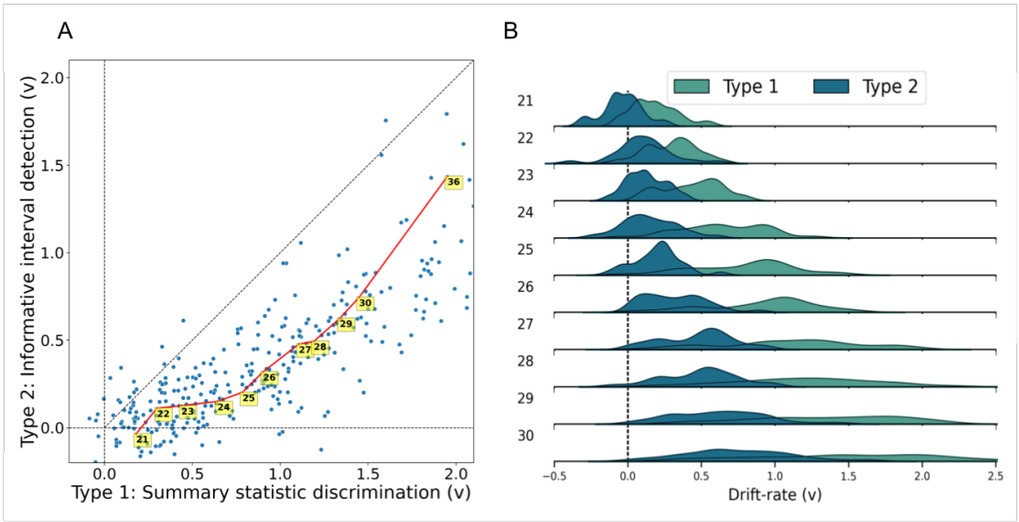

Figure 6: ***Type-1 vs Type-2 drift-rates across ensemble ratios.*** *A) Each dot represents a participant's evidence accumulation rate for type-1 (i.e., discriminating the predominant object category in the ensemble) vs type-2 task (i.e., discriminating the informative interval). The red line represents the group-level mean drift-rate for each ensemble ratio. B) type-1 and type-2 drift-rate posterior distributions for each ensemble ratio (group-level). Type-2 evidence accumulation rate for condition 21:19 was negative which means that participants were more likely to give an incorrect type-2 discrimination response. Ensemble ratios are represented by the* antecedent *term of the ratios (eg, x in x:y).*

HDDM analysis. For both type-1 and type-2 HDDM models, the Gelman-Rubin convergence test showed that the ratio of inter-chain to intra-chain variances for all the sampled parameters was close to 1 and never greater than 1.02, thus, verifying that the MCMC did find a stable posterior distribution. Regarding the drift-rate, we found that as expected, ensemble ratio had a significant effect on drift-rates ($F(2.99, 86.97) = 56.53$, $p < .001$, $\eta_2 = .253$) and that this effect on drift rate was significantly greater for type-1 than for type-2 accuracies ($F(1, 29) = 187.17$, $p < 0.001$, $\eta_2 = .288$). In line with prior results, we found a significant task x ensemble ratio interaction for drift-rates ($F(5.11, 148.39) = 26.98$, $p < .001$, $\eta_2 = .041$), that was present for both high ($F(4,116) = 2.52$, $p < .05$, $\eta_2 = .004$) and for low predominant class ratio conditions ($F(4, 116) = 25.83$, $p < .001$, $\eta_2 = .071$). This interaction is illustrated in Figure 6, showing that in conditions 22:18 – 24:16, there was a sharp type-1 drift-rate increase while the type-2 drift rate barely improved. Finally, Figure 6

also reveals that on average type-2 drift-rates in condition 21:19 were negative, which could suggest that participants were more likely to give incorrect than correct responses. However, these values did not reach statistical significance when compared to zero ($t(29) = -1.39$, $p = .087$, $d = -0.27$).

The HDDM results suggest that the type-1 and type-2 dissociation found in both experiments, alongside the interaction between task and ensemble ratio, was driven by differences in the evidence accumulation rate. More specifically, we found that type-1 information was accumulated faster and more accurately compared to the type-2 discrimination process. This is congruent with previous studies indicating that second-order processes (i.e. awareness and confidence) are less efficient compared to type-1 decisions (Pouget et al., 2016; Sanders et al., 2016).

