Numerical error monitoring

Yalçın Akın Duyan¹ · Fuat Balci¹

Published online: 6 July 2018
© Psychonomic Society, Inc. 2018

Abstract

Error monitoring has recently been discovered to have informationally rich foundations in the timing domain. Based on the common properties of magnitude-based representations, we hypothesized that judgments on the direction and the magnitude of errors would also reflect their objective counterparts in the numerosity domain. In two experiments, we presented fast sequences of “beeps” with random interstimulus intervals and asked participants to stop the sequence when they thought the target count (7, 11, or 19) had been reached. Participants then judged how close to the target they stopped the sequence, and whether their response undershot or overshot the target. Individual linear regression fits as well as the linear mixed model with a fixed effect of reproduced numerosity on confidence ratings, and participants as independent random effects on the intercept and the slope, revealed significant positive slopes for all the target numerosities. Our results suggest that humans can keep track of the direction and degree of errors in the estimation of discrete quantities, pointing at a numerical-error-monitoring ability.

Keywords Error monitoring · Number estimation · Metacognition · Magnitude estimation

Every decision is accompanied by a subjective degree of uncertainty regarding the decision’s accuracy. To address this, error monitoring (awareness of errors without feedback) and performance monitoring have been typically studied in the two-alternative forced-choice paradigm (2AFC), in which participants decide which of the two alternatives sensory evidence favors. Results of these studies showed that confidence ratings closely track the decision accuracy performance (e.g., Fleming, Dolan, & Frith, 2012).

On the other hand, many of our daily actions rely largely on approximate quantity estimates such as time intervals, numerosities, distances, and making simple decisions based on these quantitative estimates. These can be exemplified by our routine judgments regarding the earliness and lateness in meeting schedules (e.g., duration of traffic signals), counts of occurrences (e.g., number of junctions crossed) or distance

traveled (e.g., judging if you missed an exit at a known distance), and, for instance, deciding to take a given exit or not based on such estimates.

An important feature of these scenarios is that each magnitude estimation is subject to internal sources of uncertainty leading to substantial trial-to-trial variability in behavior and characterizing every magnitude-based decision as decisions made under uncertainty. A recent line of research (e.g., Çavdaroğlu, Zeki, & Balci, 2014; Çavdaroğlu & Balci, 2016) has addressed the importance of the abovementioned subjective timing and counting uncertainty for decision-making by formulating the dependence of reward-rate-maximizing decisions on the level of uncertainty. The results of these studies showed that humans and rodents can nearly optimize their quantitative decisions by integrating the level of their subjective timing and numerical uncertainty into these decisions (Çavdaroğlu & Balci, 2016; Çavdaroğlu, Zeki, & Balci, 2014). These observations suggest that internal uncertainty about magnitude estimates can be adaptively integrated into the decision process as a biasing signal.

To address the uncharted possibility, Akdoğan and Balci (2017) examined if humans could accurately guess the direction and magnitude of errors in their trial-to-trial estimates of time intervals. In a series of experiments, they asked participants to reproduce target durations as accurately as possible. Participants’ judgments provided after each trial regarding

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13423-018-1506-x>) contains supplementary material, which is available to authorized users.

✉ Fuat Balci
fbalci@ku.edu.tr

¹ Department of Psychology & Research Center for Translational Medicine, Koç University, Rumelifeneri yolu, Sarıyer, 34450 Istanbul, Turkey

confidence about the accuracy of their estimates, and whether this estimate was longer or shorter than the target interval, closely matched the participants' actual timing performance. This suggests that the information about the magnitude and the direction of the timing errors become available at some point before the confidence judgment is made. These results show that performance-monitoring ability extends to the estimation of continuous magnitudes (i.e., durations).

