Attention, Perception, & Psychophysics (2019) 81:1962–1978
https://doi.org/10.3758/s13414-019-01697-5

What makes a prototype a prototype? Averaging visual features in a sequence

Ke Tong[1] · Chad Dubé[1] · Robert Sekuler[2]

Published online: 1 May 2019
© The Psychonomic Society, Inc. 2019

Abstract

After viewing a series of sequentially presented visual stimuli, subjects can readily generate mean representations of various visual features. Although these average representations seem to be formed effortlessly and obligatorily, it is unclear how such averages are actually computed. According to conventional prototype models, the computation entails an equally weighted average taken over all the stimuli. To test this hypothesis, we had subjects estimate the running averages of some feature in a series of sequentially presented stimuli. Part way through the series, we perturbed the distribution from which stimuli were drawn, which allowed us to test alternative models of the computations behind subjects' estimates. In both explanatory and predictive tests, a model in which the most recent items had disproportionate high weight outperformed a model in which all items carried equal weight. Such recency-weighted behavior was shown consistently in multiple experiments in which subjects estimated running averages of length of vertical lines. However, the degree to which recent items were prioritized varied with the type of stimulus, such that when estimating the running averages of a series of numerals, subjects showed less recency prioritization. We conclude that previous evaluations of prototype models have made unrealistic assumptions about the nature of a prototype, and that a reassessment of prototype models of visual memory and perceptual categorization may be in order.

Keywords Summary statistical representation · Ensemble representation · Prototype model · Visual short-term memory · Perceptual categorization

Introduction

A fundamental question in cognitive neuroscience is how the brain deals so effectively with the overwhelming amount of sensory input it receives, encoding fine details of selectively attended stimuli, while also retaining a stable representation of the larger environmental context. Various theoretical accounts postulate that some balance is struck between the quality and quantity of information that is extracted and maintained, but the details of that balance between quality and quantity are unresolved.

On one hand, many influential accounts of perception, memory, and categorization assign a central, or even exclusive, role to the outcome of matches between a probe stimulus' features and features of individual items in memory (Estes, 1994; Hintzman & Ludlam, 1980; Nosofsky et al., 2011; Shiffrin & Steyvers, 1997). On the other hand, some accounts of both perceptual and short-term memory tasks assume that summary statistical representations (also called ensemble representations), extracted from sets of stimuli, are key to subjects' performance, even when memory for individual items is reduced to chance (Ariely, 2001; Corbett & Oriet, 2011). Moreover, memory for averages appear to influence memory retrieval even in the absence of instruction or encouragement to report or compute an average (Dubé & Sekuler, 2015). Such evidence suggests

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13414-019-01697-5) contains supplementary material, which is available to authorized users.

✉ Ke Tong
ketong@mail.usf.edu

[1] Department of Psychology, University of South Florida, Tampa, FL 33620, USA

[2] Volen Center for Complex Systems, Brandeis University, Waltham, MA 02453, USA

that, at least in some cases, summary representations may be obligatorily encoded, stored, retrieved, and deployed.

These findings raise important questions, such as (i) whether summary representations rely on mechanisms that are distinct from those that encode and store individual item representations; (ii) how ensemble and item representations impact memory retrieval; (iii) how statistical moments are represented at the neural level; and (iv) whether sequential and spatial averaging rely on the same or different mechanisms. To answer these questions, computational models seem to be indispensable.

Of course, any computational modeling of summary statistical representation must include (i) the number of items that enter into the computation, and (ii) the form of the computation. To answer the first question, Whitney and Leib (2018) pooled results from 21 related studies and argued that subjects effectively integrate approximately the square root of the number of all displayed objects (Whitney & Leib 2018). As our study focuses on tasks in which subjects average sequentially presented stimuli (henceforth, "sequential averaging"), we evaluated this square root relationship only for the five sequential averaging studies. Figure 1 shows that results of those five studies deviate substantially from a square root relationship. As a result, we are cautious about taking the square root relation as a given for incorporation into a model of sequential averaging.

Answers to the second question, of how the stimuli are combined in the average, are also not consistent across studies. For instance, in studies focused on perceptual classification such questions are often not even posed. Instead, it is assumed that, if an average or "prototype" were computed, it would be an equally weighted average over all prior stimuli (Nosofsky, 1987; Smith & Minda, 2000). However, results from some studies of ensemble perception are consistent with non-equally weighted averaging schemes (Juni et al., 2012; Hubert-Wallander & Boynton, 2015).

In the following, we first discuss the critical findings regarding sequential averaging in visual short-term memory (VSTM). We then present experiments designed to identify the weighting scheme that subjects apply when producing estimates of averages over trials. We conclude that the averages extracted from a sequence of stimuli reflect a recency-prioritized weighted average. We discuss the implications of our findings for existing models of memory and perceptual categorization, and underscore the need for a reassessment of prototype models in these domains.

Weighting scheme in sequential averaging

What ensemble features are encoded in sequential averaging tasks? Subjects could form representations of the mean over time when they are explicitly instructed to do so (Albrecht & Scholl, 2010; Hubert-Wallander & Boynton, 2015) or without explicit instruction (Dubé & et al. 2014; Oriet & Hozempa, 2016). Other than the mean, variance priming (Michael et al., 2014) and perceptual adaptation of variance (Norman et al., 2015) suggest implicit encoding of the variance information. One recent study, Chetverikov et al. (2016), also established that subjects implicitly encode the entire feature distribution of the distractors in visual search tasks over time.

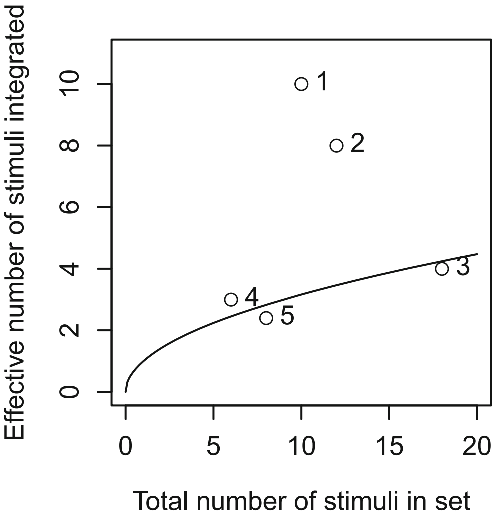

Fig. 1 The proposed square root relationship between total number of stimuli presented and the effective number of stimuli integrated (*solid line*, Whitney & Leib 2018) does not fit the sequential averaging studies very well. Labels beside data points indicate the following sources: 1. Hubert-Wallander and Boynton (2015), 2. Leib, Kosovicheva, & Whitney (2016), 3. Leib et al. (2014), 4. Piazza et al. (2013), 5. Florey, Dakin, & Mareschal (2017). Figure 1 is reproduced from Figure 4 in Whitney & Leib (2018), using data points from sequential averaging studies only

The current study focuses on the ensemble perception of sequentially presented stimuli, especially the mechanism of item integration in sequential averaging. While conventional prototype models typically assume an averaging process in which all items are weighted equally, an alternative hypothesis assumes that subjects' estimates of averages give more weight to the most recent items. A recency-prioritized weighting scheme has been demonstrated for the sequential averaging of many features, e.g., size, emotion, and motion directions, although not for the averaging of spatial locations (Hubert-Wallander & Boynton, 2015).

The recency hypothesis for averaging assumes that although the average may be stored in long-term memory, the items factoring into its computation on a given trial may be in various stages of serial position-dependent decay (Wilken & Ma, 2004; Huang & Sekuler, 2010). Since recent items have stronger memory representations at the time of an average's computation, those items will be given more weight in that computation. In fact, that differential weighting would be consistent with a mathematically optimal strategy for item integration, in which subjects assigned more weights to items at serial positions with less noise (Juni et al., 2012).

In what follows, we report four experiments meant to identify the weighting scheme that subjects use when they report running averages over sequentially presented stimuli. In doing so, we include models based on both serial positions and temporal positions, and with multiple weighting schemes. Our results support the operation of an averaging computation that is recency-weighted. This suggests that prototypes are not simple averages, and that a reconsideration of prototype models is in order.

Experiment 1

Mean-shift design

Most previous studies of sequential averaging drew stimuli randomly from a single distribution, and at the end of a series of stimuli, subjects estimated the mean of what they had seen (e.g., Hubert-Wallander & Boynton 2015). The current study introduces two design changes. The first change is that after every new stimulus, subjects report the mean value of all the stimuli they have seen up to that point (Weiss & Anderson, 1969). This greatly increases the amount and grain of the resulting data, supporting more efficient and reliable modeling analysis about the weighting schemes. In a second design change, the mean value of the distribution from which stimuli were drawn shifted midway in the sequence (Parducci, 1956). The advantage of this mean-shift design is to enhance the discriminability of alternative weighting schemes in the empirical data, by examining subjects' estimates in the aftermath of the shift.

In short, the mean-shift design allows fine-grained tracking of subjects' estimates of the running means and provides better differentiation between different item integration mechanisms. We elaborated the advantages of the design changes by a simulation with two quite different weighting schemes. Specifically, we simulated ideal observers' estimates of the running means assuming a model, *Equal*, in which subjects gave equal weight to all items, and a model, *Recency*, in which subjects utilized all previous items, but prioritized more recent items.

