Attention, Perception, & Psychophysics (2022) 84:188–204
https://doi.org/10.3758/s13414-021-02365-3

Response bias in numerosity perception at early judgments and systematic underestimation

Aslı Kılıç[1] · Aslı Bahar İnan[2]

Accepted: 26 July 2021 / Published online: 13 September 2021
© The Psychonomic Society, Inc. 2021

Abstract

Mental number representation relies on mapping numerosity based on nonsymbolic stimuli to symbolic magnitudes. It is known that mental number representation builds on a logarithmic scale, and thus numerosity decisions result in underestimation. In the current study, we investigated the temporal dynamics of numerosity perception in four experiments by employing the response-deadline SAT procedure. We presented random number of dots and required participants to make a numerosity judgment by comparing the perceived number of dots to 50. Using temporal dynamics in numerosity perception allowed us to observe a response bias at early decisions and a systematic underestimation at late decisions. In all three experiments, providing feedback diminished the magnitude of underestimation, whereas in Experiment 3 the absence of feedback resulted in greater underestimation errors. These results were in accordance with the findings that suggested feedback is necessary for the calibration of the mental number representation.

Keywords Numerosity perception · Mental number line · Speed–accuracy trade-off · Response deadline procedure

How we perceive numerosity is an intriguing subject of psychophysics with its relevance to understanding the mental number representation, magnitude perception, and mathematical concept formation (Crollen et al., 2011; Crollen et al., 2013; Dehaene et al., 2008; DeWind et al., 2019; Krueger, 1972; Mundy & Gilmore, 2009). In numerosity perception tasks, nonsymbolic stimuli, such as a collection of dots, are used as stimuli, and participants are asked to respond with a symbolic output such as Arabic numerals (Crollen et al., 2011; Izard & Dehaene, 2008; Krueger, 1972; Reinert et al., 2019). Approximate Number System (ANS) is proposed to be the mechanism that is responsible for processing nonsymbolic stimuli that are utilized in numerosity perception tasks, producing a percept of numbers without requiring to count or to calculate, but rather to estimate the numbers indirectly (Anobile et al., 2016; Dehaene, 2003; Dietrich et al., 2015; Guillaume & Gevers, 2016; Guillaume & Van Rinsveld, 2018; Mejias & Schiltz, 2013; Van den Berg et al., 2017).

In addition to findings for human adults, the ANS is also found in infants, and other species (Anobile et al., 2016; Burr et al., 2018; Dehaene, 2003; Fornaciai et al., 2016; Nieder, 2016; Piazza, 2010; Whalen et al., 1999; Xu & Spelke, 2000), which is considered as an innate mechanism (Dehaene, 2011), having evolutionary advantages. Detecting an approximate number for sets of objects may have survival value such as detecting potential predators instantaneously (Burr et al., 2018; Nieder, 2016; Norris & Castronovo, 2016; Piazza, 2010). Studies on ANS and numerosity perception suggest that humans use a mental representation of a number line that is logarithmically scaled (Castronovo & Seron, 2007; Cordes et al., 2001; Crollen et al., 2011; Crollen & Seron, 2012; Dehaene, 1992, 2011; Dietrich et al., 2015; Izard & Dehaene, 2008; Reinert et al., 2019).

In numerosity perception tasks, there is a mapping from nonsymbolic representation to symbolic representation, which directly links the logarithmically scaled mental number line to the actual linear number line (Dehaene, 2011; Piazza et al., 2007; Verguts & Fias, 2004). Consequently, the mapping of the logarithmic mental number line to the linear number line results in systematic underestimation (Crollen et al., 2011; Krueger, 1972, 1982). Since the mental number line is logarithmically compressed, the subjective representation of nonsymbolic stimuli will be mapped on the linear number line with a smaller numerical value. The underestimation of the

✉ Aslı Kılıç
askilic@metu.edu.tr

[1] Department of Psychology, Middle East Technical University, Dumplupinar Bulv. No:1, Çankaya, 06800 Ankara, Turkey

[2] Department of Psychology, Çankaya University, Ankara, Turkey

subjective magnitude compared with the objective magnitude is a result of this compression in the mental number line. However, Izard and Dehaene (2008) explains how scalar variability was still predicted by their model as the increase in numerosity results in an increase in the standard deviation of responses.

The underestimation of the subjective magnitude is a general finding when feedback was not provided to the participants (Bevan & Turner, 1964; Crollen et al., 2011; Crollen et al., 2013; Indow & Ida, 1977; Izard & Dehaene, 2008; Krueger, 1982; Reinert et al., 2019). Izard and Dehaene (2008) investigated the effects of feedback on calibrating the mental number representation by presenting inducers to the participants. In a numerosity perception experiment, they presented feedback, which they referred to as inducers, prior to the trials, and manipulated three types of inducers: overestimated, underestimated, and exact. For all the types, participants were told that the inducers contained 30 dots, while only for the exact inducers 30 dots were presented, but for the overestimated inducers 25 dots and for the underestimated inducers 39 dots were displayed. The results showed that the participants readapted their responses according to the inducers, which was the feedback that they received prior to the task. Izard and Dehaene (2008) proposed a model for numerosity perception and developed their model to account for the findings that calibration of the mental number line benefits from feedback. In their model, the width of the distribution of activation on the logarithmic mental number line corresponds to sensitivity in signal detection theory, and represents the amount of noise in the numerosity representation. The activation on the logarithmic mental number line takes place during encoding of numerosity. They postulate that the representation on the number line is then transformed into a verbal numerical response corresponding to a segment which is divided according to a list of criteria, defined as the response grid. When no feedback is provided, numerosity estimations are generated based on the spontaneous response grid. However, in the presence of feedback, an affine transformation is applied to the response grid, resulting in calibration. Therefore, the response bias in their model corresponds to the response grid, defining the position of the response criteria. The results of their study suggested that the calibration is a global process due to the response selection stage (response bias) and is not due to encoding of numerosity (sensitivity). Similar findings of calibration, or in other words a reduction of underestimation in numerosity due to usage of feedback, also comes from studies that provide feedback during the task, not being presented in the form of inducer but presented after every trial (Krueger, 1984; Price et al., 2014).

Current study

In the current study, we investigated the role of feedback on numerosity perception. After presenting randomly displayed dots on the screen, we collected responses from a two-choice decision task. In the decision task, participants either selected the option that indicated the number of dots exceeded one criterion value (50 in the current experiment), or the other option, indicating that the number of dots stayed below that criterion value. Immediately following the response, we provided the actual number of dots that were displayed on the screen. Our feedback was intended to investigate how the mental number representation was calibrated when participants received the actual number of nonsymbolic representations, and thus, if a correct mapping occurred between symbolic and nonsymbolic representations.

In addition to investigating how calibration of mental number representation benefits from presenting actual numbers tested, we examined the temporal course of numerosity perception. In traditional studies of numerosity perception, participants are given either a production task (Crollen et al., 2011) or a choice decision task (e.g., Ratcliff, 2006) that requires a selection of one response among two alternatives. The responses in two-choice decision tasks are subject to speed–accuracy trade-offs, such that fast responses are likely to be incorrect while the accurate responses are likely to be slower. The responses obtained from standard choice tasks might be biased towards either one. To obtain unbiased measures of speed and accuracy, we applied the response-deadline speed–accuracy trade-off (SAT) procedure in a two-choice numerosity decision task. As a result, we had the opportunity to observe the time course of numerosity perception, which is explained in more detail in the following section.

Finally, controlling for the speed of processing in a two-choice numerosity decision task allowed us to measure unbiased estimates of numerosity. Specifically, we can observe whether there is a systematic underestimation when the speed of responding is controlled and whether their estimates become more accurate when people are given more time to process the dot patches. To answer these questions, we employed the response deadline SAT procedure to obtain conjoint measures of accuracy (sensitivity) and response time in numerosity perception.

Response deadline speed–accuracy trade-off procedure

The response-deadline SAT procedure provides conjoint measures of speed and accuracy (Kılıç & Öztekin, 2014; Ratcliff, 2006; Ratcliff & McKoon, 2018). Due to being subject to speed–accuracy trade-offs, traditional response-time measures may provide biased measures of speed and accuracy. On the other hand, by providing the full time course of processing, SAT procedure yields independent assessment of accuracy and speed of processing (Reed, 1973).

In SAT, participants are cued to respond with a response signal presented at one of several time points, typically ranging from 60 to 3,000 ms after the display onset. The random assignment of lag between stimulus onset and the response signal to test trials provides a control for the speed of processing. In a practice session, participants are trained to give a response within 500 ms after the response cue.

SAT functions can describe changes in accuracy as a function of total processing time, the total time that passes from stimulus onset to the response after the response cue is presented. The SAT functions typically start with a period of performance where the two choices are selected randomly. Later, a rapid increase or decrease in the selection of a response, which shows the rate of information accrual over additional processing time. Finally, an asymptote is observed, indicating the response rate, which does not further improve with additional time. The shape of this function is usually well fit by an exponential approach to a limit (see Fig. 1, left panel).

Four parameters describe the SAT function for the probability of selecting either one option: (a) The first asymptote (λ_1), reflecting overall limitations of selecting either response (greater or less than the criterion value) as an increase in total processing time does not result in additional increase (or decrease) in response rates; (b) the prior asymptote (λ_2), reflecting the response rate early at decision and prior to the responses being differentiated; (c) an intercept (δ), indicating the point in time at which performance departs from randomly responding; and (d) a rate (β) of rise from random responding to a differentiated responding. Parameter λ_1 indicates the maximum response rate that can be reached and λ_2 parameter indicates bias, if there is any, towards either one of the responses, while the intercept and the rate parameters constitute the speed of processing.