As for time intervals, it has also been suggested that nonsymbolic numerosities are represented in a continuous fashion by the approximate number system, which is subject to uncertainty (Dehaene, 2011; Gallistel & Gelman, 1992). In line with this view, the discriminability of two different numerical quantities have been shown to be determined by their ratio (i.e., Weber's law; Cordes, Gallistel, Gelman, & Latham, 2007), even when participants nonverbally count rapid arrhythmic flashes (Allik & Tuulmets, 1993; Cordes, Gelman, & Gallistel, 2001; Whalen, Gallistel, & Gelman, 1999).

A number of previous findings also point at performance-monitoring abilities in the numerical domain. For instance, Gelman and Gallistel (1978) report that children often restart counting when they skip an object in the set or a number word. Another study shows that when math problems are framed nonsymbolically using two magic cups that add different numbers of items to the original set, children can infer from which of the two cups the new items came from (i.e., an addend-unknown problem; Kibbe & Feigenson, 2015). In both of these situations, children demonstrated the basic ability to compare the expected outcome with the actual one and use that error information. Moreover, Vo, Li, Kornell, Pouget, and Cantlon (2014) demonstrated performance-monitoring skills of children in numerical estimates; children made high-risk bets after correct decisions and on easier trials.

Importantly, a large amount of empirical and theoretical work has also established behavioral and neural similarities between the processing of different quantitative domains (e.g., Walsh, 2003). Based on these convergent lines of evidence, we hypothesized that quantitative error-monitoring ability would apply to numerosity estimates. To test this hypothesis, we used a numerical version of the task used in Akdoğan and Balcı (2017) with different targets over two experiments. We presented fast sequences of beeps with random interstimulus intervals and asked participants to stop the sequence when they judged the number of beeps had reached the target number. Subsequently, we collected confidence ratings on how close participants thought their estimate captured the target number and asked whether their response undershot or overshot the target.

Experiment 1

Method

Participants

Twenty-nine undergraduate students from Koç University participated in the experiment for course credit. All participants gave informed consent. The study was approved by the local ethics committee at Koç University.

Apparatus

Participants were tested in a dimly lit room, seated approximately 50 cm from a 22-inch monitor. Experiments were controlled via MATLAB (MathWorks, Natick, MA) using the Psychophysics Toolbox (Brainard, 1997) on an iMac. The experimental program and raw data collected can be accessed at the Open Science Framework (osf.io/re48n).

Procedure

We presented the sequences of beep sounds (444 Hz, 60ms) with random interstimulus intervals varying between 300 and 600 ms (uniformly distributed). Participants were asked to stop the sequence by pressing the space key when they thought the beep count reached the target number (11 or 19). Participants were prompted to provide a confidence rating by pressing the Q (low), W (medium), or E (high) keys to indicate (100 ms after their initial response) how close they thought their estimate captured the target number in that trial. They were then immediately asked whether they undershot or overshot the target by pressing the A or D keys, respectively. The intertrial interval (ITI) varied between 1.5 and 2.5 s (uniformly distributed). Participants were tested over four 13-minute blocks.

Approach to analysis

For each participant, we recorded the number of beeps before stopping the stimulus sequence for each confidence-rating pair: Under(U)-Low(L), Under(U)-Medium(M), Under(U)-High(H), Over(O)-High(H), Over(O)-Medium(M), Over(O)-Low(L). Participants' confidence judgments reflected the amount of deviation from the target, regardless of the direction of their errors. If there exists an error monitoring mechanism for numerosity judgments, trials with high confidence ratings should be closer to the actual target. Hence, the logical ordering of the confidence-rating pairs is from UL to OL; the mean reproduction should be the lowest for UL and the highest for OL. For each participant and target numerosity, we regressed six response categories (UL to UH) that reflected confidence and directionality of error judgment pairs on the estimated

numerosities. Consequently, slopes significantly higher than zero would indicate an ability to monitor the degree and the directionality of errors in the numerosity estimates. To analyze the overall effect, we also used a linear mixed model with a fixed effect of reproduced numerosity on confidence and included participants as independent random effects on the intercept and the slope.