The simulated estimates and stimuli are plotted in Fig. 2, showing that when all the stimuli were from the same Gaussian distribution (left panel), the simulated responses from the two weighting schemes (solid black for *Recency* and dashed black for *Equal*) were largely overlapped, making it difficult to tell which weighting scheme was in use. However, when there was a mean shift in the stimulus values (right panel), the two weighting schemes were clearly differentiated after the mean shift. So, compared to an experiment in which only a single stimulus distribution is used, suddenly shifting the mean of the stimulus distribution can highlight the weighting scheme that subjects are using. With this fact in mind, we incorporated the mean shift design in Experiment 1 and the other experiments.

In addition to the above benefits, we also wonder if the mean-shift design could alter subjects' item integration mechanism, specifically, promoting a higher degree of recency weighting after the mean shift. On post-shift trials, stimuli from the pre-shift distribution may be weighted less or discounted in item integration due to their significant differences from the ongoing stimulus distribution, since

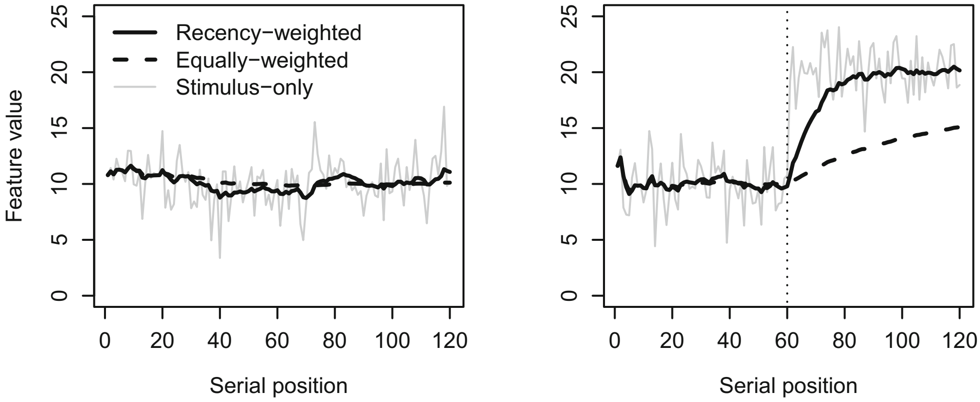

Fig. 2 Benefits of the design changes: By asking subjects to estimate the running means after presentation of every stimulus and adding a shift of the distribution mean, the simulated responses using two different pre-determined weighting schemes are more clearly separated. *Left panel*: mean = 10. *Right panel*: pre-shift mean = 10, post-shift mean = 20. The standard deviation of stimuli was kept constant (SD = 2) for the left panel and both pre- and post-shift parts of the right panel. The *Recency* scheme uses exponentially decaying weights with a rate of 0.9 (see Modeling analysis section of Experiment 1 for more details). The *vertical dotted line* in the right panel denotes trial 60, the last trial on which a stimulus was sampled from the pre-shift distribution

it has been shown that outliers may be excluded in mean estimations (Haberman & Whitney, 2010).

To address this potential confound, we asked each subject to complete three sequences with different mean-shift configurations, "Small Shift", "Large Shift", and "Large Variance" (See Methods section for details). If the mean shift does promote recency in item integration, we would expect more recency in the "Large Shift" condition where the shift is most distinctive, and less recency in the "Large Variance" condition, where the elevated variance makes the mean-shift less obvious to subjects.

Methods

Subjects

Fifteen University of South Florida undergraduates participated in the experiment for course credit (ten female, mean age = 19.53 years, SD = 1.36 years). All had normal or corrected-to-normal vision. All procedures were approved by the IRB of University of South Florida.

Procedure

Subjects were presented with a sequence of gray vertical lines, one at a time. Each line was displayed in the center of the screen for one second. After each gray line, subjects were asked to estimate the average length of *all* the gray lines they had seen to that point in the sequence by using up and down arrow keys on a computer keyboard to adjust the length of a white probe line to represent that estimate. This adjustable, white probe was presented on the screen immediately after the disappearance of each gray line. Stimuli and probes were presented in different colors to reduce the potentially confounding influence of the probe length on estimates of prior stimuli. When subjects completed an adjustment, they pressed the keyboard's Enter key to proceed. The next stimulus appeared on the screen right after the subject submitted his/her estimate. The same presentation procedure was used in Experiments 1 and 2, as detailed in Fig. 3.

Prior to the experiment, the QUEST (Watson & Pelli, 1983) routine was used to measure each subject's Weber fraction for line length. On each trial, QUEST controlled the successive presentation of two vertical lines (500 ms each, 1000 ms ISI) at the center of the screen. After each pair of stimuli, subjects judged which had been longer. Feedback was given after each response ("correct" or "wrong"). Subjects' individual Weber fractions were obtained from a QUEST run of 40 trials and were used to scale all stimuli in just noticeable difference (JND) units for the experiment.

The stimuli were scaled with individual Weber fractions and a base length of 100 pixels, so the actual stimulus size in pixels was $100*(1+wb)^x$, where wb is the Weber fraction and x is the JND value (specific JND values are detailed in the next section). The prior for the Weber faction used in QUEST had mean = .03 (Teghtsoonian, 1971) and SD = .04. The large values of SD provided a vague prior. For the 20 subjects (15 from Experiment 1 and five from Experiment 2), the mean Weber fraction was .064 and SD was .016.

Subjects were provided with detailed instructions and practice trials to ensure their understanding of the task. The instructions can be found in the Supplementary Materials.

Design and stimuli

The stimulus values presented in this paragraph are all in JND units. In the "Small Shift" condition, the pre- and post-shift means are 15 and 25, with a SD of 5. The "Large Shift" condition used a larger mean shift (pre/post-shift means = 15/30, SD = 5). The "Large Variance" condition used a larger variance across the sequence (pre/post-shift means = 15/25, SD = 8). The order of the three mean-shift conditions was counterbalanced across subjects.

For each sequence, line lengths were drawn randomly from one of two Gaussian distributions with different means but the same SD. Line lengths were sampled from the range spanned by ±2 SDs around a distribution's mean. Each

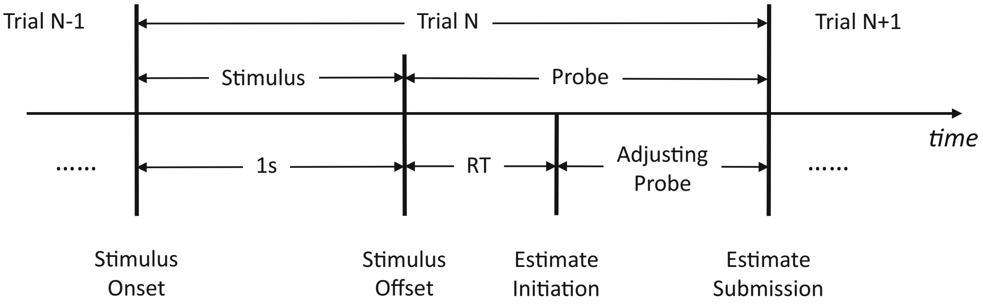

Fig. 3 Stimulus and probe presentation time in Experiments 1 and 2. A stimulus was presented for one second, and the probe was presented immediately after the stimulus, for a duration that depended on subjects' reaction time and the time for adjusting the probe. The completion time for each trial was recorded, allowing modeling of the data with both serial positions and temporal positions

entire stimulus sequence comprised 120 trials, split equally between pre- and post-shift trials. Starting on the 61st trial, the mean value of the distribution from which stimuli were drawn was altered. Subjects were not informed that stimuli might change during a sequence.

With the exception of the very first trial, the initial length of the adjustable probe line on successive trials was set to the value of the subject's immediate prior response. On the first trial, the length of the first probe was fixed at 5.4 degrees of visual angle (Experiment 2 examines whether the probe's initial value affects judgments). Stimuli were presented on a Dell 1905FP LCD computer monitor, with a resolution of 1280 × 1024, at a viewing distance of approximately 60 cm.

Because actual stimulus sizes were personalized for each subject based on their individual Weber fraction, the actual stimulus sizes were different for each subject. Across all subjects and all conditions, the mean and SD of stimulus lengths in visual angle were 9.63 and 5.85 degrees.

Modeling analysis

The data comprise subjects' successive estimates of the running means and the stimulus values. For each complete sequence, 120 data points were recorded.

The data were fitted with three models representing different item integration mechanisms, namely the *Equal*, *Recency*, and *Compression* model. The *Equal* model predicts the mean estimates to be the actual running averages of the stimuli. All items were weighted equally in the averaging process, regardless of their serial positions. We included the *Equal* model as a null model to compare with the following two models.

In the *Recency* model, subjects' estimates were modeled as the dot product of a stimulus vector and a weight vector (Weiss & Anderson, 1969; Juni et al., 2012; Hubert-Wallander & Boynton, 2015). The weight vector is recency-prioritized (Newer items have more weights). Additionally, a bias term was added to capture any systematic bias in observers' estimates (3).

$$s_i = (s_1, s_2, \ldots, s_i) \tag{1}$$

$$w_i = r^{\{1:i\}} / \sum r^{1:i} \tag{2}$$

$$Est_i = w_i \cdot s_i + bias + \varepsilon. \tag{3}$$

In the serial position-based modeling analysis, the weight vector is assumed to be strictly serial-position dependent. The *Recency* weights were modeled as a normalized exponential function defined over the serial position of successive stimuli (Brown et al., 2007). The exponents represent the serial positions of the stimuli. The rate parameter, r, ranges from 0 to 1, allowing the model to capture different degrees of recency prioritization: a smaller r indicates a higher degree of recency prioritization, and when r equals 1, w_i reduces to $1/i$ for each of the i stimuli, representing *Equal* averaging. Dividing by the summed weights of all stimuli within a trial normalizes the weight term, so that all weights sum to one for each single trial's estimation. The r parameter is responsible for the shape of the weight distributions, thus it is the parameter of interest. We aim to evaluate the best-fitting r values and compare them to the null hypothesis of an *Equal* weighting scheme ($r = 1$).