Experiment 1

In this experiment, we investigated how numerosity was perceived by making a decision about whether the number of dots presented on the screen was greater than 50 or not. There were three aims of this experiment: (1) Participants received feedback on the actual number of dots immediately following their response, which allowed them to compare their response with the probability of dots being greater than 50. This feedback gave participants the opportunity to calibrate their mental number line. (2) By employing the response deadline procedure, we controlled the speed of processing. This provided the additional advantage to measure when accuracy reached its maximum, and whether there was a tendency towards either response prior to the evidence accumulation. (3) Providing a feedback on the mental number line along with measuring the asymptotic proportion of responses allowed us to measure whether underestimation in numerosity judgments occurs, once accuracy reaches its maximum.

Method

Participants

Ten students from Middle East Technical University, with a mean age of 25 years (*SD* = 2.32), took part in the experiment and received monetary compensation for their participation. One participant dropped out of the experiment after the first session. Eight of the remaining nine participants were female, and nine were right-handed.

Stimuli

The experiment was conducted on PsychoPy2 using the *DotStim* function (Peirce et al., 2019) with 10-px sized dots presented in a 600 × 600 pixels circle field centered on a 1,280 × 720 pixels monitor. The location of each dot was sampled from a uniform distribution, which consequently suggests a random density condition for each trial. The number of dots presented was sampled randomly from the Binomial distribution with a size of 100 dots and four probability conditions, 0.30, 0.45, 0.55, and 0.70, randomly assigned for each trial. However, note that in the .45 and .55 probability conditions, the set of sampled number of dots exceeded (e.g., 53) or were under (e.g., 47) the threshold of 50 dots. These trials were still

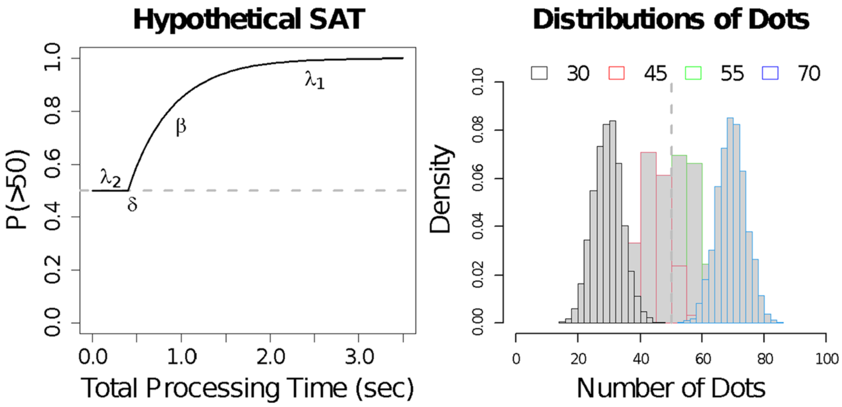

Fig. 1 Illustration of hypothetical speed–accuracy trade-off function (left) and distribution of dots presented in the experiments (right)

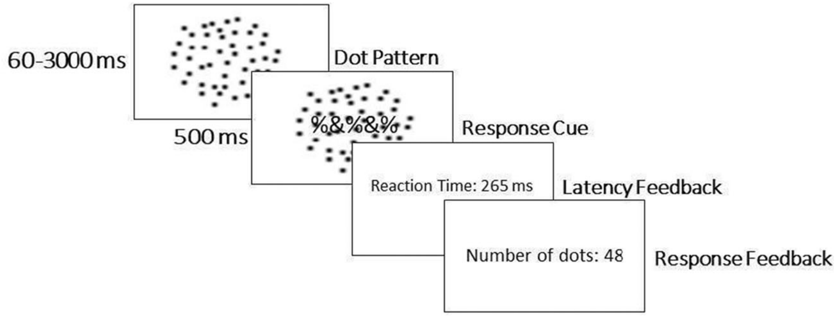

Fig. 2 Illustration of a trial presented in the experiment

considered in the .45 or in the .55 probability conditions respectively in the subsequent analysis (see Fig. 1 for the distribution of dot numbers). Dots were presented in white ink on a black background, and the response signal (i.e., '&#&#&#&#&') was presented in yellow ink.

Design and procedure

Participants completed three sessions, which included four blocks of 140 trials after the practice block in each session making a total of 1,680 trials for each participant. In each trial, dots were presented at the center of the screen. The number of dots ranged from 10 to 90 as they were sampled from four different binomial distributions, with a size of 100 and probabilities of 0.30, 0.45, 0.55, and 0.70 occurrences (see Fig. 1 for the histogram of how the dots were distributed throughout the experiment). The presentation of dots was followed by a signal that cued the participant to respond at 60, 100, 200, 500, 700, 1,500, 3,000 ms after the stimulus onset. Through the practice blocks, participants were trained to respond within 500 ms after the signal, and they received feedback on latency. If they failed to respond within 500 ms after signal onset, they received a warning saying "please respond within 500 ms," but the trials until 600 ms were counted in the subsequent analysis. If the participants responded before they were cued with a signal, they received a warning to wait for the signal, and these trials were excluded for the subsequent analysis. As for the response, participants were asked to decide whether the number of dots exceed 50, and if so, they were instructed to press "m" (otherwise, "z") on the keyboard. Later, participants were presented with the exact number of dots on the screen to calibrate their responses in the subsequent trials and required to press a key to proceed (see Fig. 2 for an illustration of a trial).

The experiment was a 7 (lag conditions: 60, 100, 200, 500, 700, 1500, and 3000 ms) × 4 (dot condition: 0.30, 0.45, 0.55, 0.70) within-subjects factorial design. There were 60 trials for each response lag and dot condition, adding up to a total of 1,680 (60 × 7 × 4) trials for each participant. All conditions were presented in a random order. Trials with latency greater than 600 ms and responses given earlier than the signal were removed for the subsequent analysis. On average, the remaining number of trials were 229.67 (*SD* = 6.38, 95% of all trials), 233.78 (*SD* = 4.74, 97%), 229.34 (*SD* = 5.64, 95%), 223.77 (*SD* = 6.88, 93%), 225.66 (*SD* = 8.25, 94%), and 218.11 (*SD* = 11.59, 91%) for 60, 100, 200, 300, 500, 700, 1,500, and 3,000 ms response lag conditions, respectively.[1]

Results and discussion

Response rates

We first analyzed the probability of accepting the number of dots as greater than 50 across response lag and dot conditions. A 7 × 4 repeated-measures analysis of variance (ANOVA) on the response rates revealed a main effect of dot condition, $F(3, 24) = 426.08$, $p < .001$, $MSE = 0.018$, $\eta_p^2 = 0.982$. The probability of responding "greater than 50" was less for the dots that were sampled from the .30 condition, $M = .114$ ($SE = 0.02$), compared with the dots that were sampled from the .45 condition (M =.391 , $SE = 0.02$), $t(24) = -11.56$, $p <$ .001, as suggested by the repeated contrasts. The probability of responding "greater than 50" increased more when compared with the dots sampled from .55 condition (M =.696, $SE = 0.02$), $t(24) = -12.68$, $p < .001$, as was the increase in the .70 condition (M =.914 , $SE = 0.02$), $t(24) = -9.131$, $p < .001$. These results imply that with an increase in the probability of dot sampling, more dots exceeded the threshold of 50, and consequently people accepted more trials as being greater than 50.

Additionally, the main effect of lag was significant, $F(6, 48) = 16.76$, p < .001, $MSE = 0.013$, $\eta_p^2 = 0.677$, suggesting a change in the probability of responding "greater than 50" as the response signal delays. The repeated contrasts indicated a decrease in the probability from the earliest lag, 60 ms (M =

[1] Note that the percentage of removed trials was comparable across deadline conditions (e.g., 60 ms versus 700 ms) when the delay after cue onset was set to 600 ms or less. This might indicate that the participants waited to fully process the screen in early cues, whereas they responded before receiving the cue in the latest deadline condition (3,500 ms). However, when the delay after cue onset was set to 500 ms or less, the percentage of the remaining trials dropped to 83%, 90%, 94%, 94%, 92%, 93%, 90% for each deadline condition, respectively. That said, the asymptotes of the best fitting model did not differ while the speed parameters showed faster processing in general. The results of the additional analysis are presented in the supplementary materials for all experiments.

.671, *SE* = 0.036), *t*(48) = 2.68, $p < .01$, to the probability at 100 ms (*M* = .599, *SE* = 0.21), averaged across all dot conditions. A similar decrease was observed from 100 ms to 200 ms (*M* = .509, *SE* = 0.18), *t*(48) = 3.39, $p < .001$, and after 200 ms, the probability of responding "greater than 50" stayed steady at around .50. Together, these findings indicate that at early lags participants responded "greater than 50" more often than they did in later lags, independent of the dot condition. Half of the trials had dots greater than 50 while the other half had less than 50. This would suggest that if participants were responding based on the perceptual evidence, they would have responded .50 in all response conditions. However, the data showed that at the earliest lags (60 ms and 100 ms) participants overestimated the number of dots presented on the screen.

The interaction between lag and dot condition was significant, *F*(18, 144) = 18.536, $p < .001$, *MSE* = 0.018, $\eta_p^2 = 0.70$. Below, we presented the SAT function that explained the interaction between lag and dot condition in more detail.

Speed–accuracy trade-off functions

Figure 3 presents the time course of responding "greater than 50," along with the function that is obtained by the best fitting SAT parameters. Decisions made at early lags indicate the probability of responding "yes" to the question before the evidence accumulation starts—in other words, performance at chance level. Here, performance at chance would be expected to be 0.5, as the probability of responding "greater than 50" would be independent of the number of dots due to lack of evidence accumulation. Values greater than .5, as shown in the ANOVA results, would indicate a bias towards the "greater than 50" response. Similarly, values less than .5 would indicate a bias towards the "less than 50" response. After a point in time, as evidence starts to accumulate, the probability of responding "greater than 50" becomes differentiated based on the dot conditions. For the 0.30 condition, the probability is expected to decrease while it is expected to increase for the .70 condition. Later, the probability of "greater than 50" response reaches an asymptote, an indication of maximum evidence accumulation. As the limits for processing are reached, allowing longer time to process the number of dots does not have additional benefits. The asymptotic probability to respond "greater than 50" is expected to change as a function of the probability parameter used in the sampling binomial distribution. For example, the probability of responding "greater than 50" is expected to be lowest in the .30 dot condition and to be highest in the .70 condition. Finally, the rate of evidence accumulation is expected to be the same across dot conditions, as is the time point that indicates when responses depart from chance. That is due to the fact that while participants were asked to respond whether the number points presented on the screen was greater than 50, they did not know about the dot conditions or the response lag conditions. Thus, there was no reason not to assume that the rate of evidence accumulation and the point at which evidence accumulation starts would be different across dot conditions.

