JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

Perceptual history biases are predicted by early visual-evoked activity

https://doi.org/10.1523/JNEUROSCI.1451-22.2023

Cite as: J. Neurosci 2023; 10.1523/JNEUROSCI.1451-22.2023

Received: 28 July 2022
Revised: 28 December 2022
Accepted: 21 February 2023

1 **Perceptual history biases are predicted by early visual-evoked activity**

2 Abbreviated title: Neural signature of perceptual history

3 Fornaciai, M.[1] *, Togoli, I.[1], Bueti D.[1]

4 [1] Cognitive Neuroscience Department, International School for Advanced Studies (SISSA), Trieste, Italy.

5

6 * Corresponding Author: Michele Fornaciai (mfornaci@sissa.it; michele.fornaciai@gmail.com)

7 Cognitive Neuroscience Department, International School for Advanced Studies (SISSA)

8 Via Bonomea, 265, 34136 Trieste (TS), Italy

9

10 Number of pages: 45

11 Number of figures: 10

12 Number of words: abstract = 246, introduction = 643, discussion = 1500

13 **Acknowledgements.** We thank Matteo Fantuz for his assistance during data collection. This project has
14 received funding from the European Union's Horizon 2020 research and innovation programme under the
15 Marie Sklodowska-Curie grant agreement No. 838823 "NeSt" to MF, from the European Research
16 Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant
17 agreement No. 682117 BIT-ERC-2015-CoG to DB, and from the Italian Ministry of University and
18 Research under the call FARE (project ID: R16X32NALR) and under the call PRIN2017 (project ID:
19 XBJN4F) to DB.

20 **Conflict of interest.** The Authors declare no competing interest.

21

ABSTRACT

What we see in the present is affected by what we saw in the recent past. *Serial dependence* – a bias making a current stimulus to appear more similar to a previous one – has been indeed shown to be ubiquitous in vision. At the neural level, serial dependence is accompanied by a signature of stimulus history (i.e., past stimulus information) emerging from early visual-evoked activity. However, whether this neural signature effectively reflects the behavioural bias is unclear. Here we address this question by assessing the neural (electrophysiological) and behavioural signature of stimulus history in human subjects (both male and female), in the context of numerosity, duration, and size perception. First, our results show that while the behavioural effect is task-dependent, its neural signature also reflects task-irrelevant dimensions of a past stimulus, suggesting a partial dissociation between the mechanisms mediating the encoding of stimulus history and the behavioural bias itself. Second, we show that performing a task is not a necessary condition to observe the neural signature of stimulus history, but that in the presence of an active task such a signature is significantly amplified. Finally, and more importantly, we show that the pattern of brain activity in a relatively early latency window (starting at ~35-65 ms after stimulus onset) significantly predicts the behavioural effect. Overall, our results thus demonstrate that the encoding of past stimulus information in neural signals does indeed reflect serial dependence, and that serial dependence occurs at a relatively early level of visual processing.

SIGNIFICANCE STATEMENT

What we perceive is determined not only by the information reaching our sensory organs, but also by the context in which the information is embedded in. What we saw in the recent past (perceptual history) can indeed modulate the perception of a current stimulus in an attractive way – a bias that is ubiquitous in vision. Here we show that this bias can be predicted by the pattern of brain activity reflecting the encoding of past stimulus information, very early after the onset of a stimulus. This in turn suggests that the integration of past and present sensory information mediating the attractive bias occurs early in the visual processing stream, and likely involves early visual cortices.

INTRODUCTION

Visual perception is not uniquely based on the sensory input received at any given moment, but also reflects the influence of the recent history of stimulation, or *perceptual history*. An increasing amount of evidence indeed shows that the perception and judgement of a current stimulus is modulated by the stimuli seen in the recent past. Namely, a current stimulus is perceived to be more similar to its preceding one than it actually is – an "attractive" bias that has been named *serial dependence* (e.g., Fischer & Whitney, 2014). This bias is thought to originate from the integration of past and present sensory evidence (e.g., Burr & Cicchini, 2014; Fischer & Whitney, 2014), and has been shown to be ubiquitous in vision (Alais, Leung, & Van der Burg, 2017; Cicchini, Anobile, & Burr, 2014; Fritsche, Mostert, & de Lange, 2017; Liberman, Fischer, & Whitney, 2014; Manassi, Liberman, Kosovicheva, Zhang, & Whitney, 2018).

However, the nature of serial dependence and its neural mechanisms are still unclear. This attractive effect has been indeed proposed to engage different putative mechanisms, spanning from sensory or perceptual processing (e.g., Fischer & Whitney, 2014; Fornaciai & Park, 2018a, 2020) to a completely post-perceptual mechanism based on memory or decision (e.g., Fritsche et al., 2017; Wehrman, Wearden, & Sowman, 2020). Overall, although serial dependence seems to operate at the perceptual level (Cicchini, Mikellidou, & Burr, 2017; Fritsche & de Lange, 2019), it also shows the hallmarks of high-level visual processes (e.g., Fornaciai & Park, 2018, 2021, 2022; Pascucci et al., 2019; Samaha et al., 2019), making indeed difficult to understand its nature.

At the neural level, previous studies mostly provide evidence supporting the idea that serial dependence is perceptual in nature. For instance, St John-Saaltink et al., 2016, using functional magnetic resonance imaging (fMRI), showed that the recent history of stimulation biases orientation representations directly in the primary visual cortex. Additionally, electroencephalography (EEG) results further support the idea of an early signature of serial dependence (Fornaciai & Park, 2018a; Fornaciai & Park, 2019b), showing that stimulus history is decodable from visual-evoked potentials early on after the onset of a stimulus,

compatibly with the timing of activity in early visual areas (e.g., Di Russo et al., 2005; Fornaciai, Brannon, Woldorff, & Park, 2017). However, whether such early signature of stimulus history actually represents a genuine correlate of the behavioural serial dependence effect is unclear.

In the present study, we address the link between the behavioural attractive effect entailed by serial dependence and the neural signature of stimulus history, aiming to pinpoint the brain processing stages involved with the behavioural bias. In Exp. 1, we used EEG in conjunction with different magnitude discrimination tasks (i.e., numerosity, duration, and size discrimination task, in three different conditions). Namely, participants had to discriminate either the numerosity, the duration, or the dot size of a constant reference dot-array stimulus in comparison to a variable probe array. To induce serial dependence effects, we presented a task-irrelevant "inducer" stimulus (always modulated in numerosity, duration, and dot size in all task conditions) before the task-relevant ones, and assessed (1) how the inducer affects the perceived magnitude of the reference, and (2) how the inducer magnitude information is carried over to the reference processing at the neural level. Following previous studies, we computed the extent to which the past (inducer) stimulus information is encoded in brain signals using a multivariate "decoding" analysis (e.g., King & Dehaene, 2014). In Exp. 2, we instead used a passive-viewing paradigm, to address the potential role of task-relevance of the stimuli in driving the neural signature of stimulus history. In this experiment, the participants simply watched a sequence of dot-array stimuli modulated in numerosity, duration, and size, while responding to occasional odd-ball stimuli defined by contrast (i.e., to avoid making the magnitude dimensions task relevant as in Exp. 1 while ensuring that the subjects attended the stimuli).

MATERIALS AND METHODS

Subjects

A total of 63 subjects took part in the study (42 females; mean age = 23.9, SD = 3.8; including the author IT), 32 of them were tested in Exp. 1, and 31 were tested in Exp. 2. Two subjects in Exp. 1 were excluded due to noisy EEG recordings (see below *Electrophysiological recording and analysis*). Two subjects were instead excluded from Exp. 2 due to equipment failure (missing EEG data). The final samples included in data analysis were thus 30 and 29 subjects, in Exp. 1 and Exp. 2 respectively. All participants had normal or corrected-to-normal vision, were naïve to the purpose of the study (with the exception of one of the authors who participated in Exp. 1), and signed a written informed consent form before taking part in the study. All experimental procedures were approved by the ethics committee of the International School for Advanced Studies (SISSA), and were in line with the declaration of Helsinki.

The sample size of the experiments was based on a previous behavioural study from our group examining the serial dependence effect in duration and numerosity perception (Togoli et al., 2021). More specifically, we estimated an average effect size (Cohen's d) from the numerosity and duration discrimination tasks tested in Togoli et al., 2021 equal to 0.55. Assuming a one-tailed distribution (based on our hypothesis concerning the direction of the effect), and a desired power of 0.9, a power analysis indicated a minimum sample size of 30 participants. Besides the power analysis based on the behavioural effects, we also considered the effect size obtained in previous EEG studies (Fornaciai & Park, 2019b). The effect size (Cohen's d) in this case was estimated to be 0.6. We thus decided to base the sample sizes on the more conservative estimate based on previous behavioural results.

Apparatus and stimuli

The experiments were performed in a quiet and dimly lit boot, equipped with a Faraday cage. Stimuli were presented on a 1920 x 1080 monitor screen running at 120 Hz, positioned about 80 cm from the

participant. All the stimuli were generated using the Psychophysics Toolbox (Kleiner et al., 2007; Pelli, 1997) on MatLab (version r2019b; The Mathworks, Inc.). In all the experimental conditions, stimuli were arrays of black and white dots (50%/50% proportion) presented on a grey background, randomly positioned within a circular area (i.e., field area). In Exp. 1, a sequence of three dot-array stimuli was presented on the screen in each trial. The first stimulus in the sequence was a task-irrelevant "inducer" stimulus used to induce serial dependence. The inducer stimulus could contain either 12 or 24 dots (numerosity), could be presented for either 140 or 280 ms (duration), and contained dots with radius equal to 4 or 8 pixels (dot size). The levels of the different magnitude dimensions of the inducer were identical in all the task conditions of Exp. 1. The second stimulus was a reference stimulus that was kept constant across all the trials and task conditions. The reference always contained 16 dots, was presented for 200 ms, and each dot had a radius equal to 6 pixels. The last stimulus in the sequence was a probe stimulus that varied according to the specific task condition. Namely, in the numerosity task, the probe contained either 8, 12, 16, 24, or 32 dots, had a duration of 200 ms, and dot size of 6 pixels. In the duration task, the probe had constant numerosity (16 dots) and dot size (6 pixels), and varied in duration (100, 140, 200, 280, or 400 ms). In the size task, the probe had constant numerosity (16 dots) and duration (200 ms), and was varied in dot size (3, 4, 6, 8, or 12 pixels). In Exp. 2, a single stimulus was presented in each trial. All the stimuli were thus modulated in numerosity, duration, and dot size, according to three levels for each dimension. Namely, each dot array could contain either 12, 16, or 24 dots, could be presented for 140, 200, or 280 ms, and had dot size of either 4, 6, or 8 pixels. Occasionally (10 trials in each block of 270 trials; 3.7% of the trials), a "catch" stimulus was presented, which was characterized by a reduced contrast (30% less compared to standard stimuli). In each experiment and condition, the field area of the dot-array stimuli was randomly modulated spanning from 200 to 400 pixel in radius.