We also tested an alternative construct of the *Recency* model using a normalized power function to model the weights. The power-based model performed worse than the exponential-based model, so we kept the *Recency* construct as specified in Eq. 2. See Supplementary materials for details.

It is worth noting that in sequential averaging studies, the same stimulus can be characterized in terms of either temporal position (e.g., the item was presented x seconds ago) or serial position (e.g., the item was presented y items back), so, the influences on the averaging computation may come from factors of either temporal or serial positions, or both. To address this issue, we ran an alternative temporal position-based modeling analysis, in which the weight vector was assumed to be strictly time-dependent. The centers of stimulus presentation durations were used as temporal positions. For each trial, the prior items' temporal distances (in seconds) to the newest item were used as the exponents over the rate parameter. So, each prior item was weighted based on its temporal distance to the current trial. This temporal position-based *Recency* model was added to the model comparison. In the following sections, we termed these two models as *Recency-s* (serial) and *Recency-t* (temporal).

The *Compression* model provides an alternative account in which subjects complete their estimation by updating their immediate previous estimation (a single "compressed" representation of the previously shown items) with the newest stimulus. This strategy is plausibly encouraged as a strategy in Experiment 1 because subjects are asked to frequently estimate the running averages and the initial value of the adjustable probe on each trial is set to subjects' estimation on the previous trial.

$$Est_i = w_{old} \cdot Est_{i-1} + w_{new} \cdot s_i + bias + \varepsilon \tag{4}$$

$$w_{old} = (i-1) \cdot f \cdot w_{new} \tag{5}$$

$$w_{old} + w_{new} = 1 \tag{6}$$

In the *Compression* model, an ideal observer should adjust the relative weights of their previous estimation and the new stimulus, by putting less weight on the new stimulus as the sequence extends to include additional stimuli. We modeled subjects' estimations as a weighted average of their previous estimation ("old") and the newly shown item ("new"), plus a constant bias term and random noise. This description is summarized in Eq. 4. Constraints on the weights for "old" and "new" items are summarized in Equations 5 and 6 below. To elaborate, the weighting relationship between the "old" and "new" terms is modulated by a factor parameter f, which ranges from 0 to 1. When $f = 1$, the new stimulus will take a weight of $1/i$ in the estimation, which is the ideal ratio for the task and relates to equally weighted averaging. When $f = 0$, the weight on a new item is 1, which means subjects rely solely on the newest item.

For each model, we separately fit each stimulus sequence from each subject. To obtain the best-fitting parameters, we computed the sum of squared error between model predictions and observed estimates and minimized the error term using the "L-BFGS-B" bounded optimization method (Byrd & et al. 1995). The initial value used for r and f was 1, with the boundaries set to (0, 1]. The initial value used for the bias term was 0, bounded at [-50, 50].

Results and discussion

Data from three representative subjects are plotted in Fig. 4 (All individual plots can be found in Supplementary Materials). Data from all mean shift types (SS for "Small Shift", LS for "Large Shift", and LV for "Large Variance") showed a general pattern, namely that subjects' estimates of the mean did not follow predictions from the *Equal* model.

After the mean shift (the vertical dotted line in the middle of the sequence denotes the final pre-shift trial), the influence of the recent stimuli grew more evident as subjects' estimates rose toward the mean of the post-shift distribution, increasingly deviating from the equally weighted moving averages. Individual differences were observed. For instance, Subject 9's estimates closely varied with the new stimulus, demonstrating a greater influence of the most recent item. Subject 15's estimates changed less in the post-shift trials. Subject 12's estimates reflected a degree of *Recency* in between. Despite the individual differences we observed, no subject showed estimates that aligned with predictions based on the *Equal* model.

These results suggest that that subjects' estimates were unlikely to arise from the *Equal* model. In the following section, we compared the performance of the two non-equal models: *Recency* and *Compression*.

Model comparison

For each model, we conducted model fitting for all sequences separately. The best-fitting parameters are shown in Table 1. To evaluate the performance of non-equal models, we conducted both explanatory and predictive tests.

In the explanatory test, all the observed data were fitted with the competing models. As a result, we obtained best-fitting parameters for each model, and compared the model fits to the data using root mean squared error (RMSE). The model with the smallest RMSE "explains" the observed data the best.

Note that a model that excels in the explanatory test can fail in the predictive test, in which part of the observed data are used to obtain parameters to predict the remaining data that are not used in parameter training. One notable reason is overfitting, suggesting that if explanatory performance is the sole mode of assessment, the model could end up fitting meaningless noise and error in the data. Unfortunately, a lack of predictive assessment is common in psychological modeling studies (Shmueli, 2010; Yarkoni & Westfall, 2017). We adopt the suggestions from Shmueli (2010) to treat explanatory and predictive performance as two dimensions of model performance assessment.

Explanatory tests Explanatory performances were summarized in Table 2. The three non-equal models (*Compression*, *Recency-s*, and *Recency-t*) consistently outperformed the *Equal* model, again suggesting the *Equal* model is least likely to capture the item integration mechanism among competing models.

Among the non-equal models, the *Recency* models (both *Recency-s* and *Recency-t*) outperformed the *Compression* model. This result suggests that subjects are more likely to integrate multiple recent items (*Recency* models) rather than updating a compressed prior estimation (*Compression* model).

The explanatory performances of the two *Recency* models are almost identical. The mean RMSE difference between the two *Recency* models is 0.003, compared to the mean RMSE difference from *Compression* model (1.19) and *Equal* model (27.99). This is due to high correlation (mean correlation coefficient = 0.99) between serial and temporal positions used in the two *Recency* models. So, from the current design and analysis, it is unclear whether the cause of *Recency* weighting is from serial positions or temporal positions. Future studies could make timing controls more specific to separate the influence from these two factors.

Going forward, we will use *Recency-s* as the representative model due to its excellent explanatory performance. The three mean-shift conditions ("Small Shift", "Large Shift", and "Large Variance") did not affect the recency rate parameters, $F(2, 28) = 0.041$, $p = 0.96$. If the abrupt up-shift of the mean had significantly influenced subjects' item integration, we would expect the modeling results to differ among the three conditions. The similar best-fit parameters among mean-shift conditions suggest this possibility is unlikely.