The probability to respond "greater than 50" is estimated with an exponential function, which provides independent and unbiased estimates of asymptotic probabilities and processing speed. The following exponential function can be employed for further investigation of the response bias observed in early lags:

$$P(Dots > 50) = \lambda_1 + (\lambda_2 - \lambda_1)\left(e^{-\beta(t-\delta)}\right), t > \delta, else\ \lambda_2, \quad (1)$$

where *P*('yes') is the probability to respond that the number of dots being greater than 50; $\lambda_1$is the asymptotic probability to accept the number of dots as being greater than 50 at the late response lags; λ_2 is the asymptotic probability to accept the number of dots as being greater than 50 at the early response lag conditions (at chance level before the information begins to

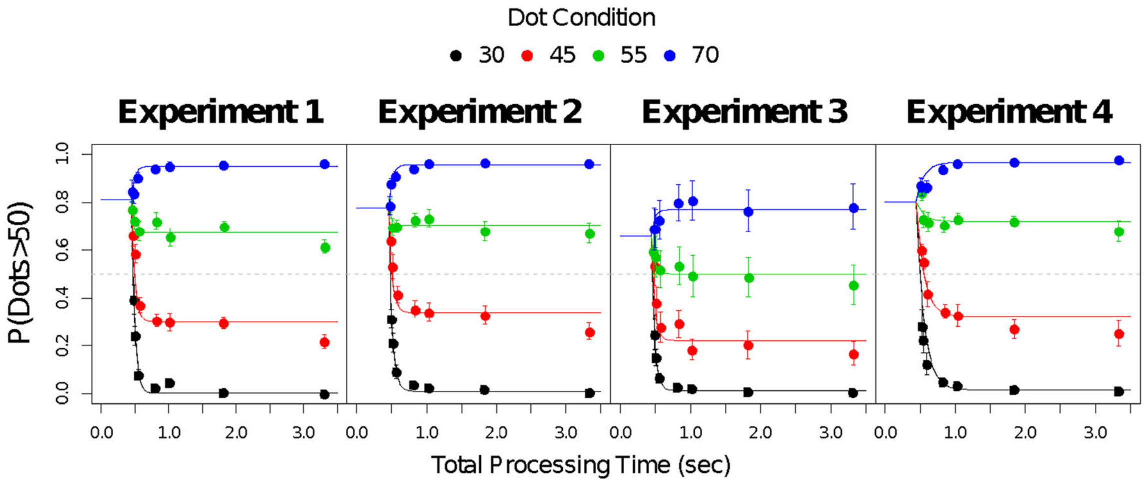

Fig. 3 Speed–accuracy trade-off functions across all experiments

accumulate); β is the rate of accumulation towards the asymptotic probability at later lags; δ is the time point when the information begins to accumulate, which shows the point when the probability departs from chance; and t is the total processing time that includes the time before signal onset and latency.

The data were fitted with the exponential function (Equation 1) using the *optim* function in R (R Core Team, 2019) to estimate the seven parameters with the maximum likelihood estimation (MLE) method. Initially, the group data, averaged across participants were fit with seven parameters: four λ_1 for each response condition (.30, .45, .55, and .70); a unique λ_2 across all dot conditions, which shows the bias prior to evidence accumulation; and, similarly, β and δ were also unique across all dot conditions. The best fitting parameter values are presented in Table 1, along with the model fit values. The parameter values of λ_1 shows an increase in the probability of responding "greater than 50" as a function of an increase in the probability of sampling the dot numbers from the Binomial distribution. For example, when the probability of sampling from the Binomial distribution increased from .30 to .70, the probability of responding "greater than 50" increased from .002 to .95. As was observed in the results of response rate analyses (ANOVA), the value of λ_2 indicates that participants had a tendency to respond "greater than 50" prior to accumulating perceptual evidence. That is, on average 81% of the trials would have received "greater than 50" responses prior to the point in time at which evidence started to accumulate. The intercept parameter (δ) also suggests that after 452 ms, evidence started to accumulate towards the λ_1 asymptote of the corresponding dot condition. Similarly, the rate parameter showed that 1% change in the probability of responding "greater than 50" occurred in 5 ms ($1/\ \beta$ = 1/21.60). That is, the evidence for greater or less than 50 points accumulated over the course of processing at 5 ms intervals. Finally, the R^2 value of .986 shows that 98% of the variance in the averaged data is explained by the proposed model parameters.

When the data obtained from each participant were fit with the same model (Equation 1), and the obtained parameters were averaged across participants, similar values were found (see Table 2). The average λ_1 values were .002 (*SD* = 0.003), .30 (*SD* = 0.057), .675 (*SD* = 0.056), and .975 (*SD* = 0.025) for the .30, .45, .55, and .70 dot conditions, respectively. The average λ_2 value was .80 (*SD* = 0.175), replicating the bias towards "greater than 50" response observed above. Similarly, the averaged speed parameters were close to the parameters obtained from the averaged data. The rate of accumulation, β, was 21.707 (*SD* = 6.18) and the point in time when information starts to accumulate, δ, was 419 ms (*SD* = 122).

Finally, we investigated the systematic underestimation of numerosity once the asymptotic probability of responding "greater than 50" was reached starting from the 500-ms lag condition. When the probability of sampling from the binomial distribution was .45, for the trials in which the number of dots exceeded 50, the average number of dots presented across lag conditions and participants was 53.17 (*SD* = 2.32). Participants responded "greater than 50" for only 63% (*SD* = .14) of those trials. This value increased to 76% (*SD* =0.07) when the probability of sampling from the binomial distribution was .55 and the average of the actual number was 56.84 (*SD* = 3.95) across all lag conditions greater than 500 ms. Finally, the highest proportion of responding "greater than 50" was for the trials in which the probability of sampling from the binomial distribution was .70 with 95% (*SD* = 0.04). As expected, the highest mean of actual number of dots presented was observed in the .70 dot condition, 69.74 (*SD* = 4.59). Note that as the average number of dots increased as a function of the probability of sampling from the binomial distribution, the proportion of responding "greater than 50" increased. Additionally, the fact that the reduced proportion compared with 1 indicates underestimation of dot numbers, even for the .70 dot condition. That is because all of these trials were in fact presenting dot patches numbered greater than 50. As Weber's law posits, the difference between the criterion value, 50 in this experiment, and the actual value is numerically close to one another, the performance gets closer to chance compared with when those numbers are farther.

To summarize, the results of Experiment 1 showed that receiving the actual number of dots presented in the preceding numerosity perception task resulted in a calibration of the

Table 1 Parameter values of the best fitting exponential function

	λ_1 30	λ_1 45	λ_1 55	λ_1 70	λ_2	β	δ	Deviance	R^2
Experiment 1	.004	.288	.669	.950	.802	21.52	0.449	−46.787	.985
Experiment 2	.010	.338	.695	.957	.765	22.88	0.459	−45.665	.980
Experiment 3	.011	.205	.492	.768	.678	29.44	0.436	−26.730	.730
Experiment 4	.015	.322	.720	.969	.799	8.80	0.432	−44.590	.943

Note. Parameter values are obtained from the fits to the data averaged across participants. λ_1 is the asymptotic probability of accepting the number of dots to be greater than 50 for each dot condition, 30, 45, 55, and 70, respectively. λ_2 is the probability of accepting the number of dots to be greater than 50 at chance and indicates the bias at early retrieval. β is the rate of accumulation towards asymptote, and δ is the time point at which the probability of accepting the number of dots start differing from bias (λ_2)

Table 2 Parameter values of the best fitting exponential function to individual data in Experiment 1

Participants	λ_1 30	λ_1 45	λ_1 55	λ_1 70	λ_2	β	δ	Deviance	R^2
1	.006	.287	.614	.930	.709	27.64	0.447	−42.447	.932
2	.005	.240	.637	.958	.542	3.15	0.100	−42.466	.957
3	.001	.264	.665	.960	.730	14.55	0.424	−41.902	.961
4	.001	.256	.687	.978	.982	30.00	0.513	−41.541	.793
5	.014	.328	.619	.902	.672	18.33	0.473	−37.545	.932
6	.001	.221	.609	.941	.990	24.69	0.465	−36.146	.763
7	.001	.358	.721	.978	.512	5.19	0.352	−43.955	.949
8	.001	.326	.789	.990	.599	25.64	0.424	−44.362	.971
9	.005	.279	.705	.965	.959	28.39	0.468	−44.031	.955

Note. Parameter values are obtained from the fits to each participant data. λ_1 is the asymptotic probability of accepting the number of dots to be greater than 50 for each dot condition, 30, 45, 55, and 70, respectively. λ_2 is the probability of accepting the number of dots to be greater than 50 at chance and indicates the bias at early retrieval. β is the rate of accumulation towards asymptote, and δ is the time point at which the probability of accepting the number of dots start differing from bias (λ_2)

mental number line. Participants were able to respond with a matching probability of the number of dots exceeding 50 in a given dot patch. Next, employing the response deadline procedure resulted in an observation of prior response bias. That is, participants had a tendency to respond greater than 50, before the accumulation of perceptual evidence starts. Finally, the parameter values obtained from the SAT curve showed that participants reached their maximum accuracy after onset of 500-ms response condition earliest. Once accuracy reached its maximum, participants still showed somewhat underestimation, even when the number of dots on average exceeded 70. In the following experiment, we further investigated whether a similar prior bias towards “greater than 50” response occurred when participants were asked to respond “less than 50.”