Our hypothesis was that judgments on the direction and the magnitude of errors would veridically reflect the nature of the actual estimation errors in the numerosity domain. Because we are primarily interested in the magnitude of errors and their relationship with subjective confidence judgments in estimates, on-target trials were excluded (26.9% and 14.73% of the trials for T11 and T19, respectively; see Supplemental Online Material Fig. S1). We excluded trials where the number of beeps were three mean absolute deviations (MAD) above or below each participant's mean (3.8% of all trials) since they could bias the results in favor of our hypothesis. The main outcomes of interest were whether the participants' signed confidence judgments were in line with their objective performance and the ratio of participants that individually exhibited significant quantitative error monitoring ability (positive slopes for reproduced numerosity as a function of signed confidence). Perfect error monitoring performance would provide a slope close to 1.

Results

Fits for the linear mixed-effects models were done separately for each target number with the `fitlme` function with default settings in MATLAB.

For both target numerosities, the main effect of reproduced numerosities on confidence was significant ($\beta = .217$, $SE = .026$, $p < .001$, $R^2 = .26$ for T11; $\beta = .109$, $SE = .01$, $p < .001$, $R^2 = .189$ for T19), indicating that confidence judgments in general followed objective performance (see Table 1). The mean standardized slopes that relate the signed numerical errors to the confidence ratings for each participant were .334 (CI [.248, .42]) for T11 and .264 (CI [.2, .332]) for T19 (see Fig.

1, top-panel for the depiction of relationship between these variables). These slopes were significantly higher than zero, $t(28) = 8$, $p < .001$, $d = 1.484$ for T11, and $t(28) = 7.915$, $p < .001$, $d = 1.467$ for T19. As a result of linear regression analysis conducted for each participant, 82.76% and 62.1% of the participants had a significant positive slope for T11 and T19, respectively (see Supplemental Online Material Table S2).

As another test of numerical error monitoring, we compared mean confidence ratings when participants' responses were on target versus off target. For both target numerosities, confidence ratings were significantly higher for on-target than for off-target responses, $t(27) = 6.326$, $p < .001$, CI [.183, .358] for T11 and $t(27) = 3.114$, $p = .004$, CI [.062, .302] for T19, respectively. One participant did not have any on-target responses for either target.

As the beeps were presented sequentially, time and number were highly correlated. To elucidate whether participants relied on time rather than numerosity, we fitted hierarchical regression to each participant's data by first entering the response times (RTs) and then the number of beeps and vice versa. Mean R^2 change was significantly higher when reproduced numbers were entered into the model secondarily for T11 ($M = .047$, CI [.013, .082]) than when the RTs were entered into the model second ($M = .005$, CI [.002, .008]). $t(28) = 2.49$, $p = .019$, $d = .462$, CI [.008, .077]). However, R^2 changes for the two different hierarchical models were comparable for target T19, $t(28) = 1.041$, $p = .31$, $BF = 3.095$.

When we regressed the confidence categories on the estimated numerosities and RTs separately, for T11, the mean standardized slopes for the estimated numerosities ($M = .322$, CI [.231, .412]) were significantly higher than the mean standardized slopes for RTs ($M = .28$, CI [.205, .355]), $t(28) = 2.587$, $p = .015$, CI [.009, .075], $d = .481$. However, the slopes obtained for RTs and numerosities were comparable for T19, $t(28) = 1.072$, $p = .292$, $BF = 2.963$. We also compared the slopes for numerosity estimates and RTs when both predictors were entered in the regression analysis. In T11, mean slope for numerosity estimates ($M = .367$, CI [.255, .458]) was significantly higher than mean slope for RTs ($M = -.030$, CI [-.126, .065]), $t(28) = 4.494$, $p < .001$, CI [.211, .563], $d = .835$. However, in T19, mean slopes were similar, $t(29) = 1.264$, $p = .217$, $BF = 2.465$.