--- FIGURE 1 HERE ---

Experimental design

In Exp. 1, participants performed three different discrimination tasks in separate conditions and in a random order (i.e., numerosity, duration, or size discrimination task, involving a 2-alternative forced-choice procedure, 2AFC). The stimulation sequence was largely identical across conditions. Namely, a sequence of three stimuli was presented in each trial: an inducer stimulus, followed by a reference (inducer-reference interstimulus interval, ISI = 650-850 ms), and finally a probe (reference-probe ISI = 600-650 ms). While the inducer and reference stimuli were identical in all the tasks (see *Apparatus and stimuli*), the probe was modulated according to the task – i.e., it varied in numerosity in the numerosity task, in duration in the duration task, and in dot size in the size task. At the end of the stimulus sequence, participants had to report which stimulus between the reference and the probe contained more dots, lasted longer in time, or contained larger dots. Participants were instructed to respond as fast and accurately as possible, and although the inducer was irrelevant for the task, to pay anyway attention to the entire sequence of the stimuli, in line with previous studies (e.g., Fornaciai & Park, 2018b). The available time to provide a response was limited to 1,250 ms. Once a participant provided a response, the next trial started automatically after 800-1,200 ms. If no response was provided within 1,250 ms from the offset of the last stimulus in the sequence, the next trial started automatically. Responses were collected by means of a standard keyboard. On average, participants missed the response in 11 (SD = 15.7) trials, out of a total of 400 trials performed in each condition. Trials in which no response was provided were excluded from behavioural data analysis, but included in the EEG analysis.

In Exp. 2, instead, participants were asked to watch a continuous stream of dot-array stimuli modulated in numerosity, duration, and dot size. The ISI between consecutive stimuli was 800 ms. To encourage participants to pay attention to the stimulus sequence, we asked them to perform a simple oddball detection task. More specifically, an occasional oddball ("catch") stimulus was presented (3.7% of the trials), with a reduced contrast compared to the majority of other ("standard") stimuli. Participants were instructed to press a button on a keyboard as fast as they could once they detected the lower-contrast

171 oddball stimulus. The detection rates in this task were on average (± SD) 93% ± 1.3%. The average
172 reaction time in correctly detected catch trials was 313 ± 11 ms.

173 In Exp. 1, participants completed 10 blocks of 40 trials in each task condition, including 20 repetitions of
174 each combination of inducer and probe magnitude. In Exp. 2, participants completed 8 blocks of 270
175 trials, for a total of 80 repetitions of each combination of stimulus magnitudes. Participants were free to
176 take breaks between different blocks. The Experiment took about 1 hour in the case of Exp. 1, and about
177 50 minutes in the case of Exp. 2.

178

179 ***Behavioural data analysis***

180 In Exp. 1, participants' performance in the three different tasks was analysed to assess to what extent the
181 perceived magnitude of the reference stimulus was affected by the inducer. First, the proportion of
182 responses in the task obtained from each participant and condition was fitted with a cumulative Gaussian
183 function according to the maximum likelihood method (Watson, 1979). The point of subjective equality
184 (PSE), representing the perceived magnitude of the reference stimulus (i.e., accuracy), was defined as the
185 median of the cumulative Gaussian function. As a measure of precision in the task, we computed the just
186 noticeable difference (JND) as the difference in probe magnitude between chance level response (50%)
187 and 75% "probe more numerous/longer/bigger" responses. As an additional measure of precision, we
188 computed the Weber fraction as JND/PSE. Finally, during the fitting procedure, we applied a finger error
189 rate correction (5%) to account for random errors and lapses of attention (Wichmann & Hill, 2001). Note
190 that the PSE and the JND were computed as a function of the different levels of the inducer magnitudes,
191 separately for each dimension (i.e., the numerosity, duration, and size of the inducer). As a final estimate
192 of the precision, we considered the average WF across the different levels of the inducer magnitudes in
193 each task. The difference in PSE obtained with different inducer magnitudes within each task (shown in
194 Fig. 2A) was assessed with a series of paired t-tests. To control for multiple comparisons, we employed a

false discovery rate (FDR) procedure, with q = 0.05. When reporting series of t-tests, we thus report the p-value adjusted by the FDR procedure ("adj-p"). The difference in WF across the tasks was assessed with a one-way repeated measures ANOVA, with factor "task."

Moreover, to better compare the serial dependence effects obtained in different tasks, we computed a serial dependence effect index according to the following formula:

$$\text{Serial dependence effect} = ((PSE_{high} - PSE_{low})/PSE_{low}) \times 100; \quad (1)$$

Where PSE_{low} refers to the PSE obtained when the inducer magnitude was low (i.e., 12 dots in the numerosity task, 140 ms in the duration task, 4 pixels in the size task), while PSE_{high} refers to PSEs obtained with a high inducer magnitude (24 dots, 280 ms, or 8 pixels, according to the task). This index was calculated separately for each participant and condition, and the average is shown in the Fig. 2B. Additional analyses (data not shown) were performed to assess whether the influence of the inducer is limited to the immediately following reference stimulus, or whether it extends across trials. To do so, we computed the PSE and the serial dependence effect as a function of the magnitudes of the inducer in the preceding trial or two trials back in the past. These analyses did not show any influence of the inducer across trials, most likely due to the presence of several intervening stimuli. Previous results concerning serial dependence in magnitude perception (although limited to numerosity perception) indeed showed that the effect is mostly limited to the immediately preceding stimulus (Cicchini et al., 2014; Fornaciai & Park, 2020, 2022). As a sanity check, we also assessed whether the reference stimulus could be affected by the inducer in the successive (future) trial. No effect was observed also in this case.

To obtain a better measure of the influence of each inducer magnitude on behavioural performance, we further performed a non-linear regression analysis. In this analysis, performed separately for each participant and condition, we used the individual responses across trials (i.e., coded as 0 or 1, as obtained in the 2AFC task) as dependent variable, and included the probe magnitude (defined according to the task), inducer numerosity, inducer duration, and inducer dot size as predictors. This allowed us to directly

assess the influence of each inducer magnitude level on the response in each trial. To perform this regression analysis, the levels of inducer magnitudes were coded as the ratio with the corresponding reference magnitude. Regarding the effects computed with this analysis, positive beta values indicate an attractive effect (i.e., increased probability of judging the reference as "bigger" than the probe with higher inducer magnitude), while negative beta values indicate a repulsive effect. The resulting beta values obtained for each participant and magnitude (see Fig. 2C; note that beta values corresponding to the effect of the probe are not shown in the figure) were then tested with a one-sample t-test to assess whether the corresponding predictor provided an effect significantly higher than zero. All statistical tests were performed in Matlab (version R2018b).

Electrophysiological recoding and analysis

In both Exp. 1 and Exp. 2, the electroencephalogram (EEG) was recorded throughout the experimental procedure, using the Biosemi ActiveTwo system (at a sampling rate of 2048 Hz), and a 32-channel cap based on the 10-20 system layout. In Exp. 2, an electro-oculogram (EOG) channel was also added below the left eye. Electrode offsets across channels were usually kept below 15 μV, but occasionally offsets up to 30 μV were tolerated.

EEG data analysis was performed offline in Matlab (version R2018b), using the EEGLAB software package (Delorme & Makeig, 2004). EEG signals were first high-pass filtered (0.1 Hz) and re-referenced to the average of all the channels used. Continuous EEG data were then segmented into epochs spanning from -200 ms to 700 ms after stimulus onset, with epochs time locked to the reference stimulus in Exp. 1, or time-locked to each stimulus in Exp. 2. In order to test the influence of the inducer stimulus (Exp. 1) or the stimulus in the preceding trial (Exp. 2), epochs were also sorted as a function of the magnitudes of the preceding stimulus. For instance, in Exp. 1, the reference epochs were sorted according to the different levels of the inducer magnitudes – i.e., separately for cases were the inducer had 12 or 24 dots, was

243 presented for 140 or 280 ms, or had dot-size of either 4 or 8 pixels. In Exp. 2, we considered as
244 "reference" of one magnitude dimension all the stimuli presenting the intermediate magnitude level (i.e.,
245 16 dots, 200 ms, 6 pixels, which corresponded to the reference stimulus in Exp. 1). We thus sorted epochs
246 according to whether such intermediate stimuli (i.e., the "current" magnitude) were preceded by a
247 stimulus having either a lower or higher magnitude, separately for each dimension (i.e., the "past"
248 magnitude). Data from both experiments were cleaned by means of an independent component analysis
249 (ICA), aimed to remove artefacts related to eye movements, blinks, or other sources of noise. After ICA,
250 we used a step-like artefact rejection procedure to remove any remaining large artefact, leading to an
251 average rejection rate of 9.8% (SD = 13.5%) in Exp. 1, and 0.6% (SD = 1.2%) in Exp. 2. Rejection rate
252 was used as a criterion for inclusion in data analysis, with a cut-off rejection rate of 35%. Two
253 participants in Exp. 1 were excluded from data analysis based on this criterion. Due to equipment failure,
254 one participant in Exp. 1 had one missing block of trials but was nevertheless included in data analysis as
255 the number of available trials was sufficient to perform the multivariate analysis. Finally, we applied a
256 low-pass filter with a cut-off of 30 Hz.

257

258 ***Multivariate pattern analysis in the time domain***

259 In order to characterize the neural signature of stimulus history during the processing of the current
260 stimulus, we employed a multivariate pattern analysis in the time domain (or "decoding" analysis), using
261 the Neural Decoding Toolbox (Meyers, 2013). This analysis has indeed proven to be very sensitive in
262 decoding stimulus history in previous studies (Fornaciai & Park, 2018; Fornaciai & Park, 2019b). In
263 general, the multivariate analysis involves the training of a pattern classifier (support vector machine,
264 SVM) on EEG data coming from multiple channels, corresponding to two specific classes of stimuli.
265 Then, the classifier is tested on an independent subset of un-labelled data to assess whether it could
266 discriminate the class of stimuli the test data belong to. The classification accuracy yielded by the
267 classifier in correctly discriminating the two classes of stimuli indicates the extent to which they generate

unique patterns of activity, and, by repeating this procedure across several time windows it hence provides an index of when the stimulus information is encoded in brain activity.