Experiment 2

The aim of this experiment was to further investigate the bias towards “greater than 50” response, by requiring participants to respond to a contrasting question, whether the number of dots was *less* than 50. Similar to Experiment 1, participants received the actual number of dots presented in the trial that precedes their response, allowing them to calibrate their mental number line. As in Experiment 1, the response deadline procedure was employed to obtain independent and unbiased measures of speed and accuracy (sensitivity) in processing of numerosity perception. Specifically, the prior bias will be measured by obtaining the probability of responding “greater than 50” as in Experiment 1 before perceptual evidence accumulation starts. The only difference in Experiment 2 was requiring the participants to answer the question of whether the number of dots presented were *less* or not, instead of asking whether they were greater or not. Afterwards, the asymptotic probability of responding “greater than 50” was measured, and similarly, were further tested on the possible underestimation results across higher number of sampling probabilities, such as .45, .55, .70.

Method

Participants

Eleven students from Middle East Technical University, with a mean age of 23 years (*SD* = 3.57), took part in the experiment and received monetary compensation for their participation. Six were female, and 11 were right-handed.

Stimuli

The stimuli were identical to those of Experiment 1.

Design and procedure

The Design and Procedure were identical to those of Experiment 1. Similar to Experiment 1, trials with latency greater than 600 ms and responses given earlier than the signal were removed for the subsequent analysis. On average, the remaining number of trials were 210.00 (*SD* = 43.40, %88 of all trials in each response deadline condition), 217.91 (*SD* = 33.17, %90), 225.09 (*SD* = 27.94, %94), 221.18 (*SD* = 23.22, %92), 218.72 (*SD* = 27.18, %90), 216.18 (*SD* = 26.32, %90), 210 (*SD* = 25.15, %88) for the 60, 100, 200, 300, 500, 700, 1,500, and 3,000-ms lag conditions, respectively, for each participant.

Results and discussion

Response rates

In this experiment, participants were asked to compare the perceived number of dots to a criterion of 50 as in Experiment 1. However, different from Experiment 1, participants were asked to judge whether the number of dots was less than 50. As they responded with the same keys—that is, pressing “m” on the keyboard for indicating that the number of dots was greater than 50, and pressing “z” on the keyboard for indicating that the number of dots was less than 50—the same analyses were conducted on the probability of responding “greater than 50” as in Experiment 1.

Similar to Experiment 1, a 7 × 4 repeated-measures ANOVA on the response rates revealed a main effect of dot condition, $F(3, 30) = 539.60$, $p < .001$, $MSE = 0.018$, $\eta_p^2 = 0.982$. The probability of responding “greater than 50” decreased for the dots that were sampled from the .30 condition, $M = .10$ ($SD = 0.13$), compared with the dots that were sampled from the .45 condition ($M = .41$, $SD = 0.18$), $t(30) = -14.24$, $p < .001$. The probability of responding “greater than 50” increased when compared with the dots sampled from .55 condition ($M = .71$, $SD = 0.12$), $t(30) = -13.91$, $p < .001$, as was the increase in the .70 condition ($M = .92$, $SD = 0.09$), $t(30) = -9.339$, $p < .001$. These results imply that with an increase in the probability of dot sampling, more dots exceeded the criterion of 50 and consequently participants accepted more trials to be greater than 50.

The main effect of lag was significant, as expected, $F(6, 60) = 8.25$, $p < .001$, $MSE = 0.124$, $\eta_p^2 = 0.461$, suggesting a change in the probability of responding “greater than 50” as the response signal was received later. The repeated contrasts indicated a marginal decrease in the probability from the earliest lag, 60 ms ($M = .63$, $SD = 0.23$), $t(60) = 1.95$, $p = .06$, to the probability at 100 ms ($M = .58$, $SD = 0.28$), averaged across all dot conditions. A decrease was also observed from 100 ms to 200 ms ($M = .53$, $SD = 0.32$), $t(60) = 1.96$, $p = .06$, and after 200 ms, the probability of responding “greater than 50” stayed steady at around .50. As in Experiment 1, participants responded “greater than 50” more often at early lags than they did in later lags, independent of the dot condition. This was a replication of the finding that participants had a bias towards the “greater than 50” response prior to accumulation of evidence as shown in Experiment 1. The interesting finding here is that in contrast to Experiment 1, participants were asked whether the number of dots was fewer than the criterion value of 50. The ANOVA results indicated a similar prior bias observed in Experiment 1.

The interaction between lag and dot condition was significant, $F(18, 180) = 19.12$, $p < .001$, $MSE = 0.005$, $\eta_p^2 = 0.66$. In the next section, the SAT functions will be analyzed in more detail for the interaction between lag and dot condition.

Speed–accuracy trade-off functions

Figure 3 depicts the time course of responding “greater than 50,” along with the function that is obtained by the best fitting SAT parameters. The probability to respond “greater than 50” was estimated with the exponential function presented in Equation 1. The modeling routine was the same as in Experiment 1. There were 7 free parameters, which measured the probability to respond “greater than 50” once the asymptote is reached, represented by λ_1, one for each dot condition, then λ_2, representing the prior bias, then the speed parameters, β and δ, which measure the rate of evidence accumulation and the point in time when the probabilities differentiate from each other. The best fitting parameter values are presented in Table 1, along with the model fit values. As in Experiment 1, when the probability of sampling from the Binomial distribution increased from .30 to .70 across conditions, the parameter values of λ_1 increased from .01 to .34, .70, .96 respectively. When the prior bias was evaluated, the value of λ_2 indicates that the participants had a tendency to respond “greater than 50” prior to accumulating perceptual evidence. On average 77% of the trials would have received “greater than 50” responses prior to the point in time at which evidence started to accumulate. The time at which evidence starts to accumulate is represented by δ, suggesting that after 458 ms, evidence started to accumulate with a rate of 4 ms ($1/\beta = 1/22.91$) per 1% increase in the evidence. That is, the evidence for greater or less than 50 points accumulated over the course of processing at 5 ms intervals similar to Experiment 1. Finally, the R^2 value of .98 shows that 98% of the variance in the averaged data is explained by the proposed model parameters.

When the data obtained from each participant was fit with the same model (Equation 1) and the obtained parameters were averaged across participants, similar values were found (see Table 3). The average λ_1 values were .01 ($SD = 0.01$), .33 ($SD = 0.09$), .71 ($SD = 0.04$), and .96 ($SD = 0.04$) for the .30, .45, .55, and .70 respectively. The average λ_2 value was .72 ($SD = 0.18$), replicating the bias towards “greater than 50” response observed in Experiment 1. Similarly, the averaged speed parameters were close to the parameters obtained from the averaged data. β was 15.69 (SD = 10.43) and δ was 390 ms (SD = 110). The best fitting parameter values obtained from the averaged data and the averaged parameter values obtained from fitting individuals’ data were in congruence with each other.

As in Experiment 1, a similar systematic underestimation of numerosity was observed when the probability of responding “greater than 50” reached its maximum starting at the 500 ms lag condition. When the sampling probability in the binomial distribution was .45 and the trials with the number of dots exceeding 50, the average number of dots presented was 53.22 ($SD = 0.55$). As in Experiment 1, participants responded “greater than 50” for only 63% ($SD = .48$) of

Table 3 Parameter values of the best fitting exponential function to individual data in Experiment 2

Participants	λ_1 30	λ_1 45	λ_1 55	λ_1 70	λ_2	β	δ	Deviance	R^2
1	.001	.169	.530	.873	.703	35.00	0.419	−40.558	.903
2	.001	.317	.763	.990	.977	35.00	0.581	−41.564	.874
3	.009	.333	.659	.944	.956	13.66	0.437	−36.076	.950
4	.001	.370	.714	.989	.692	24.31	0.463	−41.029	.963
5	.002	.367	.664	.948	.684	3.59	0.291	−43.012	.959
6	.005	.293	.626	.970	.653	26.80	0.408	−39.554	.935
7	.002	.309	.692	.962	.913	33.95	0.475	−41.940	.873
8	.004	.281	.739	.982	.755	35.00	0.422	−41.483	.963
9	.018	.386	.759	.990	.335	4.98	0.257	−40.925	.957
10	.008	.442	.778	.960	.624	10.99	0.359	−40.354	.965
11	.022	.409	.785	.977	.492	4.55	0.478	−45.177	.992

Note. Parameter values are obtained from the fits to each participant data. λ_1 is the asymptotic probability of accepting the number of dots to be greater than 50 for each dot condition, 30, 45, 55, and 70, respectively. λ_2 is the probability of accepting the number of dots to be greater than 50 at chance and indicates the bias at early retrieval. β is the rate of accumulation towards asymptote, and δ is the time point at which the probability of accepting the number of dots start differing from bias (λ_2)

the trials. An increase in the percentage of responding "greater than 50" was observed as 76% (*SD* =0.42) when the sampling probability from the binomial distribution increased to .55 and the average of the actual number size was 56.84 (*SD* = 0.28) at the asymptotic lag conditions. The highest proportion of responding "greater than 50", 95% (*SD* = 0.19), was observed for the .70 dot condition with the highest mean of actual number of dots, 70.03 (*SD* = 0.32). A similar underestimation of dot numbers was observed as in Experiment 1, even for the .70 dot condition. These results, as in Experiment 1, indicate that Weber's law holds, when the difference between the criterion value, 50 in this experiment, and the actual value is numerically close to one another, the performance gets closer to chance compared with when those numbers are farther.

To summarize, the results of Experiment 2 replicated the results of Experiment 1. First, asking subjects to make their decisions by changing the direction of comparison did not have a differential effect on the numerosity judgments. A response bias towards responding "greater than 50" was observed even though participants were asked whether the number of dots was less than 50. This finding suggests that when participants judge the numerosity of a collection of dots, they do not evaluate the direction of the judgment (greater or less). As the final point, similar to the findings observed in Experiment1, once accuracy reached its maximum, participants still showed a slight underestimation, even when the number of dots on average exceeded 70. In the following experiment, we further investigated the underestimation when participants were not given any feedback on the actual number of dots that they have been required to judge. An absence of a feedback would result in an uncalibrated mental number line, and as a result, an increased underestimation at the asymptotic probabilities.