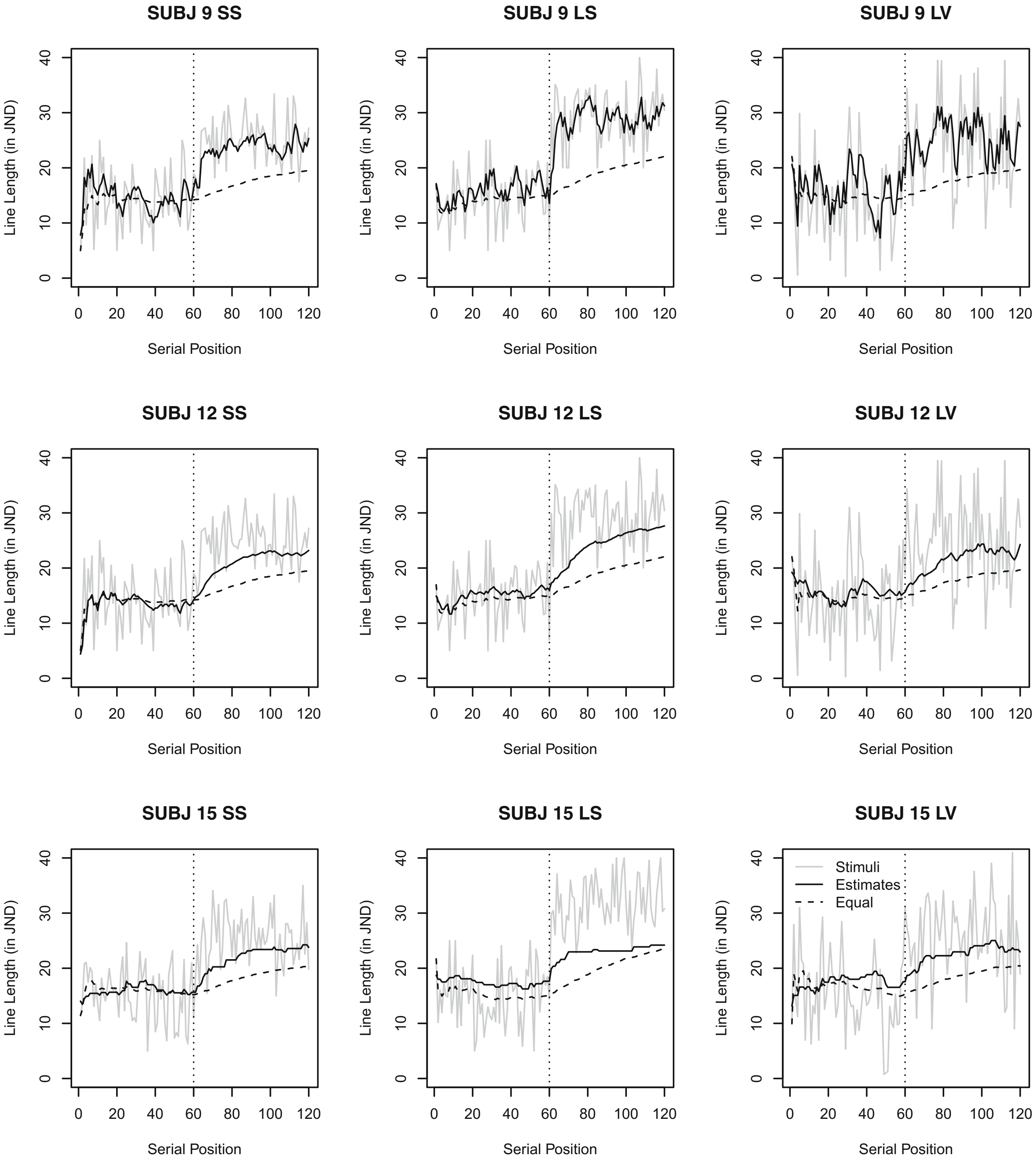

Fig. 4 Data from three representative subjects in Experiment 1. The *black solid lines* represent subjects' estimates. The *black dashed lines* represent predictions from the *Equal* model. The *gray solid lines* represent the stimuli. Before the mean shift (marked by the *vertical dotted line*, the first post-shift trial being the 61st), subjects' estimates generally hover around the mean of the pre-shift distribution. After the mean shift, subjects start to overestimate the running means, shown as the upward departure of the estimates (*black solid*) from the equally weighted moving averages (*black dashed*). This pattern is observed in all mean shift conditions (SS: small shift, LS: large shift, LV: large variance). Individual differences were observed. Subject 9's estimates closely varied with the new stimulus, demonstrating a greater influence of the most recent item. Subject 12's estimates reflected a moderate recency-weighted scheme. Subject 15's estimates were more stable in the post-shift trials. Nevertheless, none of the subjects showed estimates that overlapped with the equally weighted moving average. All individual plots can be found in Supplementary Materials

Table 1 Best-fitting parameters of Experiment 1

Model		SS	LS	LV
RS	r	0.86 (0.12)	0.86 (0.1)	0.86 (0.13)
	bias	0.7 (1.98)	0.57 (1.8)	1.8 (1.91)
RT	r	0.94 (0.06)	0.94 (0.04)	0.96 (0.06)
	bias	0.73 (1.97)	0.57 (0.88)	1.83 (1.88)
CM	f	0.35 (0.37)	0.29 (0.3)	0.45 (0.42)
	bias	0.05 (0.13)	0.07 (0.11)	0.16 (0.2)

Models are *Recency-Serial* (RS), *Recency-Temporal* (RT), and *Compression Model* (CM). The mean-shift conditions are "Small Shift" (SS), "Large Shift" (LS), and "Large Variance" (LV). Best-fitting parameters are presented in a "mean (SD)" format over subjects (N = 15)

Since the mean-shift manipulation did not affect the modeling parameters, we averaged each subject's best-fitting *r* parameters from the three conditions. The group average of the *r* parameter is 0.86, suggesting a recency prioritization (this group-averaged best-fitting *r* was used to estimate the effective number of items integrated later in this paper). A one-sample *t* test rejects the *Equal* null hypothesis, $t\,(14) = -4.71$, $p < 0.001$.

Predictive tests In the predictive tests, we used the averaged best-fitting parameters from ten randomly sampled subjects to predict data from a new subject. For training, best-fitting parameters from all mean shift conditions were averaged together. For testing, the sequence at test was randomly sampled from a new subject whose responses were not used in the parameter training. This predictive process was repeated with random sampling 100 times and predictive performance (*RMSE*) was averaged over iterations.

In general, the non-equal models outperformed the *Equal* model (Fig. 5). Among the non-equal models, the *Recency* models were better than the *Compression* model. Detailed results are reported in Table 2 (Predictive RMSE).

Table 2 Mean explanatory and predictive RMSE of Experiment 1. Models are *Recency-Serial* (RS), *Recency-Temporal* (RT), *Compression Model* (CM), and *Equal* (EQ)

Models	Explanatory RMSE	Predictive RMSE
RS	**1.90 (1.44)**	**8.67 (8.20)**
RT	1.90 (1.44)	8.68 (8.19)
CM	3.10 (2.92)	10.88 (9.23)
EQ	29.89 (18.63)	33.43 (18.76)

Mean RMSE values are presented in a "mean (SD)" format. The explanatory performance was summarized over subjects (N = 15) and the predictive performance was summarized over repetitions (N = 100). The best model performance (lowest RMSE) for each test is marked in *bold*

Model comparison results In both explanatory and predictive tests, the *Equal* model performed the worst among all models. Does the model performance of the non-equal models benefit from adding the bias term? We tested an alternative form of the *Equal* model with a bias term to capture systematic under- or over-estimation. In both explanatory and predictive tests, the *Recency-s* model outperformed this Equal model (*p*s < 0.001). So, the determinant of model performance is not the bias term but the weight distribution in item integration.

In the non-equal models, the *Recency* model outperformed the *Compression* model in both tests. The performance of *Recency* models based on serial positions and temporal positions are close to each other in both tests (Table 2).

To sum up, the *Equal* averaging scheme does not seem like a plausible explanation or prediction mechanism for subjects' estimates of the running means of sequentially presented stimuli. To estimate the running averages, subjects are more likely to utilize multiple recent representations rather than updating a single compressed representation of all prior stimuli. Subjects are likely to assign more weight to more recent items in their item integration.

Effective number of items integrated (ENI)

Both explanatory and predictive tests favored the *Recency* weighting scheme. The averaged best-fitting rate parameter for the *Recency-s* model was 0.86, suggesting that in a task of tracking the running means of sequentially displayed stimuli, subjects rely more on recent stimuli, rather than treating all items equally. The group-averaged weight distributions, plotted in the left panel of Fig. 6, show that a few of the most recent stimuli accounted for the majority of the weights, leaving other older stimuli a small fraction of the total weights to split between them.

To quantify this observation, we defined Effective Number of items Integrated (ENI) as the fewest items that are needed to accumulate weights over a certain threshold. We found that to account for 90% and 95% of the cumulative weight required the most recent 16 and 20 stimuli, respectively. Since the *Recency-s* model assumes the weights follow an exponential function of serial positions, the weights assigned to these prioritized items are not uniformly distributed over the items either (e.g., the most recent five stimuli alone provide >50% of the cumulative weights). Hence, a conclusion on ENI depends on the criterion that is used. For the current study using vertical lines, $\text{ENI}_{90} = 16$, $\text{ENI}_{95} = 20$, a small fraction of the total sequence length of 120.

Whitney and Leib (2018) suggested a square root relationship between the ENI and the total number of stimuli, although the studies cited for ENI values did not share a

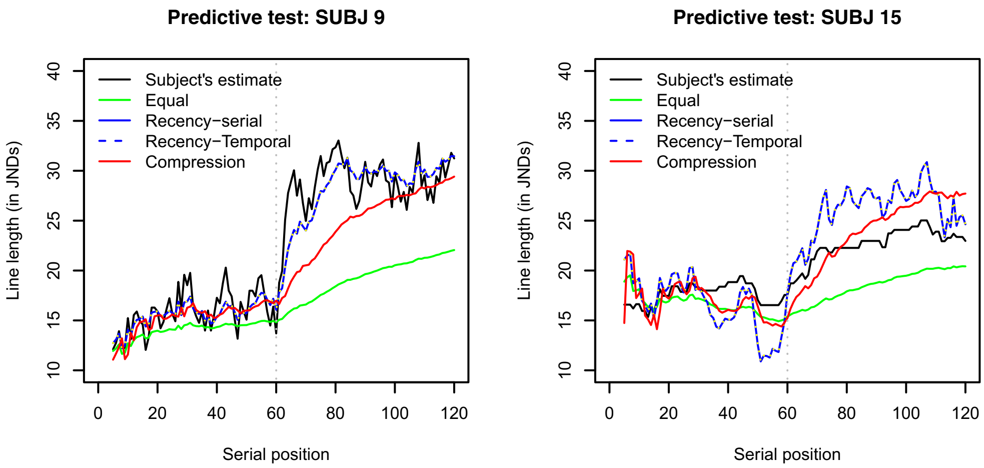

Fig. 5 Model predictions from competing models. The averaged best-fitting parameters from ten randomly sampled subjects to predict data from a new subject. For training, best-fitting parameters from all mean shift conditions were averaged together. For testing, the sequence at test was randomly sampled from a untrained subject. Both Recency and Compression model outperformed the Equal model. The Recency model showed better prediction for subjects with more recency-weighted pattern (e.g., Subject 9). The predictions from the *Recency-s* and *Recency-t* models are overlapping with each other, due to highly correlated serial and temporal positions. Note that the plots are from one random instance from the predictive test. Summarized results over repetitions are reported in Table 2

unified definition of ENI. Our data seem to suggest a ceiling of the ENI regardless of the total number of items in the sequence. Looking back at the five sequential averaging studies cited in Fig. 1, we see that the total numbers of stimuli in those studies are small (less than 20), and the estimated ENIs (max at 10) are far below the ceiling suggested in the current results (around 20). This suggests the possibility of a two-stage relationship between the ENI and total