To assess the neural signature of stimulus history, in Exp. 1 we used epoched EEG data time-locked to reference stimulus onset, with epochs sorted as a function of different inducer magnitudes. To assess the effect of inducer numerosity on the reference, we entered in the analysis epochs corresponding to the reference stimulus preceded by either a 12 or 24-dot inducer (irrespective of the other magnitudes). Similarly, in the case of duration and size, we used epochs corresponding to the reference preceded by a 140 or 280-ms inducer, or 4 or 8-pixels inducer. The classifier was thus tested in discriminating activity evoked by an identical reference stimulus as a function of the magnitude of the preceding inducer. This analysis was performed separately for each participant, task, and inducer magnitude dimension. In Exp. 2, we used epochs corresponding to the stimuli with intermediate magnitude levels (i.e., 16 dots, 200 ms, 6 pixel) preceded by either a lower or higher magnitude across the three dimensions. For instance, we used epochs corresponding to a 16-dot stimulus preceded by either 12 or 24 dots, 140 or 280 ms, or 4 or 8 pixels, to assess the effect of different magnitudes on numerosity. The same was done for the duration and size of the stimuli.

In both experiments, we implemented a series of practices to optimize the analysis and reduce noise. First, instead of using subsets of single trials to train and test the classifier, we averaged together random sets of trials (Grootswagers et al., 2016) to create averaged "pseudo-trials." The trials used to generate pseudo-trials were randomly drawn from the data set, and hence were not groups of subsequent trials. The number of individual trials averaged into pseudo-trials varied between 10 and 20 based on the number of available trials after artifact rejection (on a subject-by-subject basis), with an average (± SD) of 18.1 ± 1.8 trials in Exp. 1, and 19.1 ± 2.9 in Exp. 2. The training of the classifier was thus performed on a set of such pseudo-trials, and testing was performed on a remaining pseudo-trial according to a leave-one-out procedure. The specific number of pseudo-trials used in the cross-validation procedure was fixed to 10 in Exp. 1 (i.e., nine trials for training and the remaining one for testing) and varied in Exp. 2 according to the

293 available data (mean ± SD = 11.5 ± 2.1 pseudo-trials). To avoid overfitting (i.e., the classifier learning an
294 overly specific pattern that fails to generalise to the test set), we also performed a feature selection
295 procedure prior to the decoding analysis, restricting the decoding procedure to the five most informative
296 EEG channel as determined with a univariate ANOVA performed on the training set (Grootswagers et al.,
297 2016). Note that such feature selection procedure does not make the analysis circular. Indeed, the feature
298 selection was performed only on the training data set, leaving the test set independent. Since the selection
299 of specific channels during the analysis does not provide information about how they contribute to the
300 decoding performance, we chose not to explicitly assess the frequency of channel selection and the
301 topography of the most frequently selected channels. Moreover, instead of performing the analysis at
302 individual time points, we averaged activity across a series of 100-ms time windows, with a step size of
303 20 ms. Finally, the decoding procedure was repeated 30 times using different subsets of trials for training
304 and testing and for creating pseudo-trials, and the average of all the iterations of the analysis was taken as
305 the final estimate of classification performance. Classification accuracy measures obtained throughout the
306 epoch reflect to what extent the pattern classifier was able to classify stimuli according to the pattern of
307 activity, and hence, considering our comparisons, to what extent information from the previous stimulus
308 is encoded in the brain responses to a current one.

309 Decoding results were tested for significance by taking the average across two latency windows: an early
310 window spanning from 50 to 200 ms post-stimulus (based on previous studies; Fornaciai & Park, 2018a;
311 Fornaciai & Park, 2019b), and a late latency window spanning from 500 to 650 ms post-stimulus. This
312 second window was chosen to span late latencies capturing post-perceptual processes like working
313 memory encoding (e.g., see for instance Oh et al., 2020). Decoding results averaged in these two latency
314 windows were tested using a series of one-sample t-tests (against an empirical measure of chance level,
315 see below), followed by a three-way repeated measures ANOVA in both Exp. 1 and Exp. 2. In Exp. 1, we
316 entered as factors “task” (numerosity, duration, and size task), “inducer magnitude” (numerosity,
317 duration, and size), and “latency window” (early vs. late). In Exp. 2, we entered as factors “current

318 magnitude" (i.e., representing stimuli with the intermediate level of either numerosity, duration, or size),
319 "past magnitude" (i.e., representing the stimuli in the immediately preceding trial with the extreme levels
320 of either numerosity, duration, or size), and "latency window." To assess interaction effects in these two
321 ANOVAs, we further used simpler ANOVA models and paired t-tests. Finally, to directly compare the
322 decoding results obtained in the two experiments, we used a mixed model ANOVA with "current
323 magnitude" (i.e., either the task-relevant magnitude in Exp. 1, or the magnitude considered in the current
324 trial in Exp. 2), "past magnitude" (i.e., the different inducer magnitudes in Exp. 1, or the corresponding
325 different magnitudes of the preceding stimulus in Exp. 2), and "latency window" as within-subject
326 factors, and "experiment" (Exp. 1 vs. Exp. 2) as between-subject factor.

327 Besides our main decoding analysis, we also performed a control ("null") decoding analysis. In this
328 analysis, we replicated our main procedure with the exception that we shuffled the labels of the trials in
329 the training set prior to the classification procedure. The results of this null decoding analysis were then
330 used to set the chance level empirically. One-sample t-tests against chance level were thus performed
331 against the corresponding average classification accuracy of the null analysis, rather than against the 50%
332 probability level.

333 Finally, to address the link between the behavioural serial dependence effect and the classification
334 accuracy obtained in the decoding procedure, we performed a series of tests based on a linear mixed-
335 effect regression model, defined as follows:

336 $$Eff \sim CA + (1 \mid Subj); \qquad (2)$$

337 Where "Eff" represents the index of the behavioural serial dependence effect computed according to
338 Equation 1, "CA" represents the classification accuracy of the decoding procedure, and (1 | Subj)
339 represents the random effect (subjects). First, we performed regression tests considering the average
340 classification accuracy in each of the two latency windows used in the analysis (50-200 ms and 500-650
341 ms after stimulus onset). For instance, we assessed the relation between the classification accuracy of the

342 inducer numerosity and the behavioural effect caused by the inducer numerosity in each task, and so on
343 for all the other inducer dimensions. Additionally, we also applied this regression model in a more
344 comprehensive fashion, to more precisely assess the latencies at which a relationship between neural and
345 behavioural effect might emerge. Namely, the same regression model was applied at each time window
346 throughout the reference epoch (i.e., corresponding to the 100-ms time windows used in the decoding
347 analysis). In this procedure, we only considered significant clusters of at least two consecutive time
348 windows showing a significant relation between the behavioural effect and classification accuracy. To
349 ensure the robustness of these clusters of significant time windows, we used a cluster-based non-
350 parametric test. In this test, we examined each of the clusters identified in the main procedure separately.
351 Namely, we shuffled the distribution of classification accuracy at each time window within the cluster and
352 the distribution of behavioural effect across the group, and applied the same LME model on these
353 shuffled data, again separately for each time window within each cluster. The procedure was repeated
354 10,000 times, shuffling the datasets each time and collecting the results. To assess the robustness of the
355 actual clusters, we measured how many times the simulated cluster showed the same number of
356 contiguous significant time windows as the actual cluster, which represents the p-value of the test. As a
357 threshold for determining the significance of each test performed within a simulated cluster, we used the
358 minimum t-value obtained in each actual cluster. All the analyses were performed in Matlab (version
359 r2018b).

360

361 ***Data availability***

362 The data generated during the experiments described in this manuscript is available on Open Science
363 Framework following this link: https://osf.io/ju78r/.

364

365

366 **RESULTS**

367 ***Experiment 1***

368 In Exp. 1, we addressed the link between serial dependence and the neural signature of stimulus history
369 across different magnitude dimensions and task conditions. The experiment (depicted in Fig. 1A) was
370 divided into three different conditions (numerosity, duration, size), performed by participants (N = 30) in
371 random order. In all the task conditions, participants discriminated the relevant magnitude (either
372 numerosity, stimulus duration, or dot size) of a reference dot-array (always 16 dots, lasting 200 ms, and
373 with dot size equal to 6 pixels) compared to a variable probe (ranging from 8 to 32 dots, from 100 to 400
374 ms, and from 3 to 12 pixel in the numerosity, duration, and size task, respectively). To induce serial
375 dependence, a task-irrelevant inducer stimulus was presented before the reference, and was modulated
376 across the three dimensions (with a numerosity of either 12 or 24 dots, a duration of 140 or 280 ms, and
377 dot size of 4 or 8 pixel) in all the conditions.

378

379 --- FIGURE 2 HERE ---

380

381

382 To assess the effect at the behavioural level, we first computed the point of subjective equality (PSE, a
383 measure of the perceived magnitude of the reference stimulus; Fig. 2A) within each task condition and
384 assessed how the perception of the reference stimulus is modulated by the corresponding inducer
385 magnitude (i.e., numerosity in the numerosity task, and so on). To do so, we performed a series of paired
386 t-tests within each condition, controlling for multiple comparisons using a false discovery rate (FDR)
387 procedure with q = 0.05. The p-values reported thus reflect the p-values adjusted by the FDR (adj-p). The
388 results show a significant difference in PSE as a function of inducer magnitude in the numerosity ($t(29) =$

2.60, adj-p = 0.022, Cohen's d = 0.48) and size (t(29) = 7.79, adj-p = 0.003, Cohen's d = 1.45) task, suggesting that the higher the inducer magnitude (more numerous, bigger dot size), the higher the perceived magnitude of the reference. No effect was instead observed in the duration task (t(29) = 0.47, adj-p = 0.64). To compare the effect across different conditions, we also computed a serial dependence effect index based on the normalised difference in PSE (Fig. 2B). On average, the serial dependence effect resulted to be of 3.42% ± 1.26%, 0.92% ± 1.53%, and 7.58% ± 1.18%, for the numerosity, duration, and size task respectively. The effect in the size task was significantly higher compared to the other two conditions (t(29) = 2.56, adj-p = 0.016, d = 0.47, and t(29) = 3.37, adj-p = 0.004, d = 0.61, respectively for size vs. numerosity and vs. duration).