Experiment 3

The aim of this experiment was to investigate whether not receiving feedback would harm the calibration of the mental number line. In Experiment 1 and 2, participants received the actual number of dots presented on the preceding numerosity trial and their probability of responding "greater than 50" matched with the probability of the trials having dots greater than 50. Although there was a slight underestimation, especially after when the probabilities reached its asymptote, a greater underestimation would be expected in an experiment when no feedback was provided to the participant. The current experiment applies the same procedures of Experiment 1, where there were 4 dot conditions with 4 different probabilities to sample the number of dots in each trial. Participants were asked whether the number of dots exceeded the criterion value, 50. Different from the previous experiments, participants were not provided with feedback on the actual number of dots presented in the experiment. The response deadline procedure was employed as was in Experiment 1 and 2, in order to obtain independent measures of speed and accuracy in the numerosity task.

Method

Participants

Ten students from Middle East Technical University with a mean age of 22.5 (*SD* = 1.97) took part in the experiment and received monetary compensation for their participation. Data from one participant was removed from the subsequent analysis because they did not comply with the instructions. Five of the remaining participants were female and 8 were right-handed.

Stimuli

The stimuli were identical to those of Experiment 1 and 2.

Design and procedure

The Design and Procedure were identical to those of Experiment 1, except that the participants did not receive any feedback in the current experiment. Similar to the previous experiments, trials with latency greater than 600 ms and responses given earlier than the signal were removed for the subsequent analysis. On average, the remaining number of trials were 215.66 (*SD* = 28.22, 89% of the trials), 227 (*SD* = 18.41, 95%), 229.44 (*SD* = 10.88, %96), 227.88 (*SD* = 9.48), 225.22 (*SD* = 9.84), 220.22 (*SD* = 9.98, %92), 217 (*SD* = 12.96, %90) ms for 60, 100, 200, 300, 500, 700, 1500, and 3000 ms lag conditions, respectively averaged across participants.

Results and discussion

Response rates

In this experiment, participants were asked to make a decision on whether the number of dots exceeded 50, as in Experiment 1. Different from the previous experiments, however, participants did not receive any feedback regarding the actual number of dots presented for the trial that they have just responded to.

Similar to the previous experiments, a 7 × 4 repeated-measures ANOVA on the response rates revealed a main effect of dot condition, $F(3, 24) = 71.12, p < .001, MSE = 0.08, \eta_p^2 = 0.90$. The probability of responding “greater than 50” decreased for the dots that were sampled from the .30 condition, $M = .08$ ($SE = 0.06$), compared with the dots that were sampled from the .45 condition ($M = .29$, $SE = 0.06$), $t(24) = -4.41, p < .001$. The probability of responding “greater than 50” increased when compared with the dots sampled from .55 condition ($M = .52$, $SE = 0.06$), $t(24) = -4.73$, $p < .001$, as was the increase in the .70 condition ($M = .75$, $SE = 0.06$), $t(24) = -4.67$, $p < .001$. Note the relatively smaller probabilities of responding “greater than 50” in the .55 and the .70 dot conditions compared with those in Experiments 1 and 2, where the feedback was provided. These results might imply that the increase in number of dots in a trial might have resulted in the tendency to underestimate when the mental number line was not calibrated by feedback.

The main effect of lag was significant, $F(6, 48) = 7.57, p < .001, MSE = 0.02, \eta_p^2 = 0.48$, suggesting a change in the probability of responding “greater than 50” as the response signal was delayed. The repeated contrasts indicated a marginal decrease in the probability from the earliest lag, 60 ms ($M = .0.52$, $SE = 0.06$), $t(48) = 2.31, p = .03$, to the probability at 100 ms ($M = .45$, $SE = 0.06$), averaged across all dot conditions. After the 100 ms lag condition, no significant difference was observed across lag conditions. The average probability of responding “greater than 50” stayed steady around 0.38, suggesting that the participants used the “lower than 50” response more often than the other response.

The interaction between lag and dot condition was significant, $F(18, 144) = 9.11, p < .001, MSE = 0.006, \eta_p^2 = 0.53$. In the next section, the SAT functions will be analyzed in more detail for the interaction between lag and dot condition.

Speed–accuracy trade-off functions

Figure 3 depicts the time course of responding “greater than 50,” along with the function that is obtained by the best fitting SAT parameters. The probability to respond “greater than 50” was estimated with the exponential function presented in Equation 1. The modeling routine was the same as in the previous experiments. The best fitting parameter values are presented in Table 1, along with the model fit values. As in previous experiments, when the probability of sampling from the binomial distribution increased from .30 to .45, the probability of responding “greater than 50,” λ_1 , increased from .01 to .22. Similarly, as the sampling probability increased from .45 to .55, the probability of responding “greater than 50” increased to .51 and finally, this value increased to .77 for the .70 dot condition. Again, note that the λ_1 values in this experiment were lower than the corresponding values in the earlier experiments, suggesting an underestimation in the absence of feedback.

The value obtained for $\lambda_2$suggests that on average 66% of the trials would have received “greater than 50” responses prior to the point in time at which evidence started to accumulate. In relation to underestimation, the value of λ_2 indicates a bias towards responding “greater than 50”; however, λ_2 was found to be lower in comparison with that of in earlier experiments. That is, the response bias was also affected by the general underestimation in the absence of feedback and resulted in a relatively low value compared with when feedback was provided. The speed parameters were in accordance with the speed parameters obtained from the earlier experiments. δ, suggested that after 452 ms, evidence started to accumulate with a rate of 3 ms ($1/\beta = 1/29.52$) per 1% increase in the evidence. Finally, the R^2 value of .73 shows that 73% of the variance in the averaged data is explained by the proposed model parameters.

When the data obtained from each participant was fit with the same model (Equation 1) and the obtained parameters were averaged across participants, similar values were found (see Table 4). The average λ_1 values were .01 ($SD = 0.01$), .22

Table 4 Parameter values of the best fitting exponential function to individual data in Experiment 3

Participants	λ_1 30	λ_1 45	λ_1 55	λ_1 70	λ_2	β	δ	Deviance	R^2
1	.003	.138	.511	.911	.728	35.00	0.440	−26.034	.964
2	.001	.227	.602	.898	.559	20.45	0.424	−25.440	.922
3	.020	.421	.800	.966	.806	4.73	0.240	−23.576	.950
4	.001	.100	.300	.800	.200	18.25	0.454	−26.529	.814
5	.001	.229	.544	.931	.990	17.89	0.341	−25.392	.950
6	.005	.208	.630	.942	.990	35.00	0.448	−25.123	.943
7	.002	.100	.455	.800	.990	27.95	0.444	−24.642	.764
8	.001	.100	.300	.800	.200	4.22	0.299	−25.894	.853
9	.046	.406	.692	.935	.501	18.20	0.452	−24.317	.932

Note. Parameter values are obtained from the fits to each participant data. λ_1 is the asymptotic probability of accepting the number of dots to be greater than 50 for each dot condition, 30, 45, 55, and 70, respectively. λ_2 is the probability of accepting the number of dots to be greater than 50 at chance and indicates the bias at early retrieval. β is the rate of accumulation towards asymptote, and δ is the time point at which the probability of accepting the number of dots start differing from bias (λ_2)

(*SD* = 0.14), .51 (*SD* = 0.24), and .78 (*SD* = 0.25) for the .30, .45, .55, and .70 respectively. The average λ_2 value was .63 (*SD* = 0.33), replicating the bias towards "greater than 50" response observed above. Similarly, the averaged speed parameters were relatively similar as the parameters obtained from the averaged data. The rate of accumulation, β, was 19.63 (*SD* = 9.71) and δ was 407 ms (*SD* = 70). The best fitting parameter values obtained from the averaged data and the averaged parameter values obtained from fitting individuals' data were in congruence with each other.

As in previous experiments, systematic underestimation of numerosity was observed when the probability of responding "greater than 50" reached its maximum starting at the 500 ms lag condition. When the sampling probability in the binomial distribution was .45 and the trials with the number of dots exceeding 50, the average number of dots presented was 53.22 (*SD* = 2.34). Participants responded "greater than 50" for only 47% (*SD* = .50) of the trials. When the sampling probability from the binomial distribution increased to .55 and the average of the actual number size was 56.67 (*SD* = 3.88) at the asymptotic lag conditions, the percentage of responding "greater than 50" increased to 55% (*SD* =0.50). Finally, the highest proportion of responding "greater than 50," 79% (*SD* = 0.40), was observed in the .70 dot condition, which had the highest mean of actual number of dots, 69.80 (*SD* = 4.78). When compared with the results in previous experiments, when feedback was provided, these results indicate a systematic underestimation. The underestimation was observed mostly for the larger number of dots, consistent with the literature. Even for the .70 dot condition, participants responded 78% of the trials as greater than 50, while the expected response rate was 1 due to the number of dots being greater than 50 for all trials. Similar to the results of the previous experiments, Weber's law holds, the difference between the criterion value (50 in this experiment) and the actual value is numerically close to one another, the performance gets closer to chance compared with when those numbers are farther. This result was observed in the .45 condition, when the number of dots exceeded 50, the response rate was lower than chance, .47. That is, even though the number of dots exceeded 50, participants were less likely to respond as "greater than 50," which is a demonstration of underestimation.

To summarize, when participants were not presented with a feedback, such as the actual number of dots in the preceding trial, underestimation of dots was observed. That indicates how the calibration of the mental number line is affected by the absence of feedback. Apart from the underestimation observed especially for larger dot number conditions, the results obtained from the SAT function in Experiment 3 replicated the results from the previous experiments. Namely, despite an underestimation across dot conditions, there was a bias towards the "greater than 50" response prior to accumulation of evidence. Finally, the speed parameters were comparable with the previous experiments, proposing that the rate of accumulation on evidence and the point in time when accumulation starts were similar across feedback conditions. In the following experiment, we further investigated the prior bias towards the "greater than 50" response by reversing the response-key assignment.