General Discussion

In the present study we find that conscious awareness of the task relevant features is not a necessary prerequisite for ensemble perception. In a series of two psychophysical experiments using a criterion bias-free paradigm (Peters and Lau, 2015) we find that participants could discriminate the predominant semantic category better than chance without awareness of the informative interval: participants' ability to discriminate the informative from the non-informative interval was at chance level at lower ensemble ratios, yet perceptual identification of the more frequent class in the sequence was above chance level. Importantly, we provide compelling evidence that this dissociation was not driven by type-2 inefficiencies (Shekhar and Rahnev, 2021) or a systematic disadvantage in type-2 decision making (Allik et al., 2014). Instead, the observed dissociation between type-1 and type-2 performance at the lowest ratios is best explained in terms of unconscious ensemble perception.

We also observed that type-1 ensemble discrimination and type-2 performance were differently modulated by the ensemble ratio. Although both functions exhibited a monotonic pattern, type-1 performance displayed a linear trend characterised by a constant rate of change. In contrast, type-2 performance demonstrated a non-constant rate of change. The type-2 performance plateaus indicates that participants were introspectively blind to type-1 performance improvements associated with increasing ensemble ratios.

These results have important implications for understanding the mechanisms of ensemble perception (Whitney and Leib, 2018). Prior studies found that ensembles can be processed with minimal attention (Alvarez and Oliva, 2008; Bronfman et al., 2014; Chong and Treisman, 2005a), suggesting that ensembles are encoded pre-attentively Pascucci et al., 2021 and likely automatically (Alvarez, 2011).

However, whether visual awareness is necessary for ensemble perception has remained unclear because prior studies addressing this issue (Choo and Franconeri, 2010; Sekimoto and Motoyoshi, 2022) have predominantly relied on subjective measures of awareness (Eriksen, 1960; Lloyd et al., 2013), which are susceptible to criterion biases. Notably, some studies have shown that observers can accurately extract ensemble information even they are unable to report the identity of individual items within the ensemble (Haberman and Whitney, 2011; Leib et al., 2016; Parkes et al., 2001).

Our study goes beyond these findings by demonstrating that participants can utilize ensemble information even when they are unaware of the task-relevant features defining the ensemble. Crucially, in our investigation, the lack of awareness was established by chance level type-2 performance in a 2IFC paradigm that is known to mitigate confounding effects from criterion biases (Knotts et al., 2018; Mamassian, 2020; Peters and Lau, 2015). This results provide evidence that even higher-order ensembles can be encoded without awareness. This observation lends support to the hypothesis that dedicated mechanisms are involved in computing summary statistics (Alvarez, 2011).

More generally, the present study demonstrates the utility of robust measures for studying visual consciousness without underestimating the contribution of unconscious processing mechanisms. This is achieved by assessing awareness of the task relevant features as opposed to the mere detection of a strongly masked stimuli (Michel, 2022; Newell and Shanks, 2014). The latter can severely constraint the possibility of isolating behavioural markers of unconscious perception (Lau, 2022). This may be one of the reasons why prior studies using the 2IFC paradigm did not observe evidence of unconscious processing (Knotts et al., 2018; Peters and Lau, 2015; Peters et al., 2017).

In these studies, as soon as observers could discriminate better than chance the task relevant feature (i.e., a grating orientation) they could also detect the informative interval. However, awareness of something (i.e. detection) does not entail awareness of the task relevant feature and crucially, equating awareness to the mere detection of stimulus presence can lead to the underestimation of unconscious perception (for discussion on the criterion content fallacy see (Michel, 2022). To avoid this, the type-2 judgements should indicate whether observers were conscious of the task-relevant features. Accordingly, in the present study using visible sequences we showed that participants could discriminate the semantic category of the temporal ensemble, while being unaware of the ensemble ratio that determined which interval contained the perceptual ensemble. This is akin to a bias-free 'blindsight' effect (Weiskrantz, 1986), in which the absence of awareness of the task-relevant feature is accompanied by above-chance discrimination of the category of the ensemble. Importantly, this observation was replicated in two independent experiments.

However, we note that the dissociation between the efficiency of type-1 and type-2 performance across the different ratios could also reflect differences at the decisional level rather than at the visual awareness level. The signal, in this case, the ensemble ratio, might exert a greater impact on type-1 compared to type-2 decisions (Michel, 2022). For instance, Vlassova et al. (2014) reported that unconsciously processed information can modulate perceptual performance without altering metacognitive sensitivity. We propose that a similar phenomenon may have occurred in our experiments, whereby the unconscious integration of the sequence items guides type-1 perceptual decisions in the absence of awareness.