Table 1 Summary results of the linear mixed-effects models (full model table and diagnostic plots are provided in Supplemental Online Material Figs. S3.1–S3.4 and Tables S3.1–S3.4)

	Estimate	Standard error	df	t value	p	CI	AIC	Log likelihood
Experiment 1								
T11	.217	.026	3,452	8.443	<.001	.167–.267	11,904	−5947
T19	.109	.01	2,915	11.157	<.001	.09–.129	10,683	−5336.5
Experiment 2								
T7	.483	.08	1,615	6.017	<.001	.326–.641	5,525.7	−2,757.9
T11	.323	.063	1,809	5.131	<.001	.199–.446	6,361.5	−3,175.7

Fig. 1. Relationship between numerosity estimates and signed confidence ratings. Average signed confidence ratings (−3: UL, −2: UM, −1: UH, 1: OH, 2: OM, 3: OL) as a function of z -score transformed numerosity estimates (including all responses) for Experiment 1 (top panel) and Experiment 2 (bottom panel). We calculated z scores for all reproduced numbers separately for each participant. We then computed the mean confidence for each z -score

transformed numerosity separately for each participant. Colored markers indicate a participant's mean confidence for a given z score. Different colors correspond to different participants. Fitted lines show the robust regression fits (with Huber weights) to the data for depiction purposes. T7 = Target 7, T11 = Target 11, T19 = Target 19. Note that T11 was included both in Experiment 1 and Experiment 2. (Color figure online)

Note that these analyses were done even though participants were instructed to rely on numerosity only. In fact, in timing tasks participants tend to rely on (prioritize) counts (e.g., Fraisse, 1963) presumably because counting is a useful strategy for reducing variance in timed responses (Grondin, Meilleur-Wells, & Lachance, 1999). Thus, participants are typically instructed not to count in timing tasks (e.g., Akdoğan & Balci, 2017), which has been shown to be an effective method in and of itself (Rattat & Droit-Volet, 2012).

Cordes, Gelman, Gallistel, and Whalen (2001) reports that scalar variability is violated when participants count their key presses out loud in a number reproduction task, with coefficient of variation (CV) decreasing as the inverse square root of the target number. Consequently, we calculated the CV for each participant's numerosity judgments and compared the resulting CVs between targets. The results showed that participants' CVs for both targets were comparable, $t(28) = .913$, $p = .367$, CI [.012, .032], $BF = 3.45$. Finally, numerical CVs were lower than RT CVs for both targets, $t(29) = 8.624$, $p < .001$, CI [.009, .015], $d = 1.601$ for T11, and $t(29) = 5.672$, $p < .001$, CI [.005, .010], $d = 1.053$ for T19, indicating that participants used numerical information rather than relying on time (see Supplemental Online Material Table S5.1).

Experiment 2

Method

Participants

Fifteen undergraduate students from Koç University participated in the experiment for course credit. All participants gave informed consent. The study was approved by the local ethics committee at Koç University.

Procedure

All procedures in Experiment 2 were identical to those in Experiment 1, except that the numerical targets were 7 (T7) and 11 (T11).

Results

Data exclusion criteria were identical to those in Experiment 1. On-target responses were excluded (44.54% and 28.36% trials for T7 and T11, respectively). Trials where the

reproduced number was three MADs above or below each participant's mean were also excluded (4.87% of trials).

To test the overall effect of numerical reproduction performance on confidence judgments, we used the same linear mixed-effect model with the reproduced number as the linear predictor and participant as a random effect on the slope and the intercept. For both targets, the main effect of the reproduced numerosities on confidence was significant ($\beta = .483$, $SE = .08$, $p < .001$, $R^2 = .262$ for T7; $\beta = .323$, $SE = .063$, $p < .001$, $R^2 = .178$ for T11; see Table 1).