Experiment 4

In the previous three experiments, a prominent bias towards responding "greater than 50" was observed especially prior to evidence accumulation starts. In this experiment, we investigated whether such bias would occur if the response-key assignment was reversed such that the "greater than 50" response was assigned on the left side of the keyboard while the "less than 50" response was assigned on the right side of the keyboard. The question was if the response keys would be incongruent with the spatial correspondence of the mental number line, would we still observe a bias towards the "greater than 50" response prior to evidence accumulation. If so, we could claim an automatic bias towards accepting a cloud of dots as greater than 50, before evidence accumulation starts.

Method

Participants

Thirteen students from Çankaya University, with a mean age of 25.6 years (*SD* = 6.8), volunteered to take part in the experiment. Data from four participants were removed from the subsequent analysis because three of them used the incorrect response-key assignment, and one performed at chance. All the remaining participants were female, and eight were right-handed.

Stimuli

The stimuli were identical to those of the previous experiments.

Design and procedure

The Design and Procedure were identical to Experiment 1, except that participants were instructed to press "m" to respond "the number of dots are less than 50" and press "z" to respond "the number of dots are greater than 50" on the keyboard. Similar to the previous experiments, trials with latency greater than 600 ms, and responses given earlier than the signal were removed for the subsequent analysis. On average, the remaining number of trials were 190.88 (*SD* = 58.43, 79% of trials), 209.77 (*SD* = 45.62, 87%), 223.11 (*SD* = 21.81, 93%), 228.00 (*SD* = 6.55, 95%), 225.66 (*SD* = 7.48, 94%), 220.22 (*SD* = 18.04, 92%), 216.00 (*SD* = 15.04, 90%) for 60, 100, 200, 300, 500, 700, 1,500, 3,000 lag conditions, respectively, averaged across participants. Note that the average number of trials was the lowest for 60-ms response lag condition across all conditions and experiments. That indicates the role of congruency between mental number line and response-key assignment. When participants were required to use opposite keys for "greater" and "less" responses, they failed to respond within the allotted time especially in 60 ms response lag condition.

Results and discussion

Response rates

In this experiment, participants were asked to make a response on whether the number of dots presented on the screen exceed the criterion, 50. Different from the previous experiments, participants were instructed to use a reverse response-key assignment.

Similar to the previous experiments, a 7 × 4 repeated-measures ANOVA on the response rates revealed a main effect of dot condition, $F(3, 24) = 275.161$, $p < .001$, $MSE = 0.03$, $\eta_p^2 = 0.97$. The probability of responding "greater than 50" decreased for the dots that were sampled from the .30 condition, $M = .11$ ($SE = 0.02$), compared with the dots that were sampled from the .45 condition ($M = .40$, $SE = 0.02$), $t(24) = -9.34$, $p < .001$. The probability of responding "greater than 50" increased when compared with the dots sampled from .55 condition ($M = .73$, $SE = 0.02$), $t(24) = -10.96$, $p < .001$, as was the increase in the .70 condition, ($M = .92$, $SE = 0.02$), $t(24) = -6.15$, $p < .001$. Even though the response-key assignment was incongruent, the initial results on response rates were compatible with the results in Experiments 1 and 2, when participants received feedback and used congruent response-key assignment.

The main effect of lag was significant, $F(6, 48) = 14.80$, $p < .001$, $MSE = 0.01$, $\eta_p^2 = 0.65$, suggesting a change in the probability of responding "greater than 50" with an increase in the response signal. The repeated contrasts indicated a decrease in the probability from the earliest lag, 60 ms ($M = .65$, $SE = 0.06$), $t(48) = 2.52$, $p = .02$, to the probability at 100 ms ($M = .45$, $SE = 0.06$), averaged across all dot conditions. Similarly, the rate of responding greater 50 decreased from 100 ms to 200 ms lag condition ($M = 0.59$, $SE = 0.02$), $t(48) = 2.30$, $p < .01$. For the longer response lag conditions, no significant difference was observed. The average probability of responding "greater than 50" stayed steady around 0.48, suggesting that the participants calibrated response selection.

The interaction between lag and dot condition was significant, $F(18, 144) = 9.72$, $p < .001$, $\mathrm{MSE} = 0.006$, $\eta_p^2 = 0.55$. In the next section, the SAT functions will be analyzed in more detail for the interaction between lag and dot condition.

Speed–accuracy trade-off functions

Figure 3 depicts the time course of responding "greater than 50," along with the function that is obtained by the best fitting SAT parameters. The probability to respond "greater than 50" was estimated with the exponential function presented in Equation 1. The modeling routine was the same as in the previous experiments. The best fitting parameter values are presented in Table 1, along with the model fit values. As in previous experiments, when the probability of sampling from the binomial distribution increased from .30 to .45, λ_1 increased from .02 to .32. Similarly, as the sampling probability increased from .45 to .55, λ_1 increased to .72 and finally, this value increased to .97 for the .70 dot condition.

The value obtained for $\lambda_2$suggests that on average 80% of the trials would have received "greater than 50" responses prior to the point in time at which evidence started to accumulate. This value, compared with the prior bias values in Experiments 1 and 2, indicate that the bias towards the "greater than 50" response was not affected by the response-key assignment. When an incongruent response-key assignment was employed, the direction of response bias at early lags did not change. This suggests that participants initially have a prior belief that the number of dots was greater than 50, and once they start processing the perceptual environment, they underestimate the number of dots.

Incongruent response-key assignment slowed the rate of evidence accumulation. In this experiment, the rate of evidence accumulation was found to be 11 ms (1/ β = 1/8.80) per 1% increase in the evidence. When the response-key assignment was congruent with the mental number line, rate was found to be 4, and a comparison between these values suggests the incongruency between mental number line and response-key assignment slows responding. Finally, $\delta$was similar across all experimental conditions. The R^2 value of .93 shows that 93% of the variance in the averaged data was explained by the proposed model parameters.

When the data obtained from each participant was fit with the same model (Equation 1) and the obtained parameters were averaged across participants, similar values were found (see Table 5). The average λ_1 values were .02 (*SD* = 0.02), .33 (*SD* = 0.08), .72 (*SD* = 0.05), and .97 (*SD* = 0.02) for .30, .45, .55, and .70, respectively. The average λ_2 value was .84 (*SD* = 0.14), replicating the bias towards "greater than 50" response observed above. Similarly, the averaged speed parameters were relatively similar as the parameters obtained from the averaged data; β was 13.42 (*SD* = 9.53) and δ was 451 ms (*SD* = 0.05). The best fitting parameter values obtained from the averaged data, and the averaged parameter values obtained from fitting individuals' data were in accordance with each other Table 5.

As in previous experiments, systematic underestimation of numerosity was observed when the probability of responding "greater than 50" reached its maximum starting at the 500-ms lag condition. When the sampling probability in the binomial distribution was .45 and the trials with the number of dots exceeding 50, the average number of dots presented was 52.99 (*SD* = 2.08). Participants responded "greater than 50" for 62% (*SD* = 0.13) of the trials. When the sampling probability from the binomial distribution increased to .55 and the average of the actual number size was 56.55 (*SD* = 3.83) at the asymptotic lag conditions, the percentage of responding "greater than 50" increased to 77% (*SD* =0.06). Finally, the highest proportion of responding "greater than 50," 96% (*SD* = 0.40), was observed in the .70 dot condition, which had the highest mean of actual number of dots, 69.83 (*SD* = 4.64). These results indicate a systematic underestimation especially for the trials that the probability of sampling number size was less than .70. The response rate at the asymptote was less than 1, suggesting even though the number of dots was greater than 50, people did not perceive the size of the patch as greater than 50. However, when these results were compared with the results presented in the previous experiments, the probability of responding "greater than 50" was the greatest in the current experiment, suggesting a limited underestimation.

To summarize, when the response-key assignment was reversed and resulted in an incongruent scale, the SAT functions showed that the accumulation of evidence slowed compared with congruent response-key assignments in the previous experiments. However, apart from this change in the parameters, prior response bias observed in all experiments were comparable. There is a tendency to respond "greater than 50" even when the response-key assignment was reversed. This finding suggests this bias towards "greater than 50" response is independent of where the response keys are located or whether the spatial location of the keys are congruent with the mental number line. In either case, the bias to respond "greater than 50" is observed.

Table 5 Parameter values of the best fitting exponential function to individual data in Experiment 4

Participants	λ_1 30	λ_1 45	λ_1 55	λ_1 70	λ_2	β	δ	Deviance	R^2
1	.022	.412	.768	.950	.706	5.04	0.452	−37.939	.948
2	.001	.264	.682	.990	.990	20.29	0.408	−37.414	.947
3	.014	.423	.745	.961	.990	6.09	0.458	−36.825	.772
4	.001	.241	.703	.947	.770	8.37	0.385	−39.135	.901
5	.013	.294	.696	.972	.839	9.85	0.505	−38.106	.820
6	.042	.341	.800	.980	.977	30.00	0.495	−39.313	.944
7	.052	.423	.761	.982	.672	8.57	0.502	−38.480	.913
8	.001	.218	.629	.940	.954	26.38	0.507	−41.244	.915
9	.001	.335	.702	.983	.630	6.25	0.349	−41.190	.970

Note. Parameter values are obtained from the fits to each participant data. λ_1 is the asymptotic probability of accepting the number of dots to be greater than 50 for each dot condition, 30, 45, 55, and 70, respectively. λ_2 is the probability of accepting the number of dots to be greater than 50 at chance and indicates the bias at early retrieval. β is the rate of accumulation towards asymptote, and δ is the time point at which the probability of accepting the number of dots start differing from bias (λ_2)

General discussion

The current study examined the time course of numerosity perception by employing the response deadline SAT procedure in a two-choice decision task. In the numerosity perception task that we administered, participants were presented with dot patches, of which the dot numbers were sampled from four probability distributions. Four experiments followed the same routine. Participants were presented with a set of dots, later asked to select the option whether the number of dots exceeded 50 or not, except Experiment 2, in which they were asked whether the number fell below 50. Immediately after, they were given feedback on the actual number of dots, except in Experiment 3, dot number feedback was withheld. Finally, in Experiment 4, the response-key assignment was reversed, such that the "greater than 50" response was mapped on the left-hand side of the keyboard and the "lower than 50" response was mapped on the right-hand side. The most prominent finding from all experiments was the response bias observed at early lags, which indicates a tendency towards accepting "greater than 50" response prior to evidence accrual.