Furthermore, it is noteworthy that while 2IFC tasks efficiently address criterion-related concerns, they can impose a greater processing load on type-2 compared to type-1 processing (Peters and Lau, 2015). This is due to the necessity of monitoring two intervals of stimulation in order to perform the type-2 task. However, our Gaussian process regression analyses effectively ruled out the possibility that the observed dissociation between ensemble perception and awareness of the task-relevant features was solely driven by systematic type-2 disadvantages. Future studies employing the 2IFC paradigm should keep in mind this limitation (Michel, 2022) and should incorporate additional measures to ensure that any potential unconscious effects cannot be explained by systematic disadvantages, as demonstrated in the present study.

Finally, since visual perception goes beyond single-object recognition, we would like to emphasize the potential of ensemble perception paradigms to study visual awareness and metacognition in more ecological settings. For instance, characterizing type-2 processes associated with confidence computations in ensemble perception aligns with the need of studying metacognition in more complex tasks beyond the presentation of simple, isolated stimuli (Rahnev et al., 2022). Besides, ensemble perception provides a way of manipulating the stimulus uncertainty without degrading the low-level visual features. This may well provide a solution to tackle the distinction between non-conscious and conscious processing mechanisms without underestimating unconscious processing (Michel, 2022), which can have further ramifications in addressing unresolved theoretical issues concerning the operation of working memory on non-conscious input (Soto and Silvanto, 2016; Soto et al., 2011; Stein et al., 2016). Ensemble stimuli avoids the need of using masking or any other unconscious rendering technique to prevent awareness of the task-relevant features. Consequently, stimulus ensembles can provide a stronger internal signal for unconscious perception studies (Lau, 2022). Further, stimulus summary statistical representations are more robust to noise (Alvarez, 2011) than single-object representations thus they may be more suitable to study visual consciousness for unattended or peripheral presentations.

In summary, here we showed the contributions of unconscious and conscious pro-

cessing mechanisms to ensemble perception. Categorical ensemble perception can unfold unconsciously without awareness of the task relevant features (i.e. the ensemble ratio that defines the ensemble). Further, the results suggest that awareness lags well behind the categorization processes that support ensemble perception. We also note that these results have been observed in healthy young adults and it would be relevant for future work to address the role of awareness in ensemble perception across different neurodevelopmental trajectories, from childhood (Sweeny et al., 2015)into adulthood and later in aging.

The present work represents an attempt of studying consciousness and metacognition in more complex contexts without underestimating the contribution of unconscious processing in an otherwise conscious stimulus.

Acknowledgements

P.E. acknowledges support from the Basque Government PREDOC grant. D.S. acknowledges support from the Basque Government through the BERC 2022-2025 program, from the Spanish Ministry of Economy and Competitiveness, through the 'Severo Ochoa' Programme for Centres/Units of Excellence in R & D (CEX2020-001010-S) and also from project grants PID2019-105494GB-I00. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests statement

The authors declare no competing interests.

References

Allik, J. (1992). Competing motion paths in sequence of random dot patterns. *Vision Research*, *32*(1), 157–165. https://doi.org/10.1016/0042-6989(92)90123-Z

Allik, J., Toom, M., & Rauk, M. (2014). Detection and identification of spatial offset: Double-judgment psychophysics revisited. *Attention, Perception, & Psychophysics*, *76*, 2575–2583.

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

Alvarez, G. A., & Oliva, A. (2008). The Representation of Simple Ensemble Visual Features Outside the Focus of Attention. *Psychological science*, *19*(4), 392–398. https://doi.org/10.1111/j.1467-9280.2008.02098.x

Ariely, D. (2001). Seeing Sets: Representation by Statistical Properties [Publisher: SAGE Publications Inc]. *Psychological Science*, *12*(2), 157–162. https://doi.org/10.1111/1467-9280.00327

Attarha, M., & Moore, C. M. (2015). The capacity limitations of orientation summary statistics. *Attention, Perception, & Psychophysics*, *77*, 1116–1131.