To further assess the serial dependence effect both within and across different dimensions, we performed a non-linear regression analysis aimed to quantify the contribution of different inducer magnitudes to the discrimination judgment in each trial (Fig. 2C). The results showed a systematic influence of the inducer on perceptual judgements, in a task-specific fashion. Namely, we observed a significant effect of inducer numerosity in the numerosity task (beta = 0.030 ± 0.060; one-sample t-test against zero, t(29) = 2.67, adj-p = 0.037, d = 0.5), and inducer size in the size task (beta = 0.075 ± 0.033; t(29) = 12.31, adj-p < 0.001, d = 2.3). Interestingly, besides the attractive (positive) effects, we also observed a repulsive (negative) effect provided by the inducer duration in the numerosity task (beta = -0.017 ± 0.042), which however did not reach significance after controlling for multiple comparisons (t(29) = -2.22, adj-p = 0.051, d = 0.4). This effect however showed a medium effect size (d = 0.4) similar to the effect of numerosity (d = 0.5), suggesting that it might reflect a genuine repulsive effect. No other significant contribution to behavioural responses was observed (all t(29) ≤ 1.87, all adj-p ≥ 0.107).

Does the pattern of behavioural effects reflect differences in task difficulty? In terms of participants' precision in the task (Weber's fraction, WF), we actually observed an opposite pattern of results compared to the effect. Indeed, the most difficult task resulted to be the duration task (i.e., highest Weber's fraction, WF [mean ± SEM] = 0.17 ± 0.017), followed by the numerosity task (WF = 0.12 ±

0.009), and finally the size task where precision was the highest (WF = 0.07 ± 0.003). A one-way repeated measure ANOVA showed that there is indeed a significant difference between precision in the different tasks ($F(2,58) = 34.03$, $p < 0.001$, $\eta_p^2 = 0.281$). The different difficulty of the three tasks is also reflected by different average response times (RTs), which were longest in the duration task (mean ± SEM = 498 ± 21 ms), again followed by the numerosity (378 ± 21 ms) and size task (358 ± 19 ms). There was a significant difference across RTs in different tasks (one-way repeated measures ANOVA, main effect of condition; $F(2,29) = 116.75$, $p < 0.001$, $\eta_p^2 = 0.80$), with the duration task showing significantly slower responses compared to the other two conditions (paired t-test; duration vs. numerosity: $t(29) = 12.15$, adj-p = 0.001; duration vs. size, $t(29) = 14.10$, adj-p = 0.001). This suggests that serial dependence in this context is unlikely to be associated with poorer performance, but it is instead strongest in the easiest task. Nevertheless, serial dependence might still be associated with poorer perceptual precision (Cicchini et al., 2018) within each task condition. A series of tests performed within each condition however did not show a consistent association between the serial dependence effect and the WF, neither in the numerosity ($r = 0.09$, $p = 0.63$) nor in the duration task ($r = 0.31$, $p = 0.09$). In the size task we did observe a significant (negative) correlation, which was however driven by a single data point with particularly low WF (difference > 2 SD from the average). When excluding such data point, no significant correlation was observed ($r = -0.04$, $p = 0.83$).

--- FIGURE 3 HERE ---

To characterize the encoding of stimulus history (i.e., the magnitude information conveyed by the inducer) during the reference processing, we performed a multivariate pattern analysis in the time domain (e.g., King & Dehaene, 2014). The analysis involved the training and testing of a classifier (support

vector machine) on EEG epochs time-locked to the reference stimulus, sorted according to the magnitude of the preceding inducer (i.e., low vs. high numerosity, duration, or size). The resulting classification accuracy provides a measure of whether and to what extent the inducer magnitude is decodable from the brain responses to the reference stimulus. Fig. 3A shows the average classification accuracy of the different task conditions, relative to the numerosity, duration, and size of the inducer. As shown in the figure, the decoding procedure yielded on average a good level of classification accuracy, showing both similarities (especially at late latencies, around 600 ms post-stimulus) and difference (at earlier latencies) across the different conditions. In addition to our main decoding analysis, we also performed a control, "null" decoding analysis, in which the labels of the conditions being compared were shuffled before training the classifier. The classification accuracies obtained with this analysis (see for instance Fig. 3B) were used to set the chance level empirically, and to control for spurious results.

--- FIGURE 4 HERE ---

To assess the neural signature of stimulus history and its pattern across magnitude dimensions and tasks, classification accuracies obtained throughout the reference epoch were averaged across two different time windows, spanning 50-200 ms and 500-650 ms (marked with shaded grey areas in Fig. 3A), separately for each inducer magnitude in each of the condition. Then, we checked for potential differences in the signature of the different inducer magnitudes across the three tasks. Fig. 4A-B shows the average classification accuracies of the different inducer magnitudes for the three tasks and the two latency windows. First, to assess the pattern of results across the different conditions and dimensions, we performed a series of one-sample t-tests against the corresponding average classification accuracy obtained in the null decoding analysis (marked in Fig. 4A-B with dotted lines). To control for multiple

462 comparisons, we applied a false discovery rate (FDR) procedure within each time window ($q = 0.05$). In
463 the early latency window, the effect of inducer numerosity was significantly higher than the (empirical)
464 chance level in the numerosity (one-sample t-test, $t(29) = 2.82$, adj-$p = 0.025$) and in the duration ($t(29) =$
465 4.83, adj-$p < 0.001$) task, but not in the size task ($t(29) = 1.54$, adj-$p = 0.201$). The effect of inducer
466 duration was significant only in the size task ($t(29) = 4.29$, adj-$p < 0.001$), and not in the numerosity and
467 duration task ($t(29) = 1.38$-1.98, adj-$p = 0.129$-0.228). Finally, the effect of stimulus size was not
468 significantly higher than chance level in any of the tasks (max $t(29) = -0.138$-1.686, min adjusted $p =$
469 0.184). In the late latency window, with the exception of duration in the numerosity task ($t(29) = 1.38$,
470 adj-$p = 0.064$) and size in the duration task ($t(29) = 1.96$, adj-$p = 0.077$), all the classification accuracies
471 were significantly above the respective chance level measured empirically ($t(29) = 3.10$-5.01, adj-$p \leq$
472 0.025). These results suggest that the stimulus history information encoded in brain responses does not
473 necessarily match the pattern of behavioural serial dependence effects, as for instance we observed a
474 significant decoding even for dimensions that did not significantly affect behaviour.

475 We then performed a three-way repeated measures ANOVA on the average classification accuracy
476 computed in the two latency windows and in the different tasks, with factors "task" (i.e., numerosity,
477 duration, and size task), "inducer magnitude" (i.e., inducer numerosity, duration, and size), and "latency
478 window" (i.e., early vs. late). This analysis showed no main effect of task ($F(2,58) = 1.093$, $p = 0.342$), no
479 main effect of inducer magnitude ($F(2,58) = 1.739$, $p = 0.185$), but a significant main effect of latency
480 window ($F(1,29) = 6.927$, $p = 0.013$, $\eta_p^2 = 0.193$) and a significant interaction between inducer magnitude
481 and latency window ($F(2,58) = 3.894$, $p = 0.026$, $\eta_p^2 = 0.119$). No other interaction effect was observed
482 (max $F = 1.577$, min $p = 0.185$).

483 To better understand the nature of this interaction, we followed it up by focusing on the average effect of
484 different magnitudes at different time windows. Namely, as the task did not seem to significantly
485 modulate the pattern of classification accuracy or interact with the other factors, we collapsed (averaged)
486 the classification accuracies across the different tasks (Fig. 4C and 4D). Two separate one-way repeated

measures ANOVAs (with factor "inducer magnitude"), showed a significant main effect of inducer magnitude ($F(2,58) = 5.654$, $p = 0.006$, $\eta_p^2 = 0.163$) in the early latency window (50-200 ms; Fig. 4C), and no significant difference across the inducer magnitudes ($F(2,58) = 0.168$, $p = 0.846$) in the late latency window (500-650 ms; Fig. 4D). This difference in the pattern of decoding across the two latency windows thus explain the interaction observed in the previous test, and show that while the level of classification accuracy is more variable across magnitudes at early latencies, it becomes very similar at later latencies.

Finally, a paired t-test comparing the average decoding performance in the early vs. late latency window (i.e., average of the three bars in Fig. 4C vs. the average of Fig. 4D) showed that classification accuracies were overall significantly higher at the late window (0.542 ± 0.005 vs. 0.585 ± 0.016; $t(29) = 2.630$, $p = 0.0135$, Cohen's $d = 0.56$). This suggests that the initial pattern of activity encoding the inducer magnitude information is amplified (i.e., stronger activity, or sharper representation) at later latencies.

--- FIGURE 5 HERE ---

In our interpretation, the results of the decoding analysis show a signature of stimulus history affecting the processing of the reference. However, in the analysis we also observed relatively high classification accuracies even before the onset of the reference stimulus, especially considering the duration condition (Fig. 3A, Fig. 5B). This raises the possibility that what we are measuring may not be a signature of the effect of stimulus history on the reference representation, but a lingering trace of the inducer stimulus itself. To address this possibility, we checked the temporal generalization matrices of each inducer

magnitude effect, averaging the different tasks together (Fig. 5). The temporal generalization matrices are obtained by training and testing the classifier with brain activity at different latencies, to assess whether specific patterns of activity generalise to different latencies. Borrowing from the interpretations provided by King & Dehaene (2014), a lingering trace of the inducer stimulus is expected to emerge in the temporal generalization matrices as a relatively stable signature encompassing the pre-stimulus interval and extending to post-stimulus latencies. In the case of numerosity (Fig. 5A) and size (Fig. 5C), we observed relatively weak activity in the pre-stimulus interval, which instead starts to rise only after the onset of the reference stimulus. In line with the plots shown in Fig. 3A, which represent the diagonals of the temporal generalisation matrices, numerosity shows an early peak at around 150 ms post-stimulus followed by a later peak at around 600 ms, while size only shows a main peak at late latencies. In these two cases, there is no evidence of early decoding capturing a lingering trace of the inducer stimulus. In the case of duration (Fig. 5B), pre-stimulus classification accuracies appeared to be stronger. However, the decoding showed a pattern mostly unfolding along the diagonal, with little generalisation. This suggests the presence of distinct patterns of brain activity evolving over time (see King & Dehaene, 2014), which are more in line with an active processing of stimulus history rather than with a passive trace of the inducer stimulus. Additionally, in all three cases, the large peak observed towards the end of the stimulus epoch is not consistent with a trace of the past stimulus, which would instead be expected to decay over time. Overall, the temporal generalization plots provided little evidence for the presence of a lingering trace of the inducer stimulus. More likely, the pre-stimulus decoding could represent an anticipatory response to the reference due to the relatively narrow jittering of its onset, or a by-product of the sliding window average used in our decoding analysis. Although we cannot conclusively exclude the presence of a lingering trace of the inducer stimulus, the specific patterns of decoding observed in the analysis suggest that, if present, such a trace would likely interact with the processing of the reference stimulus.