The results obtained from all experiments showed an early bias towards responding yes, independent of the comparison question (*greater* or *less* than 50) or whether receiving feedback to calibrate the mental number line or the congruence of the key placement with the mental number line. For example, the comparison between the parameter values obtained from Experiment 1 and the parameter values obtained from Experiment 2 indicate that the participants were agnostic about whether the question was asked as *greate*r or *less* than 50. Similarly, withholding feedback in Experiment 3 did not reveal a strong difference in the early bias parameter, suggesting that the participants were still biased towards "greater than 50" response at early processing when their number line was not calibrated. Finally, which key was used to respond "yes" and "no" did not affect the strong bias towards the "yes" response even when the response-key mapping was incongruent. These results indicate a strong early bias towards "greater" responses.

The effects of time on numerosity estimation and discrimination have been investigated by Cheyette and Piantadosi (2019) by using an eye tracker, which contributed to the explanation of the ANS estimation, suggesting that instead of parallel processing, a serial accumulation process may be the underlying mechanism, since they showed that the estimates increased as visual fixations increased. In their study, participants estimated numerosity for 100, 333, 1,000 and 3,000-ms time lags and feedback was not provided. Their results indicated an underestimation for all time conditions, but this underestimation decreased for longer time lags. This was consistent with their expectation of quantity accumulation over time, which was only due to the accumulation of foveated dots and not due to time. When all the dots were foveated, the finding of underestimation was eliminated, which led them to infer that numerosity estimation is a measure of serial accumulation of foveated dots. They also replicated their findings for numerosity estimation task for their numerosity discrimination task. Their analysis showed that for short durations, foveated dots contributed almost twice as much as the peripheral dots, which provided estimation to be possible for very short times, but for longer durations, the contribution of the foveated dots increased, and for the longest durations, it was only the foveated dots that were used for numerosity estimation. When we compare our results of temporal effect on numerosity perception with their results, there is a difference in the pattern such that we come across a response bias to overestimate in the short durations, while they observed more underestimation for shorter times, which decreased as duration increased.

The underlying mechanism of the early bias can be built upon how large and small numbers are in fact processed in a unified account (Cheyette & Piantadosi, 2020) contrary to being processed in different systems (see Feigenson et al., 2004, for a review). It has been suggested that in numerosity perception there are two separate systems. One processes large numbers by utilizing an estimation of summary representations and mapping to ANS (Dehaene et al., 1999), while the other processes small numbers through subitizing resulting in a perfect and fast performance (Trick & Pylyshyn, 1994). Because larger number representations require some form of computation, the processing becomes error prone and slow. Recently, Cheyette and Piantadosi (2020) proposed a unified system suggesting that small and larger numbers can be processed with the same mechanism that relies on different processing capacities based on expectations or prior encounters of the stimuli. Thus, limited resources will be allocated differently to process small and large numbers. According to this unified system approach, it is possible that the participants had different expectancies on the large and relatively smaller number of patches in the current study. Specifically, the increase in the dot numbers were selected to be additive from 30 to 45 and from 55 to 70. That might have resulted in an unbalanced perceptual discrimination considering Weber's Law. As Weber's law also indicates, the difference from the 50-point criterion value is harder to discriminate from 70-point patches (70/50 = 1.40), while 30-point patches are easier to discriminate (50/30 = 1.67). Thus, expectancy that the smaller number of dots are processed easier might have caused the participants to initially respond with the "greater" choice, and as time passed in later deadline conditions, they corrected their response.[2] Future studies can consider sampling dots from different distributions that follow Weber's Law, which would increase multiplicatively but not additively, and test whether

[2] We would like to thank the reviewers for pointing out these possible explanations.

a positive bias is still observed under the multiplicatively increasing dot conditions.

In the present study, the numerosity perception of only relatively large numbers were tested leaving the smaller quantities out of empirical testing. The number of dots ranged between 15 and 85 across all four experiments, which resulted in an early bias to respond "greater than 50." However, the reported bias could be limited to the number of dots used in this study. Further studies using smaller ranges of dots (e.g., 1–15) are required to generalize this bias to numerosity perception of any range of quantities.

Results from Experiment 3 showed a systematic underestimation in numerosity judgment in the absence of feedback when compared with results from Experiment 1. The asymptotic response rates of the .70 condition across the two experiments showed a decrease when the participants were not given any feedback. Specifically, all the trials in the .70 dot condition contained dot patches that exceed 50 dots, but the asymptotic response rate values were lower than 1 in Experiment 3 because participants did not have a chance to calibrate their mental number representation (as shown by Izard & Dehaene, 2008; Krueger, 1984; Price et al., 2014). Similarly, for the .55 condition, only almost 30% of trials had dot patches lower than 50, but the asymptotic response rate was .50, half of all the trials in .50 dot condition in Experiment 3, while this value was .67 in Experiment 1. Although the usual finding of underestimation was still present even for Experiments 1, as the proportion of responding "greater than 50" was not 1 as expected, the magnitude of underestimation was not as high as the one observed in Experiment 3. These findings further supported the well-established benefit of feedback on numerosity judgements.

Finally, the results from Experiment 4 revealed a slowing in numerosity processing due to incongruent response-key assignment. The asymptote parameters, which indicate the total response rate that can be reached with enough processing time, did not change as a function of response-key mapping. However, the parameter values obtained for the speed parameters, especially the evidence accumulation rate parameter showed that 1% change in the response rate slowed for Experiment 4, compared with that of in other experiments. These results support the idea that congruency in multisensory associations and responses are important even at the perceptual level (Kim, Seitz, & Shams, 2008).

Ratcliff (2006) proposed a theoretical explanation for SAT functions obtained from the response deadline procedure and he applied a numerosity perception task to test the model. The model was based on the diffusion model (Ratcliff, 1978), a dynamic variant of the signal detection theory. In the diffusion model, evidence is sampled sequentially, and compared with a criterion at each time point until sampling terminates. Two responses (yes and no) are represented as boundaries and once the accumulated evidence reaches one boundary, the process is terminated. The speed of the process is measured by the amount of time that it is required to drift towards either boundary. Once the boundary is reached, the response is produced as either correct or incorrect, depending on the boundary. The bias towards either response is measured by the point at which the drift process starts—specifically, where that point lies between the two response boundaries. For example, a starting point that is proximate to the "yes" boundary would result in a tendency towards the "yes" response, as in the findings observed in the current experiment. Ratcliff further advanced the model to account for the data obtained from the response deadline procedure by including a second state. Specifically, a response is produced if evidence accumulation reaches either boundary prior to signal onset as in standard reaction time measures. In that case, the reported response is the boundary that has been reached before the signal is provided. The second state takes place if the boundary has not been reached before the signal onset. In that case, either partially accumulated evidence is used to generate a response, or the response is guessed due to the absence of partial information. The data obtained from the current study can be fit with this model and compared with the fits of the original diffusion model. Later, if the fit statistics of the partial model is preferred over the original model, we can claim that numerosity perception is automatic, such as the boundaries are reached even before the signal is presented. This claim requires further investigations.

In all experiments, the number of dots were generated randomly from the binomial distribution with means of 30, 45, 55, and 70. In the model that we fit, we kept the rate parameter constant across all dot conditions. However, the variance of the binomial distribution is in fact different for the easier conditions (30 and 70 dots) compared with the harder conditions (45 and 55 dots). Specifically, to discriminate 30 (or 70) dots from the 50 dots criterion is much easier compared with discriminating 45 (or 55) dots from the 50 dots criterion. The easier and harder dot conditions have different variance values of the binomial distribution, which are 21 (100 × 0.3 × 0.7) for the easier condition and 24 (100 × 0.45 × 0.55) for the harder condition. These two different variances in the dot distributions might have caused a difference in the rate of information accumulation across dot conditions.[3] To test this possibility, we also ran a set of models where the rate parameter differed across the 30/70 and 45/55 dot conditions. Best fitting model parameters were presented in the supplementary materials along with the model indices (AIC and BIC). The results revealed that the reduced model ($4\lambda_1$-λ_2-β-δ) is more parsimonious compared with the two-rate parameter model ($4\lambda_1$-λ_2-2β-δ), and a single β value is sufficient for explaining the SAT function.

[3] We would like to thank Reviewer 2 for pointing out this possibility.

The ability to estimate numbers using numerosity perception tasks, as in other perceptual tasks such as loudness, brightness, and distance, requires mapping from subjective internal states to formal measurements. To process all these perceptual stimuli, observing a systematic underestimation (e.g., Stevens, 1959, 1966; Teghtsoonian & Teghtsoonian, 1978) suggests a similar underestimation in other sensory mediums. That is, our perceptual system can be generalized to all sensory inputs such that they all build on a form of logarithmic mental representation for perceptual stimuli.