Bai, Y., Yamanashi Leib, A., Puri, A., Whitney, D., & Peng, K. (2015). Gender differences in crowd perception. *Frontiers in Psychology*, *6*. Retrieved October 18, 2022, from https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01300

Barthelmé, S., & Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *107*(48), 20834–20839. https://doi.org/10.1073/pnas.1007704107

Bronfman, Z. Z., Brezis, N., Jacobson, H., & Usher, M. (2014). We See More Than We Can Report. *Psychological science*. https://doi.org/10.1177/0956797614532656

Cavanagh, P. (2001). Seeing the forest but not the trees. *Nature Neuroscience*, *4*(7), 673–674. https://doi.org/10.1038/89436

Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., De Gardelle, V., Castañón, S. H., & Summerfield, C. (2014). Adaptive gain control during human perceptual choice. *Neuron*, *81*(6), 1429–1441.

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

Chong, S. C., & Treisman, A. (2005b). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, *45*(7), 891–900. https://doi.org/10.1016/j.visres.2004.10.004

Choo, H., & Franconeri, S. (2010). Objects with reduced visibility still contribute to size averaging. *Attention, perception & psychophysics*, *72*, 86–99. https://doi.org/10.3758/APP.72.1.86

Cohen, M. A., Cavanagh, P., Chun, M. M., & Nakayama, K. (2012). The attentional requirements of consciousness. *Trends in Cognitive Sciences*, *16*(8), 411–417. https://doi.org/10.1016/j.tics.2012.06.013

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

Corbett, J. E., Utochkin, I., & Hochstein, S. (2023). *The pervasiveness of ensemble perception: Not just your average review*. Cambridge University Press.

Dakin, S. C., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research*, *37*(22), 3181–3192. https://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*(5), 789–794. https://doi.org/10.3758/PP.70.5.789

Elosegi, P., & Soto, D. (2023). Characterizing the role of awareness in ensemble perception. osf.io/whr2n

Eriksen, C. W. (1960). Discrimination and learning without awareness: A methodological survey and evaluation. *Psychological review*, *67*(5), 279.

Evans, S., & Azzopardi, P. (2007). Evaluation of a'bias-free'measure of awareness. *Spatial vision*, *20*(1), 61–78.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, *41*(4), 1149–1160.

Florey, J., Dakin, S. C., & Mareschal, I. (2017). Comparing averaging limits for social cues over space and time. *Journal of Vision*, *17*(9), 17–17.

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences [Publisher: Institute of Mathematical Statistics]. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

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

Goss-Sampson, M. (2019). Statistical analysis in jasp: A guide for students.

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

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychonomic Bulletin & Review*, *18*, 855–859.

Haberman, J., & Whitney, D. (2012). *Ensemble perception: Summarizing the scene and broadening the limits of visual processing* [Pages: 349]. Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780199734337.003.0030

Haddara, N., & Rahnev, D. (2022). The Impact of Feedback on Perceptual Decision-Making and Metacognition: Reduction in Bias but No Change in Sensitivity [Publisher: SAGE

Publications Inc]. *Psychological Science*, *33*(2), 259–275. https://doi.org/10.1177/09567976211032887

Harp, T. D., Bressler, D. W., & Whitney, D. (2007). Position shifts following crowded second-order motion adaptation reveal processing of local and global motion without awareness. *Journal of Vision*, *7*(2), 15. https://doi.org/10.1167/7.2.15

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values ofd'. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51. https://doi.org/10.3758/BF03203619

Huang, L. (2015). Statistical properties demand as much attention as object features. *PloS one*, *10*(8), e0131191.

Hubert-Wallander, B., & Boynton, G. (2015). Not all summary statistics are made equal: Evidence from extracting summaries across time. *Journal of vision*, *15*, 5. https://doi.org/10.1167/15.4.5

Jachs, B., Blanco, M. J., Grantham-Hill, S., & Soto, D. (2015). On the independence of visual awareness and metacognition: A signal detection theoretic analysis [Place: US Publisher: American Psychological Association]. *Journal of Experimental Psychology: Human Perception and Performance*, *41*, 269–276. https://doi.org/10.1037/xhp0000026

Jackson-Nielsen, M., Cohen, M. A., & Pitts, M. A. (2017). Perception of ensemble statistics requires attention. *Consciousness and cognition*, *48*, 149–160.

Ji, L., & Hayward, W. G. (2021). Metacognition of average face perception. *Attention, Perception, & Psychophysics*, *83*(3), 1036–1048. https://doi.org/10.3758/s13414-020-02189-7

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological review*, *93*(2), 136.

Knotts, J. D., Lau, H., & Peters, M. A. K. (2018). Continuous flash suppression and monocular pattern masking impact subjective awareness similarly. *Attention, Perception, & Psychophysics*, *80*(8), 1974–1987. https://doi.org/10.3758/s13414-018-1578-8

Kolb, F. C., & Braun, J. (1995). Blindsight in normal observers. *Nature*, *377*(6547), 336–338.