Besides the neural signature of stimulus history per se, an important question is whether such a signature relates to the attractive effect observed behaviourally. Therefore, to establish a link between the behavioural and neural effect, we addressed whether the strength of the serial dependence effect reflects the extent to which past information is encoded in brain signals, as indicated by the classification accuracy. To do so, we performed a series of linear mixed-effect (LME) model tests aimed at assessing whether the behavioural effect (i.e., normalized difference between PSEs obtained as a function of different inducer magnitudes; see Fig. 2B) could be predicted by the classification accuracy (CA) obtained in the decoding analysis. In these tests, the behavioural effect thus represented the dependent variable, the classification accuracy the fixed effect, and the subjects were added as the random effect (Eff ~ CA + (1|subj)). This analysis was performed separately in the different task conditions and the two latency windows used in the previous analysis (50-200 ms, 500-650 ms). The results showed that the strength of the behavioural effect can be successfully predicted by the classification accuracy in the early latency window (50-200 ms) when considering the effect of numerosity in the numerosity task and the decoding of inducer numerosity ($R^2 = 0.68$, $t = 2.69$, $p = 0.006$), and the effect of size in the size task and the decoding of inducer size ($R^2 = 0.66$, $t = 2.64$, $p = 0.013$). No other test reached significance in the early latency window (t spanning from -1.08 to 0.60, min $p = 0.29$). In the late latency window, no significant relationship between the behavioural effect and the classification accuracy was observed (t spanning from -1.56 to 1.32, min $p = 0.13$).

--- FIGURE 6 HERE ---

In addition to these tests focused on the two large latency windows used in the previous analyses, we also took a more comprehensive approach and performed a series of LME tests at different latencies

559 throughout the reference epoch (-200:700 ms), to assess whether the strength of the behavioural effect
560 could be predicted by the classification accuracy. To limit the number of tests, we only considered the
561 classification accuracies and behavioural effects of the task-relevant magnitude in each task condition
562 (i.e., effect of numerosity in the numerosity task, duration in the duration task, size in the size task). The
563 pattern of classification accuracies corresponding to this subset of conditions is shown in Fig. 6. In this
564 analysis, we considered an effect significant only when showing at least two consecutive significant time
565 windows, and applied a cluster-based non-parametric test to control for multiple comparisons (see below).
566 In the numerosity task condition, the classification accuracy relative to the inducer numerosity
567 significantly predicted the behavioural effect provided by the inducer numerosity across two clusters: a
568 smaller one spanning from 65 to 85 ms after stimulus onset (2 consecutive time windows; average R^2 =
569 0.62, t-values = 2.15-2.17, p = 0.038-0.041), and a larger one from 145 to 245 ms (7 time windows;
570 average R^2 = 0.65, t-values = 2.10-2.86, p = 0.008-0.045). In the duration task, no significant effect was
571 observed (max t = 1.82, min p = 0.08). Finally, in the size task condition, we observed again two clusters
572 of significant effects showing a relationship between the behavioural effect and the classification
573 accuracy, one spanning from 35 to 85 ms (4 time windows; average R^2 = 0.63, t-values = 2.06-2.57, p =
574 0.015-0.048), and another one going from 115 to 180 ms (5 time windows; average R^2 = 0.65, t-values =
575 2.13-2.96, p = 0.006-0.042). The significant clusters are marked in Fig. 6A at the bottom of the plot, with
576 the same colour code as the main plots. To ensure the robustness of these clusters of significant effects,
577 we also performed a cluster-based non-parametric test, shuffling the data entered into the LME test and
578 assessing the proportion of times that a similar cluster could be observed by chance (10,000 repetitions
579 for each cluster). The cluster-level p-value obtained in this way was in all the cases < 0.001. Compared to
580 the broad latency windows used in the previous analysis, these tests allowed to identify with greater
581 temporal precision the early latencies at which a relationship between behavioural effect and neural
582 signature of stimulus history emerges.

583 To show the direction of the relationship captured with the regression analysis, we also plotted the (log-
584 scaled) average classification accuracy observed across the significant time windows against the
585 behavioural effect (limited to numerosity and size; Fig. 6B and 6C). Both plots show that there is a
586 positive relationship between these two measures, with stronger behavioural effects associated with
587 higher classification accuracies. In both cases, we observed a significant correlation between the two
588 measures ($r = 0.37$, $p = 0.043$ and $r = 0.39$, $p = 0.035$, respectively for numerosity and size).

589 Note that the significance of the regression analysis does not depend on the absolute value of the
590 classification accuracy, but on the pattern across the group as a function of the behavioural effect. In the
591 size condition, indeed, a significant relationship between CA and behavioural effect was observed at
592 latencies that show weak decoding, close to the 50% probability level. Despite the weak decoding,
593 differences across the group can still show a relationship with the strength of the behavioural effect.

594

595 ***Experiment 2***

596 The results of Exp. 1 showed that (1) while the behavioural serial dependence effect shows different
597 patterns according to the task performed by participants, with task-specific effects, the decoding of
598 stimulus history shows a more generalised signature also reflecting inducer's magnitudes that did not
599 yield a behavioural effect. (2) We nevertheless observed a link between the behavioural effect and its
600 neural signature, emerging at early latencies after stimulus onset.

601 In Exp. 2, we further asked whether the neural signature of stimulus history might be modulated by task-
602 related factors. Previous results show that stimulus history can be decoded from neural signals in a
603 passive-viewing paradigm (Fornaciai & Park, 2018), suggesting that the encoding of stimulus history
604 does not hinge upon the presence of an active task. However, it is unclear whether the signature of
605 stimulus history would be completely independent from what the participants are doing, or whether its
606 strength could be modulated by being engaged in a task. Our aim was thus to compare the strength and

607 pattern of stimulus history decoding with the results of Exp. 1, to address the potential modulatory role of
608 the task as opposed to a passive viewing of a series of stimuli.

609 To address this question, we employed a passive-viewing paradigm, keeping the stimulation procedure as
610 similar as possible to Exp. 1. In this experiment, participants (N = 29) observed a sequence of dot-array
611 stimuli varying in numerosity, duration, and dot size (see Fig. 1B), and responded to occasional odd-ball
612 stimuli defined by a lower contrast (i.e., in order to ensure that participants kept watching and attending
613 the stimuli). We then employed again a multivariate decoding procedure to assess the encoding of past
614 stimulus information in visual-evoked activity. Since every stimulus in the sequence varied along the
615 three dimensions (i.e., instead of having a constant reference), the sorting of trials used to perform the
616 analysis differed from Exp. 1. Namely, in different iteration of the analysis, we took all the trials in which
617 the intermediate magnitude level of each dimension was presented (i.e., either 16 dots, 200 ms, or 6
618 pixels; henceforth called "current magnitude") and assessed the effect of different magnitude dimensions
619 of the stimulus presented in the preceding trial. The decoding was thus performed on EEG activity time-
620 locked to an identical "current" stimulus with the intermediate magnitude level (equivalent to the
621 reference stimulus in Exp. 1), as a function of the magnitude of the preceding one. In different iterations
622 of the analysis, we thus compared cases where the past stimulus had 12 vs. 24 dots, a duration of 140 vs.
623 280 ms, or a dot size of 4 vs. 8 pixels (for numerosity, duration, and size respectively; henceforth called
624 "past magnitude").

625 --- FIGURE 7 HERE ---

626

627 Fig. 7A shows the average classification accuracies relative to the decoding of different numerosities,
628 durations, or dot sizes of the preceding stimulus, while Fig. 7B shows the results of the null decoding
629 analysis performed with shuffled data. Overall, although generally weaker compared to Exp. 1, the
630 analysis was able to capture the signature of the past stimulus also in this case, while the null decoding

631 analysis did not show any consistent decoding. To better assess the signatures of different magnitude
632 dimensions, we again computed the average classification accuracy in two separate latency windows
633 (early window: 50-200 ms; late window: 500-650 ms; marked with shaded areas in Fig. 7A). The
634 classification accuracies observed within these two latency windows, corresponding to the different
635 combinations of current and past magnitude, are shown in Fig. 8A-B. Each sub-panel in Fig. 8A-B refers
636 to the "current" magnitude, while each bar refers to the decoding accuracy of the "past" magnitude. For
637 example, the first sub-panel in Fig. 8A shows the effect of numerosity (12 vs. 24 dots; blue bar), duration
638 (140 vs. 280 ms; red bar) and dot size (4 vs. 8 pixels; yellow bar) on stimuli having a numerosity of 16
639 dots (i.e., the intermediate numerosity of the range). Fig. 8C-D shows instead the effect of the different
640 "past" magnitude irrespective of the "current" magnitude (the magnitude of the current stimulus). Panels
641 C and D are indeed the average of the three sub-panels of Fig. 8A and 8B, respectively.

642

643 --- FIGURE 8 HERE ---

644

645 To assess the pattern of decoding across different magnitudes, we first performed a series of one-sample
646 t-tests against the corresponding average classification accuracy observed in the null decoding analysis.
647 Additionally, we applied FDR to control for multiple comparisons ($q = 0.05$). In the early latency window
648 (Fig. 8A), the results showed that the decoding of numerosity was significantly above chance only when
649 assessing the effect on numerosity in the current trial ($t(28) = 3.30$, adj-$p = 0.008$; other tests: $t(28) = -$
650 0.20-0.003, min adj-$p = 0.943$). Duration was instead significant in all cases ($t(28) = 2.47$-3.92, adj-$p <$
651 0.036). Finally, the decoding of size showed a significantly above chance accuracy only when considering
652 the effect of size on duration ($t(28) = 4.48$, adj-$p = 0.001$; other tests: $t(28) = 0.84$-2.18, min adj-$p =$
653 0.057). In the late latency window (Fig. 8B), instead, we did not observe any significant classification
654 accuracy after correction for multiple comparisons ($t(28) = -1.78$-2.60, min $p = 0.069$).

Similarly to Exp. 1, we then performed a three-way repeated measures ANOVA, with factors "current magnitude" (i.e., numerosity, duration, size), "past magnitude," and "latency window." The results showed no main effect of current magnitude ($F(2,56) = 1.613$, $p = 0.208$), no main effect of latency window ($F(1,28) = 0.950$, $p = 0.338$), but a significant main effect of past magnitude ($F(2,56) = 5.163$, $p = 0.006$, $\eta_p^2 = 0.155$). More importantly, we observed a significant three-way interaction ($F(4,112) = 2.868$, $p = 0.026$, $\eta_p^2 = 0.094$) between the current magnitude, past magnitude, and latency window.