Conclusion

In the present study, we investigated how the time course of numerosity perception was affected by receiving feedback to calibrate the mental number line representation. We employed the response deadline procedure in a two-choice decision task, where participants were asked to judge whether the number of dots exceeded 50 (Experiments 1, 3, 4) after being presented with a patch of dots that ranged between 15 and 85 and sampled with four different probability conditions. In three experiments (Experiments 1, 2, 4), participants were given feedback on the actual number of dots in each patch, but were withheld in one experiment (Experiment 3). The results revealed that when feedback was provided, a decrease in underestimation in late processing occurred. The novel finding in this study, demonstrated by all experiments, is a strong bias towards accepting the number of dots as greater than 50 at early judgments. Whereas, the results imply that being provided with feedback allows for a calibration of mental number line representation later at processing. Additionally, being tested on a large number such as 50 resulted in overestimation at early processing. A possible explanation for this finding could be an overcorrection to avoid incorrect responses for patches with a dot number greater than 50. Such patches would present greater difficulty of comparison to patches of 70 in contrast to the relative ease of comparison to patches of 30. To test this claim, future studies can be conducted on a smaller range of dot numbers and a smaller criterion number for comparison.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13414-021-02365-3.

Author note This work has been supported by Middle East Technical University Scientific Research Projects Coordination Unit under grant number BAP-08-11-2017-036.

We would like to thank Sinem Aytaç, Hatice Dedetaş, Zeynep Erbaş and Sema Betül Türk for their assistance in data collection.

Parts of this work were presented in the International Meeting of the Psychonomic Society in 2018 and the 51st Annual Meeting of the Society for Mathematical Psychology.

Data link (https://osf.io/u3yz5/).

References

Anobile, G., Cicchini, G. M., & Burr, D. C. (2016). Number as a primary perceptual attribute: A review. *Perception*, *45*(1/2), 5–31. https://doi.org/10.1177/0301006615602599

Bevan, W., & Turner, E. D. (1964). Assimilation and contrast in the estimation of number. *Journal of Experimental Psychology*, *67*(5), 458–462. https://doi.org/10.1037/h0041141

Burr, D. C., Anobile, G., & Arrighi, R. (2018). Psychophysical evidence for the number sense. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1740), 20170045. https://doi.org/10.1098/rstb.2017.0045

Castronovo, J., & Seron, X. (2007). Numerical estimation in blind subjects: Evidence of the impact of blindness and its following experience. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(5), 1089. https://doi.org/10.1037/0096-1523.33.5.1089

Cheyette, S. J., & Piantadosi, S. T. (2019). A primarily serial, foveal accumulator underlies approximate numerical estimation. *Proceedings of the National Academy of Sciences of the United States of America, 116,* 17729–17734.

Cheyette, S. J., & Piantadosi, S. T. (2020). A unified account of numerosity perception. *Nature Human Behaviour*, *4*(12), 1265–1272.

Cordes, S., Gelman, R., Gallistel, C. R., & Whalen, J. (2001). Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin & Review*, *8*(4), 698–707. https://doi.org/10.3758/BF03196206

Crollen, V., & Seron, X. (2012). Over-estimation in numerosity estimation tasks: More than an attentional bias? *Acta Psychologica*, *140*(3), 246–251. https://doi.org/10.1016/j.actpsy.2012.05.003

Crollen, V., Castronovo, J., & Seron, X. (2011). Under- and over-estimation: A bi-directional mapping process between symbolic and non-symbolic representations of number? *Experimental Psychology*, *58*, 39–49. https://doi.org/10.1027/1618-3169/a000064

Crollen, V., Grade, S., Pesenti, M., & Dormal, V. (2013). A common metric magnitude system for the perception and production of numerosity, length, and duration. *Frontiers in Psychology*, *4*, 449. https://doi.org/10.3389/fpsyg.2013.00449

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, *44*(1/2), 1–42. https://doi.org/10.1016/0010-0277(92)90049-N

Dehaene S, Spelke E, Pinel P, Stanescu R, Tsivkin S. Sources of mathematical thinking: behavioral and brain-imaging evidence. Science. 1999 May 7;284(5416):970-4

Dehaene, S. (2003). The neural basis of the Weber–Fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences*, *7*, 145147. https://doi.org/10.1016/S1364-6613(03)00055-X

Dehaene, S. (2011). *The number sense: How the mind creates mathematics* (2nd ed.). Oxford University Press.

Dehaene, S., Izard, V., Spelke, E.S., & Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science*, 320(5880), 1217–1220. https://doi.org/10.1126/science.1156540

DeWind, N. K., Park, J., Woldorff, M. G., & Brannon, E. M. (2019). Numerical encoding in early visual cortex. *Cortex*, *114*, 76–89. https://doi.org/10.1016/j.cortex.2018.03.027

Dietrich, J. F., Huber, S., & Nuerk, H. C. (2015). Methodological aspects to be considered when measuring the approximate number system

(ANS)–A research review. *Frontiers in Psychology*, *6*, 295. https://doi.org/10.3389/fpsyg.2015.00295

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences, 8*(7), 307–314.

Fornaciai, M., Cicchini, G. M., & Burr, D. C. (2016). Adaptation to number operates on perceived rather than physical numerosity. *Cognition*, *151*, 63–67. https://doi.org/10.1016/j.cognition.2016.03.006

Guillaume, M., & Gevers, W. (2016). Assessing the Approximate Number System: No relation between numerical comparison and estimation tasks. *Psychological Research*, *80*(2), 248–258. https://doi.org/10.1007/s00426-015-0657-x

Guillaume, M., & Van Rinsveld, A. (2018). Comparing numerical comparison tasks: A meta-analysis of the variability of the weber fraction relative to the generation algorithm. *Frontiers in Psychology*, *9*, 1694. https://doi.org/10.3389/fpsyg.2018.01694

Indow, T., & Ida, M. (1977). Scaling of dot numerosity. *Perception & Psychophysics*, *22*(3), 265–276. https://doi.org/10.3758/BF03199689

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, *106*, 1221–1247. https://doi.org/10.1016/j.cognition.2007.06.004

Kılıç, A., & Öztekin, I. (2014). Retrieval dynamics of the strength based mirror effect in recognition memory. *Journal of Memory and Language*, *76*, 158–173. https://doi.org/10.1016/j.jml.2014.06.009

Kim, R. S., Seitz, A. R., & Shams, L. (2008). Benefits of stimulus congruency for multisensory facilitation of visual learning. PLoS One, 3 (1), e1532

Krueger, L. E. (1972). Perceived numerosity. *Perception, & Psychophysics*, *11*(1), 5–9. https://doi.org/10.3758/BF03212674

Krueger, L. E. (1982). Single judgments of numerosity. *Perception & Psychophysics*, *31*, 175–182. https://doi.org/10.3758/BF03206218

Krueger, L. E. (1984). Perceived numerosity: A comparison of magnitude production, magnitude estimation, and discrimination judgments. *Perception & Psychophysics*, *35*(6), 536–542. https://doi.org/10.3758/BF03205949

Mejias, S., & Schiltz, C. (2013). Estimation abilities of large numerosities in kindergartners. *Frontiers in Psychology*, *4*, 518. https://doi.org/10.3389/fpsyg.2013.00518

Mundy, E., & Gilmore, C. K. (2009). Children's mapping between symbolic and nonsymbolic representations of number. *Journal of Experimental Child Psychology*, *103*, 490–502. https://doi.org/10.1016/j.jecp.2009.02.003

Nieder, A. (2016). The neuronal code for number. *Nature Reviews Neuroscience*, *17*(6), 366–382. https://doi.org/10.1038/nrn.2016.40

Norris, J. E., & Castronovo, J. (2016). Dot display affects approximate number system acuity and relationships with mathematical achievement and inhibitory control. *PLOS ONE*, *11*(5). https://doi.org/10.1371/journal.pone.0155543

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman. E., & Lindeløv, J. K. (2019) PsychoPy2: Experiments in behavior made easy. *Behavioral Research Methods*, 51, 195–203. https://doi.org/10.3758/s13428-018-01193-y

Piazza, M. (2010). Neurocognitive start-up tools for symbolic number representations. *Trends in Cognitive Sciences*, *14*, 542–551. https://doi.org/10.1016/j.tics.2010.09.008

Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron*, *53*(2), 293–305. https://doi.org/10.1016/j.neuron.2006.11.022

Price, J., Clement, L. M., & Wright, B. J. (2014). The role of feedback and dot presentation format in younger and older adults' number estimation. *Aging, Neuropsychology, and Cognition*, *21*(1), 68–98. https://doi.org/10.1080/13825585.2013.786015

R Core Team. (2019). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Ratcliff, R. (1978). A theory of memory retrieval. Psychological Review, 85(2), 59

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology*, *53*(3), 195–237. https://doi.org/10.1016/j.cogpsych.2005.10.002

Ratcliff, R., & McKoon, G. (2018). Modeling numerosity representation with an integrated diffusion model. *Psychological Review, 125*(2), 183–217. https://doi.org/10.1037/rev0000085

Reed, A. V. (1973). Speed-accuracy trade-off in recognition memory. Science, 181(4099), 574-576

Reinert, R. M., Hartmann, M., Huber, S., & Moeller, K. (2019). Unbounded number line estimation as a measure of numerical estimation. *PLOS ONE*, *14*(3). https://doi.org/10.1371/journal.pone.0213102

Stevens S. S. (1959). Cross-modality validation of subjective scales for loudness, vibration, and electric shock. Journal of Experimental Psychology, 57(4), 201–209

Stevens, S. S. (1966). Matching functions between loudness and ten other continua1. *Perception & Psychophysics*, *1*(1), 5–8. https://doi.org/10.3758/BF03207813

Teghtsoonian, R., & Teghtsoonian, M. (1978). Range and regression effects in magnitude scaling. *Perception & Psychophysics*, *24*(4), 305–314. https://doi.org/10.3758/BF03204247

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review*, *101*(1), 80.

Van den Berg, R., Lindskog, M., Poom, L., & Winman, A. (2017). Recent is more: A negative time-order effect in nonsymbolic numerical judgment. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(6), 1084–1097. https://doi.org/10.1037/xhp0000387

Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: A neural model. *Journal of cognitive neuroscience*, *16*(9), 1493–1504. https://doi.org/10.1162/0898929042568497

Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Non-verbal counting in humans: The psychophysics of number representation. *Psychological Science*, *10*, 130–137. https://doi.org/10.1111/1467-9280.00120

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), B1–B11. https://doi.org/10.1016/S0010-0277(99)00066-9

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