The mean standardized slopes were .412 (CI [.309, .516]) and .354 (CI [.253, .455]) for T7 and T11, respectively (see Fig. 1, bottom panel for the depiction of relationship between the variables). The comparisons of these slopes to zero revealed significant differences, $t(12) = 8.685$, $p < .001$, $d = 2.07$ for T7; $t(14) = 7.531$, $p < .001$, $d = 1.967$ for T11. As a result of the linear regression analyses conducted separately for each participant, 86.67% and 80% of the participants had a significant positive slope for T7 and T11, respectively (see Supplemental Online Material Table S2). As another test of numerical error monitoring, we compared the confidence ratings when participants' responses were on target versus off target. For both target numerosities, confidence ratings were significantly higher for on-target than off-target responses, $t(14) = 4.179$, $p < .001$, CI [.164, .511], $d = 1.134$, and $t(14) = 2.649$, $p = .02$, CI [.032, .308], $d = .51$ for T7 and T11, respectively.

For T7, R^2 changes were significantly higher when the reproduced numerosity were entered secondarily into the model ($M = .052$, CI [.015, .081]) than when the RTs were entered second ($M = .007$, CI [.001, .011]), $t(12) = 2.388$, $p = .034$, CI [.004, .076], $d = .662$. For T11 too, R^2 changes were significantly higher when we entered the reproduced numerosity second ($M = .051$, CI [.021, .082]) than when we entered the RTs second ($M = .01$, CI [.004, .016]), $t(14) = 3.004$, CI [.011, .07], $p = .01$, $d = .776$. Moreover, when we regressed confidence categories separately on the reproduced numerosities and RTs the mean slopes for the reproduced numerosities ($M = .412$, CI [.309, .516]) were significantly higher than the mean slopes for the RTs ($M = .364$, CI [.283, .444]) in T7, $t(12) = 2.329$, $p = .038$, CI [.003, .094] $d = 0.317$. In T11, the mean slopes for the reproduced numerosities ($M = .354$, CI [.215, .391]) were also significantly higher than the RTs ($M = .303$, CI [.253, .455]), $t(14) = 3.34$, $p = .005$, CI [.018, .084], $d = 0.863$. When the RTs and the numerosity estimates were both entered in the regression analyses, mean slope for numerosity estimates was significantly higher than the mean slope for the RTs for both targets, $t(12) = 2.737$, $p = .018$, CI [.061, .535], $d = .759$ for T7, and $t(14) = 2.525$, $p = .024$, CI [.055, .675], $d = .652$ for T11.

The numerical CVs for the targets were similar $t(14) = 1.056$, $p = .309$, CI [−.019, .058], $BF = 2.367$ suggesting that participants did not verbally count the number of beeps. Finally, numerical CVs were lower than RT CVs for both

targets, $t(14) = 9.585$, $p < .001$, CI [.019, .03], $d = 2.475$ for T7, and $t(14) = 4.928$, $p < .001$, $d = 1.272$ for T11, indicating that participants relied on numerosities rather than time (see Supplemental Online Material Table S5.2).

Discussion

The results of the current study suggested that humans can monitor not only the magnitude but also the direction of errors in their numerosity estimations and extended the scope of previous findings regarding temporal error monitoring to the numerosity domain. Our results also showed the predictive power of numerosity for the error-monitoring performance above and beyond its RT correlate. Importantly, the study of error monitoring based on magnitude estimations provides the unique opportunity to characterize the quantitative capacity of error-monitoring ability with respect to objective quantitative errors. This cannot be achieved in the context of 2AFC behavior (with binary outputs). In fact, in 2AFC tasks participants have been reported to utilize parametric information (i.e., RTs) as a proxy for confidence judgments (Kiani, Corthell, & Shadlen, 2014). This result suggests that the quantitative capacity error-monitoring ability may constitute its default operational mode, which might be adapted to confidence judgments even in 2AFC behavior based on whatever parametric information is available. Consequently, our results show that error-monitoring is informationally richer than can be captured by earlier work in 2AFC behavior.

A contemporary line of evidence shows that humans and animals can optimize their quantitative and perceptual decisions by taking near normative account of their level of endogenous timing and counting uncertainty (Balci et al., 2011; Çavdaroglu & Balci, 2016). Our results, coupled with the earlier results on temporal error-monitoring, suggest that humans and animals might adapt their decision according to their estimate regarding the level of their uncertainty.