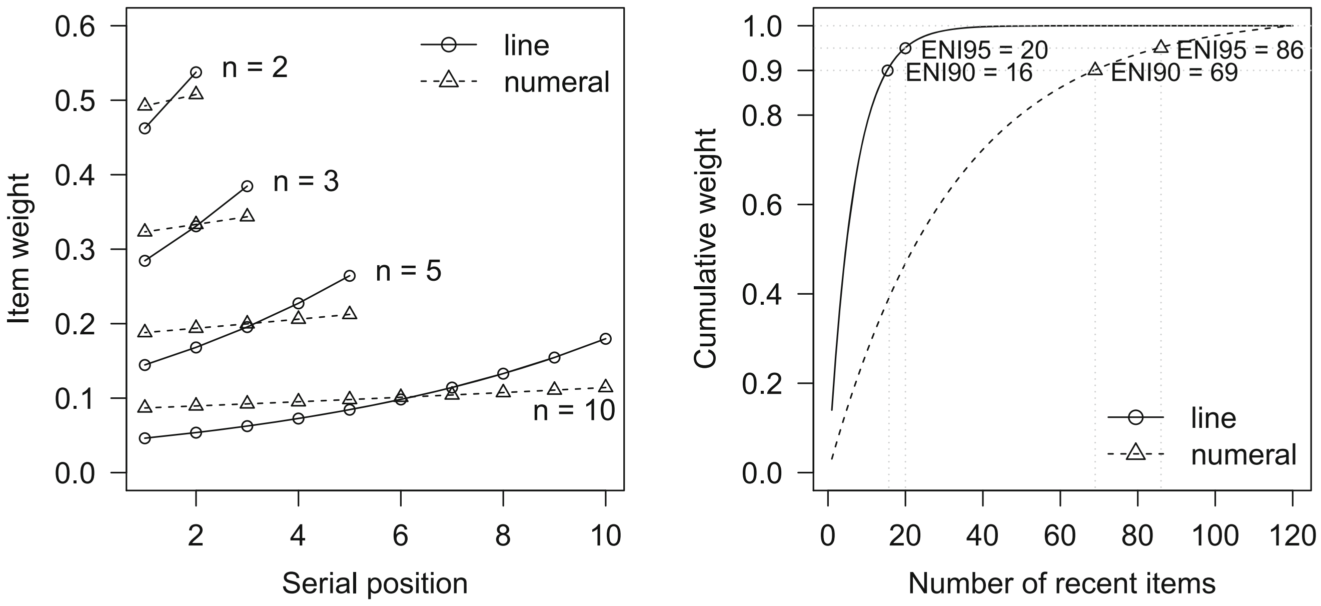

Fig. 6 *Left panel* The recency patterns in the weight distributions from the group-averaged best-fitting r parameter (mean $r = 0.86$, from the line conditions of Experiments 1 and 2). Each set of connected dots is one weight distribution. The number of lines in the sequence is denoted by each weight distribution. As the number of items increases, the weights are distributed over more items, but the recency pattern exists for sequences of all lengths. For visual convenience, only four sequence lengths (2, 3, 5, and 10) are shown. *Right panel*. The degree of recency differs in the line and numeral conditions, where subjects assigned weights to more items in the numeral condition than the line condition. The *black solid line* represents the line condition and the *black dashed line* represents the numeral condition. *Gray dotted lines* are 90% and 95% thresholds, and the projections of their crossings with the cumulative weights on the horizontal axis are the ENI for the respective level. For the line condition, $\text{ENI}_{90} = 16$ and $\text{ENI}_{95} = 20$. For the numeral condition, $\text{ENI}_{90} = 69$ and $\text{ENI}_{95} = 86$

number of stimuli, that the two may have a certain functional relationship (e.g., the square root relation), however, the maximum number of ENI may also be capped at some upper limit.

The mean-shift design

Previously in the introduction, we mentioned two concerns over the mean-shift design: (i) the distinctive shift in the mean may encourage recency in item integration, (ii) reporting the mean on every trial may encourage subjects to compress prior items into a single representation rather than averaging prior items. Modeling results ruled out these two concerns in Experiment 1.

Firstly, the best-fitting parameters from the three mean-shift conditions were not significantly different from each other, contrary to the idea that the distinctiveness of the mean-shift encourages recency in item integration. Secondly, the better model performance of the *Recency* model over the *Compression* model suggests that by requiring subjects to report the mean on every trial, the preceding stimuli are unlikely to be compressed into a single value, as suggested in the *Compression* model.

In sum, our results challenge prior treatments of prototype models as equally weighted averages over all items (Nosofsky, 1987; Smith & Minda, 2000), and call into question the conclusions drawn on the basis of those assumptions.

Experiment 2

In Experiment 1, the initial value of the adjustable probe on each trial (except the first trial) was set to the subject's response from its immediate previous trial. This may have encouraged subjects to base their estimations on their previous responses. This strategy is reasonable because the responses are naturally auto-correlated in this task, but we wonder whether the starting values were responsible for the previous results that suggest recency weighting. Existing studies using the method of adjustment have used either random starting values (Haberman & Whitney, 2010) or fixed values that are outside the range of the regular stimuli (Huang & Sekuler, 2010). In Experiment 2, we changed the starting values of the adjustable probes to a fixed small value.

Additionally, we aim to determine whether the findings in Experiment 1 are limited to line length. To this end, we include a condition in which subjects are asked to keep track of the running averages of sequentially displayed numerals. Unlike simple visual stimuli like vertical lines, numerals are symbols that carry conceptual information that is directly related to subjects' estimation responses. So, Experiment 2 afforded a direct comparison of response patterns from the line and numeral tasks.

Methods

Subjects Ten new subjects participated in the study, five each for the line and numeral conditions (Seven female, mean age = 18.7 years, SD = 1.06 years). All had normal or corrected-to-normal vision. All procedures were approved by the IRB of University of South Florida.

Procedure, design, and stimuli For the line condition, all specifications were the same as the "Large Shift" condition in Experiment 1, except that the starting value of the adjustable probe was set to a fixed small value on each trial. The value is 4 pixels (0.11 degrees in visual angle), the same size as the width of the line, so the probe looked like a dot on the screen. This starting length was chosen to ensure the probe contained as little line length information as possible. The dot probe is far outside the range of regular stimulus lines, so it is unlikely to confound the estimations.

For the numeral condition, the stimuli used were one or two-digit integer numerals, and the task was to estimate the average value of all the preceding numerals. No QUEST procedure was applied because we assumed that any difference between integer numerals is equally detectable, so data were recorded in the original numerical scale. The starting values of the probes were set to zero. The pre/post-shift means were 20/50 and the standard deviation was 5. The smallest adjustment step is one.

Results and discussion

The results, summarized in Fig. 7, show that data from the line condition generally resemble those of Experiment 1. However, subjects' estimations were closer to the equally weighted means in the number condition. This suggests that the weighting schemes for sequential averaging differ for different types of stimuli. One possible explanation for this particular difference between lines and numerals is that participants have considerable experience in estimating numerical quantities in everyday life and are subject to extensive testing of the accuracy of mental arithmetic.

To quantify the differences in the response patterns, we fit the data with the *Recency-s* model from Experiment 1. For the line condition, the mean best-fitting r was 0.86, the same value found in Experiment 1. This suggests that results in Experiment 1 were not due to the influence of the probe line's initial length.

For the numeral condition, the mean of the best-fitting r was 0.97. One of five subjects in the number condition showed recency like we found in the line conditions in