Lau, H. (2022). *In consciousness we trust: The cognitive neuroscience of subjective experience*. Oxford University Press.

Lee, S. E., & James Press, S. (1998). Robustness of bayesian factor analysis estimates. *Communications in statistics-theory and methods*, *27*(8), 1871–1893.

Leib, A. Y., Kosovicheva, A., & Whitney, D. (2016). Fast ensemble representations for abstract visual impressions [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *7*(1), 13186. https://doi.org/10.1038/ncomms13186

Lloyd, D. A., Abrahamyan, A., & Harris, J. A. (2013). Brain-stimulation induced blindsight: Unconscious vision or response bias? *PloS one*, *8*(12), e82828.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*(8), 391–400. https://doi.org/10.1016/j.tics.2013.06.006

Luo, A. X., & Zhao, J. (2018). Capacity limit of ensemble perception of multiple spatially intermixed sets. *Attention, Perception, & Psychophysics*, *80*, 2033–2047.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection Theory: A User's Guide* (2nd ed.). Psychology Press. https://doi.org/10.4324/9781410611147

Mamassian, P. (2020). Confidence Forced-Choice and Other Metaperceptual Tasks* [Publisher: SAGE Publications Ltd STM]. *Perception*, *49*(6), 616–635. https://doi.org/10.1177/0301006620928010

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Maule, J., & Franklin, A. (2016). Accurate rapid averaging of multihue ensembles is due to a limited capacity subsampling mechanism [Publisher: Optica Publishing Group]. *JOSA A*, *33*(3), A22–A29. https://doi.org/10.1364/JOSAA.33.000A22

Mei, N., Santana, R., & Soto, D. (2022). Informative neural representations of unseen contents during higher-order processing in human brains and deep artificial networks [Number: 5 Publisher: Nature Publishing Group]. *Nature Human Behaviour*, *6*(5), 720–731. https://doi.org/10.1038/s41562-021-01274-7

Michel, M. (2022). How (not) to underestimate unconscious perception. *Mind & Language*.

Michel, M. (2023). Confidence in consciousness research. *Wiley Interdisciplinary Reviews: Cognitive Science*, *14*(2), e1628.

Morales, J., & Lau, H. (2021). Confidence tracks consciousness. *Qualitative consciousness: themes from the philosophy of David Rosenthal*, 1–21.

Moreno-Martínez, F. J., & Montoro, P. R. (2012). An Ecological Alternative to Snodgrass & Vanderwart: 360 High Quality Colour Images with Norms for Seven Psycholinguistic Variables [Publisher: Public Library of Science]. *PLOS ONE*, *7*(5), e37527. https://doi.org/10.1371/journal.pone.0037527

Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition*, *128*(1), 56–63. https://doi.org/10.1016/j.cognition.2013.03.006

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *The Behavioral and Brain Sciences*, *37*(1), 1–19. https://doi.org/10.1017/S0140525X12003214

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*, 23–36. https://doi.org/10.1016/S0079-6123(06)55002-2

Oriet, C., & Corbett, J. (2008). Evidence for rapid extraction of average size in rsvp displays of circles. *Journal of Vision*, *8*(6), 13–13.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature neuroscience*, *4*(7), 739–744.

Pascucci, D., Ruethemann, N., & Plomp, G. (2021). The anisotropic field of ensemble coding | Scientific Reports. *Scientific Reports*, *11*(1), 8212. https://doi.org/10.1038/s41598-021-87620-1

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., & Louppe, G. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*.

Peirce, C. S., & Jastrow, J. (1884). On small differences in sensation.

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness [Number: 2 Publisher: Nature Publishing Group]. *Nature Neuroscience*, *10*(2), 257–261. https://doi.org/10.1038/nn1840

Peters, M. A. K., & Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli (M. Carandini, Ed.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *4*, e09651. https://doi.org/10.7554/eLife.09651

Peters, M. A., Fesi, J., Amendi, N., Knotts, J. D., Lau, H., & Ro, T. (2017). Transcranial magnetic stimulation to visual cortex induces suboptimal introspection. *Cortex*, *93*, 119–132.