Although for majority of the participants we observed significantly positive slopes using the number of beeps in the analysis, it appears that the confidence ratings for the largest target (T19) might have been affected by the total time of stimulus presentation (a correlate of numerosity). One reason for this might be the underlying uniform distribution that we used to generate the interbeep intervals. That is, CV of the presentation durations to reach a given target number decreases as that target number increases, making time a relatively more reliable source of information for larger numerosities. Alternatively, when a portion of the consecutive beeps are too closely clustered in time, participants might lose track of the count and switch to a time-based strategy instead. Furthermore, perceived numerosity is known to decrease with spatial and temporal proximity, which applies to both static patterns and sequential presentations (Allik & Tuulmets, 1993). These would occur in higher frequency in longer sequences. However, as mentioned above, earlier studies

showed that in timing tasks participants typically prioritize counting over timing (e.g., Fraisse, 1963; Rattat & Droit-Volet, 2012). Thus, in timing studies, participants are typically asked not to count, which has been shown to be sufficient to prevent counting (e.g., Akdoğan & Balci, 2017; Rattat & Droit-Volet, 2012). To this end and as intended, the current study differs from Akdoğan and Balci (2017) as it addresses nonverbal counting rather than interval timing (see Allik & Tuulmets, 1993, for evidence for counting in sequential presentations).

An interesting question that arises from these findings is why a participant with knowledge of their numerical errors would not correct their estimates in the first place. The same question also applies to temporal error monitoring. Akdoğan and Balci's (2017) multiple integrator model showed that the source of the error-related information is the comparison of the integrator that drives the current estimate and secondary integrator(s), the state of which at the time of estimate serves as a benchmark for error monitoring. In this view, error monitoring can be realized only retrospectively and thus cannot guide responding prospectively. Furthermore, the related literature already separates prospective and retrospective judgments of performance and attributes them to different information-processing dynamics and neural mechanisms (Fleming & Dolan, 2012).

An important issue in error monitoring and metacognition literature is the relationship between first-order and second-order performance. In the decision-making domain, measures of metacognitive performance often depend on actual task performance, and devising a method for obtaining a pure metacognitive score is crucial (Fleming & Lau, 2014; Maniscalco & Lau, 2012). Rounis, Maniscalco, Rothwell, Passingham, and Lau (2010) showed that application of TMS to the prefrontal cortex impairs metacognitive performance but leaves stimulus discrimination performance intact. On the other hand, Winman, Juslin, Lindskog, Nilsson, and Kerimi (2014) reported that participants with higher number sense acuity gave more realistic appraisals of their own performance relative to others in a probabilistic reasoning task. In our study, the standardized slopes from individual regression fits are solely determined by confidence judgments and therefore provide an independent measure of error monitoring performance. However, we did not observe a consistent statistically significant relationship between the CV and error-monitoring ability or improvement of performance during the experiment (see Supplemental Online Material Fig. S4, Table S6). Future studies can address if there is a relationship between participants' CVs and their judgements regarding their performance in relation to others.

Finally, another important question that arise from these findings is if the directionality and magnitude of errors are processed by the same or different cognitive/neural mechanisms across quantitative and perceptual domains paving the

path for a more complete understanding of key components of human error-monitoring ability. Future work can investigate if one can disassociate these two components of error-monitoring by using neuroimaging and neuromodulation methods.

Conclusion

The results of the current study suggest that humans can estimate the direction and degree of errors during nonverbal counting, providing evidence for the numerical error-monitoring ability. Consequently, these results coupled with earlier work in interval timing (Akdoğan & Balci, 2017) show that quantitative information in the domain of magnitude representations is accessible to the error monitoring mechanism.

Funding This work was supported by the Turkish Academy of Sciences (TÜBA GEBİP 2015) to F.B.