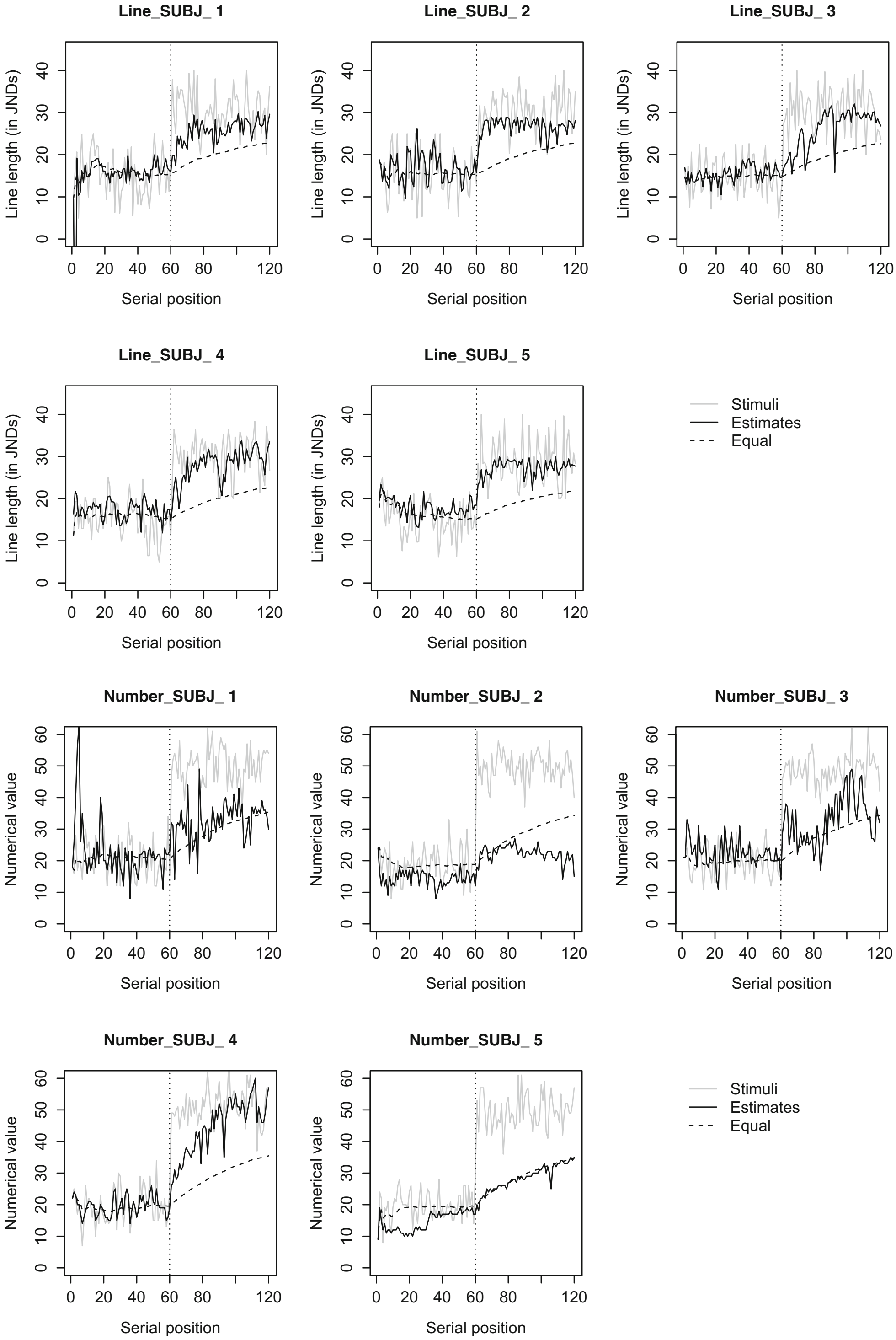

Fig. 7 Data from Experiment 2. The *black solid lines* represent subject's estimates. The *black dashed lines* represent predictions from the *Equal* model. The *gray solid lines* represent the stimuli. *Upper five panels: Line condition*. Compared with Experiment 1, our results suggest that the starting value of the probe did not affect the response pattern in the line condition. Subjects' estimates do not follow the equally weighted moving averages for any subject. *Lower five panels: Number condition*. Subjects' mean estimates for numbers were closer to the equally weighted means. This pattern is also confirmed by the modeling results. Individual differences were observed, e.g., Subject 4 showed estimates similar to those from the line condition, but the other four subjects' estimates were much closer to the equally weighted moving averages. A significant difference was found between the best-fitting rate parameters for the line and numeral conditions

Experiments 1 and 2. For the other four subjects, the best-fitting rate parameters were close to 1, suggesting equal or near-equal weighting schemes. A Welch two-sample *t*-test revealed a significant difference in the best-fitting *r* parameter between the line and numeral condition, *t* (23.86) = -3.38, *p* = 0.002.

We also found the ENI_{90} = 69 and ENI_{95} = 86 for the numeral condition, which are considerably higher than those of the line condition (ENI_{90} = 16, ENI_{95} = 20). This indicates subjects utilized many more items when estimating the running means of numerals. Despite individual differences, with some subjects showing recency weighting and others not, the overall prioritization of recent items is much less evident than in the line conditions.

To summarize, by comparing the line data from Experiments 1 and 2, we ruled out the influence from the starting values of the probes in the line condition. However, the type of stimulus, line vs. numeral, did significantly influence subjects' response patterns. Subjects differentially weight more recent items in both tasks but incorporate many more items when tracking the running average of numerals.

Experiment 3

Experiment 3 used a mean shift paradigm similar to that of Experiment 1, but the direction of the shift was manipulated (upward vs. downward shifts), to test whether the findings from previous experiments were specific to their shared direction of the shift.

Methods

Subjects Twelve new subjects participated in Experiment 3. Subject 9 was excluded due to incomplete data, so eleven subjects' data were analyzed (six female, mean age = 19.17 years, SD = 0.94 years). All had normal or corrected-to-normal vision. All procedures were approved by the IRB of University of South Florida.

Procedure, design, and stimuli Subjects were presented with a sequence of gray vertical lines, one at a time. Each line was displayed in the center of the screen for one second. After each gray line, subjects were asked to estimate the average length of all the gray lines they had seen to that point. Subjects showed their estimations by adjusting a white probe, which appeared on the screen after a one-second blank screen after each gray line, by moving the computer mouse. Stimuli and probes were presented in a different color to reduce any confounding influence of the probe length on estimates of prior stimuli. When subjects completed an adjustment, they pressed the space bar on the keyboard to proceed, triggering a 500-ms blank inter-trial buffer, which was followed by the next trial.

The lengths of the stimuli were drawn from Gaussian distributions. The mean lengths of the two halves of the sequence were 200 and 400 pixels (5.62 to 11.21 degrees in v.a.) The SD of the line lengths was 100 pixels (2.81 degrees in v.a.). Line lengths drawn outside two SDs from the mean were resampled. The number of lines in a full sequence was 60, where the mean-shift occurred on the 31st trial. The adjustable line was always displayed as a white dot, as specified in Experiment 2. Each subject went through six sequences, interleaved with upward and downward mean-shifts. The direction of the mean-shift of the first sequence was randomly assigned across subjects.

Results and discussion

Using the same *Recency-s* model in Experiment 1, we obtained the best-fitting *r* parameters for all the sequences. We averaged the parameters within subjects, so each subject had two mean best-fitting *r* parameters, one for upshift and one for downshift. The differences between the mean upshift and downshift *r* parameters were tested against zero, *t*(10) = -1.87, *p* = 0.09. The upward sequences (mean = 0.83) show a smaller mean best-fitting *r* parameter, suggesting a higher degree of recency than the downward sequences (mean = 0.89), but the difference was not significant. Both r parameters are significantly different from 1, indicating recency-weighting is both upshift and downshift sequences

To further compare the difference between upshift and downshift conditions, we calculated the signed post-shift estimation error for upshift and downshift sequences for each subject. The estimation error was linearly scaled to [-1, 1] for each sequence from each subject, using the signed difference between estimates and equally weighted averages, divided by the maximum value of the absolute difference.

The signed estimation error allows us to see how subjects estimate in different directions of mean-shifts. Under the *Recency* model, subjects would show overestimation in the up-shift conditions and underestimation in the down-shift conditions in the post-shift trials, because their estimates relied more on the recent stimuli which deviate with the pre-shift means. Most subjects' data confirmed this prediction, but individual differences were also observed (Fig. 8).

We calculated the mean estimation error for upward and downward blocks for each subject. If there was no magnitude difference between the upward and downward blocks, the sum of mean upward and downward errors should be around zero for each subject, since upward block generally showed overestimation and downward showed underestimation. We conducted a one-sample *t*-test on the summed mean error from upward and downward blocks

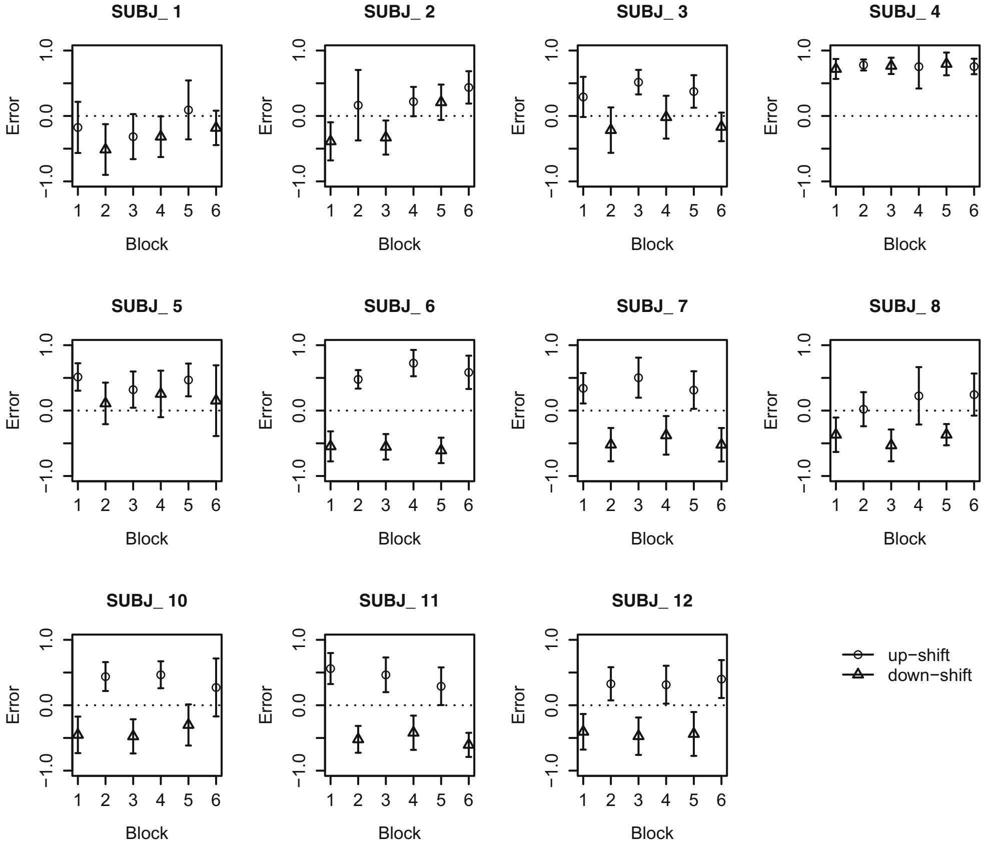

Fig. 8 Mean and SD of normalized post-shift estimation error for all subjects and all sequences. *Horizontal axes* are the block order (1 is the oldest). Subject 9 was excluded due to incomplete data. For most subjects, the estimation error is positive (overestimation) for up-shift blocks and negative (underestimation) for down-shift blocks. Individual differences (e.g., Subject 1, 4, and 5) were also observed. There is no significant difference in the magnitude of estimation error between upward and downward blocks

against zero. Results showed no significant difference, t (10) = 0.85, p = 0.41. To conclude, the direction of the mean shift did not influence subjects' estimations of the running means.