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374. https://doi.org/10.1038/nn.4240

Rahnev, D., Balsdon, T., Charles, L., de Gardelle, V., Denison, R., Desender, K., Faivre, N., Filevich, E., Fleming, S. M., Jehee, J., Lau, H., Lee, A. L. F., Locke, S. M., Mamassian, P., Odegaard, B., Peters, M., Reyes, G., Rouault, M., Sackur, J., … Zylberberg, A. (2022). Consensus goals in the field of visual metacognition. *Perspec-*

tives on psychological science : a journal of the Association for Psychological Science, *17*(6), 1746–1765. https://doi.org/10.1177/17456916221075615

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Reder, L. M., & Schunn, C. D. (2014). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In *Implicit memory and metacognition* (pp. 45–77). Psychology Press.

Rodieck, R. W. (1998). *The first steps in seeing* [Pages: xi, 562]. Sinauer Associates.

Rosenthal, D. (2019). Consciousness and confidence. *Neuropsychologia*, *128*, 255–265.

Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron*, *90*(3), 499–506. https://doi.org/10.1016/j.neuron.2016.03.025

Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, *85*, 1–16.

Sekimoto, T., & Motoyoshi, I. (2022). Ensemble perception without phenomenal awareness of elements | Scientific Reports. *Scientific Reports*, *12*(1), 11922. https://doi.org/10.1038/s41598-022-15850-y

Shekhar, M., & Rahnev, D. (2021). The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review*, *128*(1), 45.

Soto, D., & Silvanto, J. (2016). Is conscious awareness needed for all working memory processes? *Neuroscience of Consciousness*, *2016*(1), niw009.

Soto, D., Mäntylä, T., & Silvanto, J. (2011). Working memory without consciousness. *Current Biology*, *21*(22), R912–R913.

Soto, D., Sheikh, U. A., & Rosenthal, C. R. (2019). A novel framework for unconscious processing. *Trends in Cognitive Sciences*, *23*(5), 372–376.

Stein, T., Kaiser, D., & Hesselmann, G. (2016). Can working memory be non-conscious? *Neuroscience of Consciousness*, *2016*(1).

Stein, T., Utz, V., & Van Opstal, F. (2020). Unconscious semantic priming from pictures under backward masking and continuous flash suppression. *Consciousness and Cognition*, *78*, 102864.

Sweeny, T. D., Wurnitsch, N., Gopnik, A., & Whitney, D. (2015). Ensemble perception of size in 4–5-year-old children. *Developmental science*, *18*(4), 556–568.

Tibshirani, H. R., James, G., & Trevor, D. W. (2017). *An introduction to statistical learning*. springer publication.

Tiurina, N., & Markov, Y. (2022). Ensemble representation of animacy is based on mid-level visual features. *https://psyarxiv.com/h2b4a/*.

Vlassova, A., Donkin, C., & Pearson, J. (2014). Unconscious information changes decision accuracy but not confidence. *Proceedings of the National Academy of Sciences*, *111*(45), 16214–16218.

Ward, E. J., Bear, A., & Scholl, B. J. (2016). Can you perceive ensembles without perceiving individuals?: The role of statistical perception in determining whether awareness overflows access. *Cognition*, *152*, 78–86.

Watamaniuk, S. N. J., & Duchon, A. (1992). The human visual system averages speed information. *Vision Research*, *32*(5), 931–941. https://doi.org/10.1016/0042-6989(92)90036-I

Watamaniuk, S. N. J., Sekuler, R., & Williams, D. W. (1989). Direction perception in complex dynamic displays: The integration of direction information. *Vision Research*, *29*(1), 47–59. https://doi.org/10.1016/0042-6989(89)90173-9

Webster, J., Kay, P., & Webster, M. A. (2014). Perceiving the average hue of color arrays [Publisher: Optica Publishing Group]. *JOSA A*, *31*(4), A283–A292. https://doi.org/10.1364/JOSAA.31.00A283

Weiskrantz, L. (1986). *Blindsight: A Case Study and Implications*. Oxford University Press. https://academic.oup.com/book/2388

Wender, K. F., & Rothkegel, R. (2000). Subitizing and its subprocesses. *Psychological Research*, *64*, 81–92.

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

Wiecki, T., Sofer, I., & Frank, M. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*. Retrieved November 21, 2022, from https://www.frontiersin.org/articles/10.3389/fninf.2013.00014

Williams, J. R. (2008). The declaration of helsinki and public health. *Bulletin of the World Health Organization*, *86*, 650–652.

Yeshurun, Y., Carrasco, M., & Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision research*, *48*(17), 1837–1851.