References

- Akdoğan, B., & Balci, F. (2017). Are you early or late?: Temporal error monitoring. *Journal of Experimental Psychology: General*, 146(3), 347–361. <https://doi.org/10.1037/xge0000265>
- Allik, J., & Tuulmets, T. (1993). Perceived numerosity of spatiotemporal events. *Perception & Psychophysics*, 53(4), 450–459. <https://doi.org/10.3758/BF03206789>
- Balci, F., Freestone, D., Simen, P., deSouza, L., Cohen, J. D., & Holmes, P. (2011). Optimal temporal risk assessment. *Frontiers in Integrative Neuroscience*, 5, 56. <https://doi.org/10.3389/fnint.2011.00056>
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 433–436. <https://doi.org/10.1163/156856897X00357>
- Cordes, S., Gallistel, C. R., Gelman, R., & Latham, P. (2007). Nonverbal arithmetic in humans: Light from noise. *Perception & Psychophysics*, 69(7), 1185–1203. <https://doi.org/10.3758/BF03193955>
- 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>
- Çavdaroglu, B., & Balci, F. (2016). Mice can count and optimize count-based decisions. *Psychonomic Bulletin & Review*, 23(3), 871–876. <https://doi.org/10.3758/s13423-015-0957-6>
- Çavdaroglu, B., Zeki, M., & Balci, F. (2014). Time-based reward maximization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1637), 20120461–20120461. <https://doi.org/10.1098/rstb.2012.0461>
- Dehaene, S. (2011). The number sense : How the mind creates mathematics. New York, NY: Oxford University Press. <https://doi.org/10.1080/00029890.1998.12004997>
- Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594), 1280–1286. <https://doi.org/10.1098/rstb.2012.0021>
- Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8, 1–9. <https://doi.org/10.3389/fnhum.2014.00443>
- Fraisse, P. (1963). Psychology of time. New York, NY: Harper & Row.

- Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, 44(1/2), 43–74. [https://doi.org/10.1016/0010-0277\(92\)90050-R](https://doi.org/10.1016/0010-0277(92)90050-R)
- Gelman, R., & Gallistel, C. (1978). *The child's understanding of number*. Cambridge, MA: Harvard University Press.
- Grondin, S., Meilleur-Wells, G., & Lachance, R. (1999). When to start explicit counting in a time-intervals discrimination task: A critical point in the timing process of humans. *Journal of Experimental Psychology: Human Perception and Performance*, 25(4), 993–1004. <https://doi.org/10.1037/0096-1523.25.4.993>
- Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, 84(6), 1329–1342. <https://doi.org/10.1016/j.neuron.2014.12.015>
- Kibbe, M. M., & Feigenson, L. (2015). Young children 'solve for x' using the approximate number system. *Developmental Science*, 18(1), 38–49. <https://doi.org/10.1111/desc.12177>
- Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1), 422–430. <https://doi.org/10.1016/j.concog.2011.09.021>
- Rattat, A.-C., & Droit-Volet, S. (2012). What is the best and easiest method of preventing counting in different temporal tasks? *Behavior Research Methods*, 44(1), 67–80. <https://doi.org/10.3758/s13428-011-0135-3>
- Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., & Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive neuroscience*, 1(3), 165–175. <https://doi.org/10.1080/17588921003632529>
- Vo, V. A., Li, R., Kornell, N., Pouget, A., & Cantlon, J. F. (2014). Young children bet on their numerical skills: Metacognition in the numerical domain. *Psychological Science*, 25(9), 1712–1721. <https://doi.org/10.1177/0956797614538458>
- Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, 7(11), 483–488. <https://doi.org/10.1016/j.tics.2003.09.002>
- Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*, 10(2), 130–137. <https://doi.org/10.1111/1467-9280.00120>
- Winman, A., Juslin, P., Lindskog, M., Nilsson, H., & Kerimi, N. (2014). The role of ANS acuity and numeracy for the calibration and the coherence of subjective probability judgments. *Frontiers in Psychology*, 5, 851. <https://doi.org/10.3389/fpsyg.2014.00851>