To further explore this three-way interaction, we assessed the pattern of effects across different combinations of current and past stimulus magnitude separately at early (Fig. 8A) and late (Fig. 8B) latencies, performing two independent two-way ANOVAs with factor "current magnitude" and "past magnitude." In the early latency window, we observed no main effect of either current and past magnitude, and no interaction between the two factors (max $F = 2.533$, min $p = 0.089$), suggesting that although the decoding of numerosity (blue bars in Fig. 8A) appears to be weaker than the decoding of the other magnitudes (especially when considering duration and size as current magnitudes; middle and leftmost panel of Fig. 8A), the overall difference is not large enough to reach significance. In the late latency window (Fig. 8B) instead, we observed a significant main effect of the "past" magnitude ($F(2,56) = 3.611$, $p = 0.030$, $\eta_p^2 = 0.113$), no main effect of the "current" magnitude, and no interaction (max $F = 2.245$, min $p = 0.069$), suggesting a greater difference in the decoding accuracy between different past magnitudes. Thus, differently from the results of Exp. 1 (where we found that classification accuracies were more variable at early latencies), here there was a greater difference in the decoding accuracy between magnitudes at later latencies.

--- FIGURE 9 HERE ---

Since classification accuracies seemed again to increase above chance level before the stimulus onset, we assessed the temporal generalization matrices (Fig. 9) to better understand the nature of this effect, similarly to Exp. 1. In all three cases (concerning the different magnitude dimensions) the temporal generalisation provided little evidence for the presence of a lingering trace of the previous stimulus. Indeed, although the classification accuracies started increasing during the pre-stimulus interval, the temporal generalisation did not show any relatively stable pattern independent from the onset of the current stimulus. In all cases, the classification accuracy peaked after the onset of the current stimulus, suggesting an interaction between stimulus history and the processing of the current stimulus. As in Exp. 1, the pattern of classification accuracies before the onset of the current stimulus might represent either an anticipatory activation, or a by-product of the sliding window average used in the decoding analysis. If present, any lingering trace of the past stimulus would most likely interact with the processing of the current stimulus, showing an increase in classification accuracy at post-stimulus latencies rather than a decay.

Comparison between Experiment 1 and Experiment 2

Finally, we directly compared the results of Exp. 1 and Exp. 2, in order to assess the influence of task context on the neural signature of stimulus history. To do so, we used a mixed model ANOVA, with "reference magnitude" (i.e., coding for the different task conditions in Exp. 1, and for the different "current magnitudes" in Exp. 2), "past magnitude" (i.e., inducer magnitude in Exp. 1, and magnitudes of the previous trial in Exp. 2), and "latency window" (early vs. late) as within-subject factors, and "experiment" (Exp. 1 vs. Exp. 2) as between-subject factor. The results showed a main effect of past magnitude ($F(2,194) = 4.192$, $p = 0.018$, $\eta_p^2 = 0.068$), a main effect of latency window ($F(1,57) = 4.530$, $p = 0.038$, $\eta_p^2 = 0.073$), and a main effect of experiment ($F(1,57) = 21.723$, $p < 0.001$, $\eta_p^2 = 0.276$). We also observed a significant interaction between latency window and experiment ($F(1,57) = 7.665$, $p = 0.008$, $\eta_p^2 = 0.118$), suggesting that the signature of stimulus history at different latencies depends on whether

the experiment involves an engaging magnitude task, or a passive viewing of the visual stimuli. No other main effect or interaction reached significance (max $F = 2.289$, min $p = 0.061$).

--- FIGURE 10 HERE ---

We further performed a series of post-hoc tests assessing the average classification accuracy in the two temporal windows separately for each experiment (Fig. 10). First, a series of one-sample t-tests against chance level (i.e., average classification accuracy of the null decoding analysis), corrected with FDR, showed that in all cases classification accuracy was significantly higher than chance level (all t-values $\geq$ 3.15, all adj-$p \leq 0.004$). Then we performed a paired t-test (corrected with FDR) comparing the classification accuracy at different temporal latencies within each experiment. The results showed that while in Exp. 1 there was a significant difference between temporal windows, with higher classification accuracy at later temporal latencies ($t(29) = 2.63$, adj-$p = 0.026$, $d = 0.56$), there was no significant difference between latencies in Exp. 2 ($t(28) = 0.953$, adj-$p = 0.35$). Furthermore, we averaged together classification accuracies corresponding to different latency windows within each experiment, and compared the two experiments against each other. This test shows that classification accuracy are on average significantly higher in Exp. 1 compared to Exp. 2 (0.56 ± 0.009 vs 0.52 ± 0.003, respectively; independent-sample t-test, $t(57) = 4.66$, $p < 0.001$, $d = 1.25$).

DISCUSSION

In this study, we addressed the link between behavioural serial dependencies observed in perceptual tasks (Fischer & Whitney, 2014), and the neural signature of stimulus history emerging from visual-evoked potentials (Fornaciai & Park, 2018a). In doing so, we had two main goals: (1) understanding whether the neural signature of stimulus history is a direct correlate of the behavioural effect, and (2) pinpoint the

727 brain processing stages linked to the emergence of the behavioural bias. The nature of serial dependence
728 is indeed debated, and interpretations on the origin of this effect span from basic sensory/perceptual
729 processes (Burr & Cicchini, 2014; Fischer & Whitney, 2014; Michele Fornaciai & Park, 2018a) to high-
730 level processes related to memory or decision-making (Bliss et al., 2017; Fritsche et al., 2017; Pascucci et
731 al., 2019). Our study first demonstrates that while serial dependence is selective for the task at hand, the
732 neural signature of stimulus history encompasses multiple dimensions of the past stimulus, including
733 dimensions that do not yield a behavioural effect. Second, we observed a relationship between the
734 strength of the behavioural effect and the decoding accuracy of past stimulus information. Third, the
735 encoding of past information is weaker during passive-viewing, suggesting that attention or task-
736 relevance modulate the extent to which stimulus history is encoded in brain signals.

737 Regarding the behavioural results, our data highlight three main features of serial dependence in
738 magnitude perception. First, consistently with previous results (Togoli et al., 2021; Van der Burg et al.,
739 2019), the effect appears to be specific for the task. This in turn suggests that attention or other factors
740 like the task set likely play a role in determining the effect when different dimensions are simultaneously
741 manipulated, in line with the idea that serial dependence depends on attention (Collins, 2019; Fornaciai &
742 Park, 2018; Fritsche & de Lange, 2019; Manassi & Whitney, 2022). Second, we observed a repulsive
743 effect, akin to perceptual adaptation (Kohn, 2007), provided by the inducer duration in the numerosity
744 task. Although not statistically significant after correcting for multiple comparisons, this effect still
745 showed a medium effect size, and is in line with previous results (Togoli et al., 2021). Third, duration
746 perception seems in this context free from serial dependence, although previous results showed significant
747 effects in a similar task (Togoli et al., 2021). However, this is likely explained by the small range of
748 inducer durations, making the inducer too similar to the reference stimulus to yield robust effects.

749 At the neural level, EEG results show a clear signature of stimulus history, reflecting the encoding of past
750 information into the responses to a current stimulus. Strikingly, our results show a neural signature that is
751 more generalised compared to the behavioural effect. Brain responses to the reference stimulus indeed

752 incorporate not only the "task-relevant" information yielding attractive effects (numerosity in the
753 numerosity task, size in the size task), but a more complete representation of the different inducer
754 dimensions. This suggests a partial dissociation between the neural encoding of stimulus history and the
755 behavioural effect. Alternatively, the generalised neural effect might reflect influences that are actually
756 occurring at the behavioural level, but are too small to measure reliably. However, similarly to previous
757 studies (Togoli et al., 2021; Tsouli et al., 2019), our paradigm was able to capture an opposite (repulsive)
758 effect of duration on numerosity, with a similar effect size compared to the effect of numerosity. This
759 suggests that a lack of sensitivity of the current paradigm to cross-dimensional biases is unlikely to
760 explain the absence of those effects.

761 How can we explain such dissociation between behavioural and neural effects? First, our decoding
762 procedure might capture a lingering trace of the past stimulus rather than the stimulus history information
763 that affects perception/behaviour. Although classification accuracies in some cases increase even before
764 the stimulus onset, the temporal generalisation pattern observed is not consistent with a lingering trace of
765 the past stimulus. Second, another possibility concerns the effect of central tendency (e.g., Jazayeri &
766 Shadlen, 2010), which is similarly based on stimulus history and has been recently linked to serial
767 dependence (Tong & Dubé, 2022). However, since our analyses compared stimuli embedded in very
768 similar temporal contexts, we believe that it is unlikely that central tendency contributed to the observed
769 results. Finally, another possibility is that serial dependence would reflect only a subset of the past
770 stimulus dimensions encoded in brain signals due to the involvement of an active gating mechanism.
771 Namely, while the encoding of stimulus history carries a more complete representation of the past
772 stimulus (similarly to the encoding of task-relevant and task-irrelevant stimulus features in working
773 memory; Bocincova & Johnson, 2019), the mechanisms involved in generating serial dependence effects
774 would select and implement only the relevant information according to which dimension is highlighted by
775 attention and/or task demands.

776 Although weaker, a consistent signature of stimulus history is also evident in the passive-viewing
777 paradigm of Exp. 2. This suggests that performing a specific task is not strictly necessary to observe such
778 an effect (in line with previous results; Fornaciai & Park, 2018a), but that attention and/or task-related
779 processes may play a role in modulating it. For instance, attention might explain a stronger effect at
780 earlier latencies (i.e., via top-down modulation on early visual activity; Grothe et al., 2018; Somers, Dale,
781 Seiffert, & Tootell, 1999). On the other hand, the later amplification of classification accuracy observed
782 in Exp. 1 is completely missing in the results of Exp. 2, suggesting that it is specifically related to
783 performing an active task. A possibility is that this later peak might reflect the memory storage of
784 information after decision-making, which could make the biased stimulus representation to be encoded in
785 a more stable form compared to earlier perceptual processing (Oh, Kim, & Kang, 2019).

786 Importantly, we observed a link between the serial dependence effect and the brain signals encoding past
787 stimulus information. Indeed, at least in the numerosity and size task, the strength of the effect could be
788 significantly predicted by the classification accuracy obtained in the decoding analysis. The timing of this
789 significant relationship is early after the onset of the reference stimulus, starting at around ~35-65 ms and
790 persisting until ~250 ms. This in turn suggests that serial dependence emerges from early perceptual
791 computations, potentially starting at the earliest levels of visual processing. Such an early correlate is
792 consistent with previous studies proposing that the effect operates at the perceptual rather than decision-
793 making level (Cicchini et al., 2017; Manassi et al., 2018; Manassi & Whitney, 2022; Collins, 2020). Our
794 results provide direct evidence that early visual-evoked activity is indeed effectively linked to the
795 behavioural effect.