Experiment 4

In the previous experiments, subjects reported the running means after every stimulus. Does this requirement affect the item integration mechanism when subjects report the means? In Experiment 4, subjects only report one mean value at the end of each of the sequences. We estimated the weight distributions with multiple regression and compared the estimates with predictions from the *Recency* model used in previous experiments.

Methods

Subjects Twenty new subjects participated in Experiment 4 (Five male, mean age = 19.85 years, SD = 1.04 years). All had normal or corrected-to-normal vision. All procedures were approved by the IRB of University of South Florida.

Procedure Experiment 4 had two conditions varying in the number of lines in the sequence (five-line and ten-line). The twenty subjects were equally divided between the two conditions. There were 140 and 80 trials per subject in the five-line and ten-line conditions, respectively.

In the experiment, a series of vertical lines was displayed sequentially on the center of the screen. Each line was presented for 0.5 s and followed by a 1-s blank screen before the next line. After the blank screen following the last line in the sequence, an adjustable "dot" appeared in the center of the screen.

The task was to estimate the average line length of the sequence of lines, by adjusting the probe dot using the mouse. When subjects moved the mouse up or down, the probe dot expanded vertically in both directions at the same rate into a line, so the line was always displayed in the center of the screen. When satisfied with their estimates, subjects pressed the space bar on the keyboard to submit. No time limit was set for the adjustment phase. A trial ended with the subject's submission, followed by a 1-s blank screen before the next trial. As in Experiment 2, the design choice of making the adjustable probe a “dot” aims to minimize the influence of the initial probe values on subjects’ adjustments.

Stimuli The sequence means were randomly sampled from a uniform distribution between 200 and 400 pixels. The line lengths within the sequences were sampled from Gaussian distributions centered at the sequence means with a fixed SD of 100 pixels. All random lengths were taken absolute values to ensure that there were no negative values.

Results and discussion

Weight distribution estimation Unlike Experiments 1–3, where subjects reported the running means after each stimulus, in Experiment 4 subject reported one mean for each sequence. This design makes it difficult to estimate the weight distribution for each sequence. To compensate, we added a large number of sequences for each subject, and use a multiple regression method to estimate the weight distribution (Juni et al., 2012; Hubert-Wallander & Boynton, 2015). Subjects’ responses were regressed onto the stimuli at different serial positions with no interaction terms. The group-averaged regression coefficients were then standardized to sum to one as an estimate of the weight distribution.

Figure 9 shows that the estimated weight distributions of the five-line (solid triangle) and ten-line (solid circle) conditions in Experiment 4 match the predictions from the Recency model (open circle), using the averaged best-fitting r parameter (r = 0.86) from the line conditions in Experiments 1 and 2. Individual weight distributions can be found in the Supplementary Materials. Subject 6 from the five-line condition was excluded from the group average because data suggested that this subject was reproducing the last item, rather than estimating the average of the five items, due to the overly high coefficient to the most recent item (greater than 0.96) and near-zero coefficients for other serial positions.

From Fig. 9, we learn that subjects also show recency weighting in tasks where they report one mean at the end of the sequentially presented stimuli. More importantly, the estimated weighting schemes are well predicted by the best-fitting parameter in the *Recency* model from previous experiments, suggesting that the same item integration mechanism may be in use across the sequential averaging experiments, whether or not subjects report the means after every stimulus.

The item integration mechanism could still be influenced by task differences (report running means vs. report one mean) in a manner that the current analysis is incapable of revealing. Subjects could be reinforced by their frequent reports in the running mean task. Future studies may investigate how keeping track of the running means affects the performance of sequential averaging.

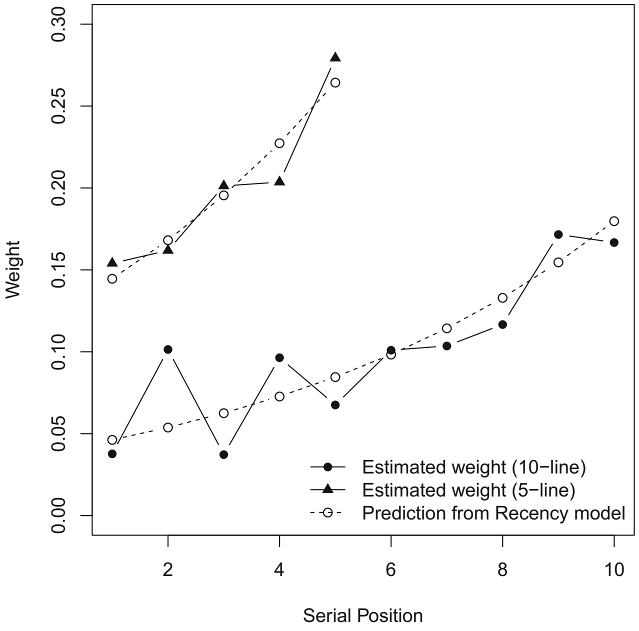

Fig. 9 The estimated weight distributions of the five-line (*solid triangle*) and ten-line (*solid circle*) conditions in Experiment 4 match the predictions from the *Recency* model (*open circle*), using the averaged best-fitting r parameter (r = 0.86) from the line conditions in Experiments 1 and 2

General discussion

Feature-matching models of memory and perception, such as the exemplar-based random walk model (EBRW, Nosofsky & Palmeri 1997; Nosofsky et al. 2011), assume that decisions about probes entail a parallel match to stored memory representations of individual items. Importantly, such models exclude any role for memory representations of central tendency. Indeed, the generalized context model, EBRW’s core, has been used in competitive fits in order to rule out models assuming central tendency representations (prototypes) factor into such decisions (e.g., Nosofsky & Zaki 2002; Smith & Minda 2000). Yet these efforts have typically assumed the prototype to be a simple arithmetic mean over the prior stimulus features. If this assumption is wrong, then a reassessment of prototype theory may be necessary and prior conclusions rejecting prototype theory may need to be revised.

Our results suggest that a limited number of recent stimuli contribute to subjects’ estimates of the mean value of sequentially presented stimuli. Specifically, fewer than ten stimuli account for over half of the cumulative weight subjects use, and fewer than 20 stimuli account for nearly all of the cumulative weight, for averaging the length of vertical lines. The classical assumption of prototypes as arithmetic means in studies of perceptual categorization provided a poorer account of the data for visual features like line length than a recency-weighted weighting scheme. On one hand, our data call for a reassessment of prototype models used in perceptual categorization research. On the other hand, due to the differences in tasks and stimuli used in our experiments from many categorization tasks, further research is necessary to make strong claims.

Recency-prioritization in sequential averaging

In a sequential averaging task, ideal observers are assumed to keep equal fidelity and assign equal weight to all the previous items, however, neither assumption is likely to hold for human subjects. We speculate that the recency-prioritization seen in sequential averaging may in part reflect the temporal decay of memory strength (Wilken & Ma, 2004). Specifically, when the average is computed over a set of items presented up to a particular point in time, the strength of the individual item representations and/or their association to the current temporal context is likely to determine the degree to which they influence the computation they are a part of (Howard & Kahana, 2002).

In addition to decreasing weight over serial positions (as in *Recency-s* model), a recency-prioritized weighting scheme could be realized by alternative constructs, e.g., applying a relatively homogeneous weighting scheme to a short list of recent items. It is also interesting to know how subjects adjust their item integration mechanism upon explicit instructions. For example, when asked to average over the most recent five or ten items in a sequence, will subjects accurately limit their item inclusion according to the instruction? Results of such inquires in the context of active item integration may shed light on the discussion of the nature of capacity limits for visual working memory (Luck & Vogel, 2013; Ma et al., 2014).

In addition to the fidelity of internal item representations, external noise and feedback may also influence the weighting scheme. Juni et al. (2012) manipulated the noise level of different serial positions in a sequential averaging study and found that the weights of the serial positions are negatively correlated with their noise level. In the current study, the external noise level (stimulus SD within each condition) was kept constant, so this external factor cannot play a major role in our results. That suggests recency-prioritization may be mainly attributed to the temporal decay of internal memory representations. Corrective feedback may induce more equally distributed weights over sequential items (Juni et al., 2010). Since feedback was not incorporated in the current study, a future study may further investigate feedback's role in sequential averaging.

In the current study, the temporal positions and serial positions are highly correlated with each other, due to fixed between-stimulus intervals and low response time variability, so the modeling analysis cannot distinguish the influences from the serial or temporal factors. Future studies may incorporate designs that disentangle the contributions of serial and temporal positions to better understand the source of the recency weighting scheme.