796 If a link between the neural signature of stimulus history and the behavioural serial dependence effect
797 emerges so early in the visual processing stream, why does serial dependence often show the properties of
798 a high-level effect? Previous results indeed show that serial dependence is partially independent from the
799 low-level features of the stimuli (Fischer et al., 2020), and works even across completely different stimuli
800 (Fornaciai & Park, 2019c). Additionally, it often relies on prior choices rather than prior stimuli (e.g.,

Pascucci et al., 2019), and works according to the perceived rather than physical properties of a stimulus (Fornaciai & Park, 2021). An interesting possibility explaining this aspect of the effect is a dissociation between where the bias *originates* from, and where it *operates*. Indeed, the bias itself may originate from high-level computations well before the onset of the current stimulus, and be transmitted back to earlier sensory stages via re-entrant feedback signals (Fornaciai & Park, 2019a; Fornaciai & Park, 2021), affecting perception directly. Our findings thus support the idea that while serial dependence likely originates from high-level computations, it operates at an early processing stage biasing the phenomenological appearance of a stimulus rather than just how we judge or remember it (see also Collins, 2019; Manassi & Whitney, 2022).

Finally, it is interesting to note that serial dependence and its neural signature were measured in our paradigms on a very short timescale, with sub-second intervals across successive stimuli. Previous studies show serial dependence effects on a longer timescale spanning multiple seconds (e.g., Fischer & Whitney, 2014). Thus, an interesting question is whether a signature of serial dependence at early perceptual stages, like the one here, would be present for serial effects measured at different timescales. In general, we believe that the neural signature shown here should not depend on the timescale of the effect, but this remains an open question that requires dedicated experiments.

To conclude, our findings highlight a set of important properties of serial dependence in magnitude perception and its neural signature. First, our results suggest the existence of an active mechanism gating past stimulus information according to its relevance. Second, we show that while an active task is not necessary for the encoding of stimulus history in brain signals, performing a task amplifies this neural signature. Finally, we show a link between behavioural serial dependence effects and brain activity at very early latencies, suggesting that serial dependence emerges during early perceptual processing.

Author contribution

MF and DB conceived the study and devised the stimuli. MF and IT collected and analysed the data, and prepared the figures. All the Authors wrote and revised the manuscript.

REFERENCES

Alais, D., Leung, J., & Van der Burg, E. (2017). Linear Summation of Repulsive and Attractive Serial Dependencies: Orientation and Motion Dependencies Sum in Motion Perception. *The Journal of Neuroscience*, *37*(16), 4381–4390. https://doi.org/10.1523/JNEUROSCI.4601-15.2017

Bliss, D. P., Sun, J. J., & D'Esposito, M. (2017). Serial dependence is absent at the time of perception but increases in visual working memory. *Scientific Reports*, *7*(1), 14739. https://doi.org/10.1038/s41598-017-15199-7

Bocincova, A., & Johnson, J. S. (2019). The time course of encoding and maintenance of task-relevant versus irrelevant object features in working memory. *Cortex*, *111*, 196–209. https://doi.org/10.1016/j.cortex.2018.10.013

Burr, D., & Cicchini, G. M. (2014). Vision: Efficient adaptive coding. *Current Biology*, *24*(22), R1096–R1098. https://doi.org/10.1016/j.cub.2014.10.002

Cicchini, G. M., Anobile, G., & Burr, D. C. (2014). Compressive mapping of number to space reflects dynamic encoding mechanisms, not static logarithmic transform. *Proceedings of the National Academy of Sciences*, *111*(21), 7867–7872. https://doi.org/10.1073/pnas.1402785111

Cicchini, G. M., Mikellidou, K., & Burr, D. (2017a). Serial dependencies act directly on perception. *Journal of Vision*, *17*(14), 6. https://doi.org/10.1167/17.14.6

847 Cicchini, G. M., Mikellidou, K., & Burr, D. C. (2017b). Serial dependencies act directly on perception.
848 *Journal of Vision*, *17*(14), 6. https://doi.org/10.1167/17.14.6

849 Cicchini, G. M., Mikellidou, K., & Burr, D. C. (2018). The functional role of serial dependence.
850 *Proceedings. Biological Sciences*, *285*(1890), 20181722. https://doi.org/10.1098/rspb.2018.1722

851 Collins, T. (2019). The perceptual continuity field is retinotopic. *Scientific Reports*, *9*(1), 18841.
852 https://doi.org/10.1038/s41598-019-55134-6

853 Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG
854 dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–
855 21. https://doi.org/10.1016/j.jneumeth.2003.10.009

856 Di Russo, F., Pitzalis, S., Spitoni, G., Aprile, T., Patria, F., Spinelli, D., & Hillyard, S. A. (2005).
857 Identification of the neural sources of the pattern-reversal VEP. *NeuroImage*, *24*(3), 874–886.
858 https://doi.org/10.1016/j.neuroimage.2004.09.029

859 Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience*, *17*(5),
860 738–743. https://doi.org/10.1038/nn.3689

861 Fornaciai, M., Brannon, E. M., Woldorff, M. G., & Park, J. (2017). Numerosity processing in early visual
862 cortex. *NeuroImage*, *157*, 429–438. https://doi.org/10.1016/j.neuroimage.2017.05.069

863 Fornaciai, M., & Park, J. (2018). Attractive Serial Dependence in the Absence of an Explicit Task.
864 *Psychological Science*, *29*(3), 437–446. https://doi.org/10.1177/0956797617737385

865 Fornaciai, M., & Park, J. (2019a). Spontaneous repulsive adaptation in the absence of attractive serial
866 dependence. *Journal of Vision*, *19*(5), 21. https://doi.org/10.1167/19.5.21

867 Fornaciai, M., & Park, J. (2019b). Serial dependence generalizes across different stimulus formats, but
868 not different sensory modalities. *Vision Research*, *160*, 108–115.
869 https://doi.org/10.1016/j.visres.2019.04.011

870 Fornaciai, M., & Park, J. (2020). Neural Dynamics of Serial Dependence in Numerosity Perception.
871 *Journal of Cognitive Neuroscience*, *32*(1), 141–154. https://doi.org/10.1162/jocn_a_01474

872 Fornaciai, Michele, & Park, J. (2018b). Serial dependence in numerosity perception. *Journal of Vision*,
873 *18*(9), 15. https://doi.org/10.1167/18.9.15

874 Fornaciai, Michele, & Park, J. (2019). Neural dynamics of serial dependence in numerosity perception.
875 *Journal of Cognitive Neuroscience*. https://doi.org/10.1162/jocn_a_01474

876 Fornaciai, Michele, & Park, J. (2020). Attractive serial dependence between memorized stimuli.
877 *Cognition*, *200*, 104250. https://doi.org/10.1016/j.cognition.2020.104250

878 Fornaciai, Michele, & Park, J. (2021). Disentangling feedforward versus feedback processing in
879 numerosity representation. *Cortex*, *135*, 255–267. https://doi.org/10.1016/j.cortex.2020.11.013

880 Fornaciai, Michele, & Park, J. (2022). The effect of abstract representation and response feedback on
881 serial dependence in numerosity perception. *Attention, Perception, & Psychophysics*, *84*(5), 1651–
882 1665. https://doi.org/10.3758/s13414-022-02518-y

883 Fritsche, M., & de Lange, F. P. (2019). The role of feature-based attention in visual serial dependence.
884 *Journal of Vision*, *19*(13), 21. https://doi.org/10.1167/19.13.21

885 Fritsche, M., Mostert, P., & de Lange, F. P. (2017). Opposite Effects of Recent History on Perception and
886 Decision. *Current Biology*, *27*(4), 590–595. https://doi.org/10.1016/j.cub.2017.01.006

887 Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2016). Decoding Dynamic Brain Patterns from
888 Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series
889 Neuroimaging Data. *Journal of Cognitive Neuroscience*, *26*(3), 1–21.
890 https://doi.org/10.1162/jocn_a_01068

891 Grothe, I., Rotermund, D., Neitzel, S. D., Mandon, S., Ernst, U. A., Kreiter, A. K., & Pawelzik, K. R.
892 (2018). Attention Selectively Gates Afferent Signal Transmission to Area V4. *The Journal of*

Neuroscience, *38*(14), 3441–3452. https://doi.org/10.1523/JNEUROSCI.2221-17.2018

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*, *13*(8), 1020–1026. https://doi.org/10.1038/nn.2590

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210. https://doi.org/10.1016/j.tics.2014.01.002

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3? *Perception ECVP 2007 Abstract Supplement*, *36*(14), 1–16. https://doi.org/10.1068/v070821

Kohn, A. (2007). Visual adaptation: physiology, mechanisms, and functional benefits. *Journal of Neurophysiology*, *97*(5), 3155–3164. https://doi.org/10.1152/jn.00086.2007

Liberman, A., Fischer, J., & Whitney, D. (2014). Serial dependence in the perception of faces. *Current Biology : CB*, *24*(21), 2569–2574. https://doi.org/10.1016/j.cub.2014.09.025

Manassi, M., Liberman, A., Kosovicheva, A., Zhang, K., & Whitney, D. (2018). Serial dependence in position occurs at the time of perception. *Psychonomic Bulletin & Review*, *25*(6), 2245–2253. https://doi.org/10.3758/s13423-018-1454-5

Manassi, M., & Whitney, D. (2022). Illusion of visual stability through active perceptual serial dependence. *Science Advances*, *8*(2). https://doi.org/10.1126/sciadv.abk2480

Meyers, E. M. (2013). The Neural Decoding Toolbox. *Frontiers in Neuroinformatics*, *7*(May), 8. https://doi.org/10.3389/fninf.2013.00008

Oh, B.-I., Kim, Y.-J., & Kang, M.-S. (2019). Ensemble representations reveal distinct neural coding of visual working memory. *Nature Communications*, *10*(1), 5665. https://doi.org/10.1038/s41467-019-13592-6

916 Pascucci, D., Mancuso, G., Santandrea, E., Della Libera, C., Plomp, G., & Chelazzi, L. (2019). Laws of
917 concatenated perception: Vision goes for novelty, decisions for perseverance. *PLOS Biology*, *17*(3),
918 e3000144. https://doi.org/10.1371/journal.pbio.3000144

919 Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into
920 movies. *Spatial Vision*, *10*(4), 437–442. https://doi.org/10.1163/156856897X00366

921 Samaha, J., Switzky, M., & Postle, B. R. (2019). Confidence boosts serial dependence in orientation
922 estimation. *Journal of Vision*, *19*(4), 25. https://doi.org/10.1167/19.4.25

923 Somers, D. C., Dale, a M., Seiffert, a E., & Tootell, R. B. (1999). Functional MRI reveals spatially
924 specific attentional modulation in human primary visual cortex. *Proceedings of the National*
925 *Academy of Sciences of the United States of America*, *96*(4), 1663–1668.
926 https://doi.org/10.1073/pnas.96.4.1663

927 St John-Saaltink, E., Kok, P., Lau, H. C., & de Lange, F. P. (2016). Serial Dependence in Perceptual
928 Decisions Is Reflected in Activity Patterns in Primary Visual Cortex. *The Journal of Neuroscience :*
929 *The Official Journal of the Society for Neuroscience*, *36*(23), 6186–6192.
930 https://doi.org/10.1523/JNEUROSCI.4390-15.2016

931 Togoli, I., Fedele, M., Fornaciai, M., & Bueti, D. (2021). Serial dependence in time and numerosity
932 perception is dimension-specific. *Journal of Vision*, *21*(5), 6. https://doi.org/10.1167/jov.21.5.6

933 Tong, K., & Dubé, C. (2022). A Tale of Two Literatures: A Fidelity-Based Integration Account of Central
934 Tendency Bias and Serial Dependency. *Computational Brain & Behavior*, *5*(1), 103–123.
935 https://doi.org/10.1007/s42113-021-00123-0

936 Tsouli, A., Dumoulin, S. O., te Pas, S. F., & van der Smagt, M. J. (2019). Adaptation reveals unbalanced
937 interaction between numerosity and time. *Cortex*, *114*(April), 5–16.
938 https://doi.org/10.1016/j.cortex.2018.02.013

939 Van der Burg, E., Rhodes, G., & Alais, D. (2019). Positive sequential dependency for face attractiveness
940 perception. *Journal of Vision*, *19*(12), 6. https://doi.org/10.1167/19.12.6

941 Watson, A. B. (1979). Probability summation over time. *Vision Research*, *19*(5), 515–522.
942 https://doi.org/10.1016/0042-6989(79)90136-6

943 Wehrman, J. J., Wearden, J., & Sowman, P. (2020). Decisional carryover effects in interval timing:
944 Evidence of a generalized response bias. *Attention, Perception & Psychophysics*.
945 https://doi.org/10.3758/s13414-019-01922-1

946 Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of
947 fit. *Perception & Psychophysics*, *63*(8), 1293–1313. https://doi.org/10.3758/BF03194544

948

949

950 **FIGURE LEGENDS**

951 ***Figure 1. Experimental procedure.*** *(A) Stimulation procedure used in Exp. 1. In Exp. 1, participants*
952 *performed either a numerosity, a duration, or a size discrimination task, in separate sessions. In each*
953 *trial, we presented a sequence of three dot-array stimuli modulated in numerosity, duration, and dot size:*
954 *a task-irrelevant "inducer" (either 12 or 24 dots, 140 or 280 ms, 4 or 8 pixels), a constant reference*
955 *(always 16 dots/200 ms/6 pixels), and a variable probe (varying in either numerosity, duration, or dot*
956 *size according to the task). At the end of each trial, participants reported which stimulus between the*
957 *reference and the probe either contained more dots, lasted longer in time, or had bigger dots*
958 *(respectively for the numerosity, duration, and size task). (B) In Exp. 2, we employed a passive-viewing*
959 *paradigm. Participants watched a stream of dot-array stimuli modulated in numerosity, duration, and dot*
960 *size. Each stimulus comprised either 12, 16, or 24 dots, was presented for 140, 200, or 280 ms, and*

961 *included items with size of 4, 6, or 8 pixels. Participants were instructed to attend the sequence and*
962 *respond to occasional oddball stimuli defined by a lower contrast.*

963

964 ***Figure 2. Behavioural results of Experiment 1.*** *(A) Behavioural results in terms of point of subjective*
965 *equality (PSE; i.e., representing accuracy in the task and the perceived magnitude of the reference) as a*
966 *function of different inducer magnitudes, limited to the task-relevant magnitude. (B) Average serial*
967 *dependence effect indexes computed as the normalized difference between PSEs in the two corresponding*
968 *inducer conditions, transformed into percentage. (C) Effect of the different inducer magnitudes on*
969 *behavioural performance, computed with a non-linear regression analysis (i.e., contribution of different*
970 *inducer magnitudes to the behavioural response in each trial). In this analysis, positive beta values*
971 *indicate an attractive effect (i.e., increased chance of responding "reference bigger" as the inducer*
972 *magnitude increases), and negative results indicate a repulsive (opposite) effect. Error bars are SEM.*

973

974 ***Figure 3. Average classification accuracy in Experiment 1.*** *Classification accuracy obtained in the*
975 *multivariate analysis, showing the signature of the three magnitude dimensions of the inducer decoded*
976 *from the EEG responses evoked by the reference, averaged across the three task conditions. The*
977 *classification accuracy shown here reflects the ability of a classifier (support vector machine) to*
978 *successfully classify the pattern of brain activity across multiple EEG channels evoked by the reference,*
979 *according to the inducer magnitude (i.e., low vs. high inducer numerosity, for example). Such procedure*
980 *was repeated across several 100-ms time windows throughout an epoch (-200:700 ms) time-locked to the*
981 *reference onset. (A) Classification accuracies obtained in the actual decoding analysis. The grey shaded*
982 *areas mark the two latency windows selected to perform further analyses (50-200 ms and 500-650 ms).*
983 *(B) Classification accuracies obtained in the "null" decoding analysis performed as a control and to*
984 *determine the chance level empirically. The vertical dashed line marks the onset of the reference stimulus,*

while the horizontal dashed line indicates the level of 50% accuracy. The coloured shaded areas represent the SEM.

Figure 4. Decoding results of Experiment 1 at early and late latencies. *(A) Average classification accuracies at the early latency window (50-200 ms), corresponding to the effect of inducer numerosity, duration, and size across the three task conditions (from the left to the right panel: numerosity, duration, and size task). (B) Average classification accuracies in the three task conditions at the late latency window (500-650 ms). The dotted line at each bar shows the empirical chance level computed from the null decoding analysis. (C) Classification accuracy corresponding to the effect of numerosity, duration, and size at the early latency window, averaged across the three tasks. (D) Classification accuracies at the late latency window, averaged across the three tasks. Error bars are SEM.*

Figure 5. Average temporal generalization matrices in Exp. 1*. The temporal generalization matrices are obtained by training and testing the classifier with brain activity at different latencies, to show the extent to which a given pattern of brain activity generalizes to later latencies. (A) Temporal generalization matrix concerning the effect of inducer numerosity, averaged across task conditions. (B) Temporal generalization matrix concerning the effect of duration. (C) Temporal generalization matrix concerning the effect of size. The horizontal and vertical dashed lines mark the off-diagonal generalization direction corresponding to the reference onset. The diagonal dashed line corresponds to the training and testing of the classifier at the same latency. The classification accuracies shown in Fig. 3A correspond to the diagonal of the temporal generalization matrices.*

Figure 6. Classification accuracies of the task-relevant dimensions in the three task conditions. *Pattern of classification accuracies obtained from the decoding of the magnitude dimension of the inducer*

consistent with the task performed, in the three task conditions. Lines at the bottom of the plot mark the significant time windows observed in the regression analysis, showing the latency windows at which the behavioural effect could be predicted from classification accuracy. The shaded areas represent the SEM. (B) Log-scaled classification accuracy plotted against the serial dependence effect, in the numerosity task condition. (C) Log-scaled classification accuracy plotted against the serial dependence effect, in the size task condition. Black lines are linear fit to the data.

Figure 7. Average classification accuracy in Experiment 2. *In Exp. 2, participants passively watched a stream of dot-array stimuli that were concurrently modulated in numerosity, duration, and dot size in a trial-by-trial fashion. In the multivariate decoding procedure, in different iterations of the analysis, we pooled all the stimuli with the intermediate level of either numerosity, duration, or dot size (named "current" magnitude in the description of the results), and decoded the signature of the preceding stimulus by contrasting trials in which the previous stimulus had a lower magnitude (i.e., either 12 dots, 140 ms, or 4 pixels, to assess the effect of numerosity, duration, or dot size, respectively; named "past" magnitude) against trials in which the previous stimulus had a larger magnitude (i.e., either 24 dots, 280 ms, or 8 pixels). The decoding was thus performed by considering the activity time-locked to an identical stimulus with an intermediate magnitude level, differing only in the magnitude of the preceding stimulus – similarly to the procedure used in Exp. 1. (A) Classification accuracies observed in the actual decoding analysis, corresponding to the different magnitudes of the past stimulus. The two grey shaded areas mark the latency windows selected to perform data analysis, as in Exp. 1. (B) Classification accuracies obtained in the null decoding analysis, serving as a control and to determine the chance level empirically. The vertical dashed line marks the onset of the reference stimulus, while the horizontal dashed line indicates the level of 50% accuracy. Coloured shaded areas represent the SEM.*

Figure 8. Decoding results of Exp. 2 at early and late latencies. *(A-B) For each magnitude dimension of the current stimulus, i.e., "current magnitude" (different subplots) we plotted the decoding accuracy of the "past" magnitude in the early (A, 50 to 200 ms) and late (B, 500 to 650 ms) latency window. Namely, from left to right, the bars indicate the classification accuracy corresponding to the effect of numerosity (blue), duration (red), and size (yellow) of the past stimulus on the numerosity (leftmost panel), duration (middle panel), and size of the current stimulus (rightmost panel). The dotted lines shown at each bar mark the empirical chance level computed from the null decoding analysis. (C) Average effect of the magnitudes of the previous stimulus on the current one in the early latency window, collapsing together the different magnitudes of the current stimulus. (D) Average effects of different magnitudes in the late latency window. Error bars are SEM.*

Figure 9. Temporal generalization matrices of Exp. 2*. (A) Temporal generalization matrix concerning the effect of the numerosity of the past stimulus. (B) Temporal generalization matrix concerning the effect of duration. (C) Temporal generalization matrix concerning the effect of size. The horizontal and vertical dashed lines mark the off-diagonal generalization direction corresponding to the current stimulus onset. The diagonal dashed line corresponds to the training and testing of the classifier at the same latency. The classification accuracies shown in Fig. 7A correspond to the diagonal of the temporal generalization matrices.*

Figure 10. Comparison between Exp. 1 and Exp. 2. *(A) Average classification accuracy across all the conditions and dimensions tested in Exp. 1. The results of Exp. 1 show that while a signature of the inducer magnitude information is on average measurable from neural signals very early after the reference stimulus onset (50-200 ms), such a signature is strongly amplified at later latencies (500-650 ms). (B) Average classification accuracies obtained in Exp. 2. Although significantly higher than chance*

level (0.5), in Exp. 2 the decoding performance shows an overall weaker encoding (lower classification accuracy) of the inducer magnitudes during reference processing, compared to Exp. 1. At the late latency window, no amplification of these signals was observed. Error bars are SEM.