Shifting the distribution mean

One unique contribution of the current study is the use of the mean shift to evaluate the number of trials used in sequential averaging. For the pre-shift trials, all stimuli were from the same distribution, producing relatively low estimation error. This low error on the first distribution is consistent with the ensemble representation literature which shows that judgments of the statistics of a single distribution of stimuli are accurate (Albrecht & Scholl, 2010; Haberman & Whitney, 2009). However, after the mean shift, errors started to reflect a mixture of influences from the current and prior distribution. Such an influence of prior items was likely to escape detection by prior studies of ensemble representation. In those studies, stimuli were usually drawn from a single distribution (but see Morgan et al. 2000), from which a small sample was likely to produce close approximation to the mean, so long as the variance of the stimulus distribution was not too great (referred to as the "subsampling problem" in Simons & Myczek 2008).

Subsampling

The summary statistical effects could be produced through a subsampling strategy: subjects could merely attend to two or three items in a given sequence or ensemble, for which an estimation of the average would likely be close to the true average of the ensemble at least in cases with low variance and normally distributed items (Simons & Myczek, 2008; Myczek & Simons, 2008). Although there have since been several demonstrations of averaging effects under conditions that attempt to rule out subsampling (Corbett & Oriet, 2011), the simulation-based argument of Myczek and Simons suggests that under many circumstances the two approaches may be difficult to discriminate, as in the pre-shift phase in the current study.

In the current study, if subjects subsampled "properly", i.e., sampled a subset of items from the whole sequence regardless of serial positions, their estimation error would have been much lower than what was observed. Our data suggest that the mean shift caused subjects to quickly change their "sampling pool" to the new distribution, if in fact the subsampling strategy was used.

Sequential vs. simultaneous averaging

One key hypothesis about ensemble representation is that it involves mechanisms that are qualitatively different from those used to process individual objects (Alvarez, 2011; Whitney & Leib, 2018). Evidence supporting such qualitative differences are mostly from simultaneous averaging studies in which observers performed better at tasks targeting ensemble features than tasks targeting individual stimulus features (e.g., Ariely 2001).

Subjects are assumed to employ parallel processing in simultaneous averaging tasks (Alvarez, 2011), however, in sequential averaging studies, individual objects are not presented at the same time, thus they cannot be processed in parallel (barring, say, some form of perseveration or resonance in short-term memory). This may explain why studies of sequential averaging have received less attention in the ensemble representation literature than studies of simultaneous averaging (for review, see Dubé & Sekuler 2015). Nonetheless, we know of no definitive evidence suggesting that the qualitative distinguishing factor of ensemble representation has to be, solely, parallel processing. Qualitative differences might include other aspects such as the mechanisms of computation, storage, or retrieval.

What directions do our results suggest for simultaneous averaging studies? One possibility is to use the recency weighting scheme to test the hypothesis that regardless of whether the stimuli are presented simultaneously or sequentially, the averaging process can be decomposed into processing each item serially and integrating their representations into an ensemble representation. If subjects' responses in simultaneous averaging tasks could be reliably and accurately predicted using the recency weighting scheme from sequential averaging and external measures such as eye-tracking to determine the serial order of item processing, such results would challenge the parallel processing hypothesis in ensemble representation.

Generalizations and limitations

Similar recency-weighted evaluation may occur in everyday conditions. For example, memory-based evaluation of events can be distorted by the order of the experiences (Redelmeier et al., 2003). We wonder whether this is a generalization of the recency-weighted summary statistical representation demonstrated in our study.

Stimulus serial position is a key determinant of memory performance (Baddeley, 2003) and thus influences item contributions in the sequential averaging task. We incorporated this knowledge into the *Recency* model, however, the model is far from inclusive of all relevant factors. The current model assumes a static weight distribution governing item integration, however, this process may well be dynamic, having variable resource distribution in item integration, depending on multiple trial-wise features, such as stimulus saliency (Brown et al., 2007; VanRullen, 2003; Itti & Koch, 2000), perceptual sequential dependency (Fischer & Whitney, 2014; Fornaciai & Park, 2018) or decisional sequential dependency (Alais et al., 2017; Fritsche et al., 2017). Future modeling work may benefit from incorporating separate sources of noise, to parse out different forms of errors that contribute to subjects' judgments of summary statistical properties.

The data and materials for the experiments reported here is available upon request at ketong@mail.usf.edu. None of the experiments was preregistered.

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. *Journal of Neuroscience*, 4601–15.

Albrecht, A. R., & Scholl, B. J. (2010). Perceptually averaging in a continuous visual world: Extracting statistical summary representations over time. *Psychological Science*, *21*(4), 560–567.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences 15.3*, 122–131.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*(2), 157–162.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539–576.

Byrd, R., et al. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208.

Chetverikov, A., Campana, G., & Kristjánsson, Á. (2016). Building ensemble representations: How the shape of preceding distractor distributions affects visual search. *Cognition*, *153*, 196–210.

Corbett, J. E., & Oriet, C. (2011). The whole is indeed more than the sum of its parts: Perceptual averaging in the absence of individual item representation. *Acta Psychologica 138*(2), 289–301.

Dubé, C., & Sekuler, R. (2015). Obligatory and adaptive averaging in visual short-term memory. *Journal of Vision 15*(4):13, 1–13.

Dubé, C., et al. (2014). Similarity-based distortion of visual short-term memory is due to perceptual averaging. *Vision Research 96*, 8–16.

Estes, W. K. (1994). *Classification and cognition*. London: Oxford University Press.

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience 17*(5), 738–743.

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

Fornaciai, M., & Park, J. (2018). Attractive serial dependence in the absence of an explicit task. *Psychological Science 29*(3), 437–446.

Fritsche, M., Mostert, P., & de Lange, F. P. (2017). Opposite effects of recent history on perception and decision. *Current Biology 27*(4), 590–595.

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

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception and Psychophysics 72*(7), 1825–1838.

Hintzman, D. L., & Ludlam, G. (1980). Differential forgetting of prototypes and old in stances: Simulation by an exemplar-based classification model. *Memory and Cognition 8*(4), 378–382.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology 46*(3), 269–299.

Huang, J., & Sekuler, R. (2010). Distortions in recall from visual memory: Two classes of attractors at work. *Journal of Vision 10*(2), 24–24.

Hubert-Wallander, B., & Boynton, G. M. (2015). Not all summary statistics are made equal: Evidence from extracting summaries across time. *Journal of Vision 15*(4), 5–5.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research 40*(10), 1489–1506.

Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (2010). Integration of visual information across time. *Journal of Vision 10*(7), 1213–1213.

Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (2012). Effective integration of serially presented stochastic cues. *Journal of Vision 12*(8), 12–12.

Leib, A. Y., Fischer, J., Lui, Y., Qui, S., Robertson, L., & Whitney, D. (2014). Ensemble crowd perception: a viewpoint-invariant mechanism to represent average crowd identity. *Journal of Vision*, *14*(8), 26–26.

Leib, A. Y., Kosovicheva, A., & Whitney, D. (2016). Fast ensemble representations for abstract visual impressions. *Nature Communications 7*(13186).

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

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience 17*(3), 347–356.

Michael, Elizabeth, De Gardelle, V., & Summerfield, C. (2014). Priming by the variability of visual information. *Proceedings of the national academy of sciences*, 201308674.

Morgan, M. J., Watamaniuk, S. N. J., & McKee, S. P. (2000). The use of an implicit standard for measuring discrimination thresholds. *Vision Research 40*(17), 2341–2349.

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception and Psychophysics 70*(5), 772–788.

Norman, L. J., Heywood, C. A., & Kentridge, R. W. (2015). Direct encoding of orientation variance in the visual system. *Journal of Vision 15*(4), 3–3.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition 13*(1), 87–108.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review 104*(2), 266–300.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology. Learning, Memory, and Cognition 28*(5), 924–940.

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review 118*(2), 280–315.

Oriet, C., & Hozempa, K. (2016). Incidental statistical summary representation over time. *Journal of Vision 16*(3), 3.

Parducci, A. (1956). Direction of shift in the judgment of single stimuli. *Journal of Experimental Psychology 51*(3), 169–178.

Piazza, E. A., Sweeny, T. D., Wessel, D., Silver, M. A., & Whitney, D. (2013). Humans use summary statistics to perceive auditory sequences. *Psychological Science*, *24*(8), 1389–1397.

Redelmeier, D. A., Katz, J., & Kahneman, D. (2003). Memories of colonoscopy: a randomized trial. *Pain 104*(1), 187–194.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM–retrieving effectively from memory. *Psychonomic Bulletin and Review 4*(2), 145–166.

Shmueli, G. (2010). To explain or to predict? *Statistical science 25*(3), 289–310.

Simons, D. J., & Myczek, K. (2008). Average size perception and the allure of a new mechanism. *Perception and Psychophysics 70*(7), 1335–1336.

Smith, D. J., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition 26*(1), 3–27.

Teghtsoonian, R. (1971). On the 1 exponents in Stevens' law and the constant in Ekman's law. *Psychological Review 78*(1), 71–80.

VanRullen, R. (2003). Visual saliency and spike timing in the ventral visual pathway. *Journal of Physiology-Paris. Neurogeometry and Visual Perception 97*(2), 365–377.

Watson, A. B., & Pelli, D. G. (1983). Quest: A Bayesian adaptive psychometric method. *Perception and Psychophysics 33*(2), 113–120.

Weiss, D. J., & Anderson, N. H. (1969). Subjective averaging of length with serial presentation. *Journal of Experimental Psychology 82*(1), 52–63.

Whitney, D., & Leib, A. Y. (2018). Ensemble perception. *Annual Review of Psychology 69*(1), 105–129.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision 4*(12), 11–11.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science 12*(6), 1100–1122.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.