ERROR MONITORING IN MAGNITUDE REPRESENTATIONS

by

YALÇIN AKIN DUYAN

Submitted to Koç University's Graduate School of Social Sciences and Humanities

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Psychology

Koç University

November 2019

Koç University
Graduate School of Social Sciences and Humanities

This is to certify that I have examined this copy of a doctoral dissertation by

Yalçın Akın Duyan

and have found that it is complete and satisfactory in all respects,

and that any and all revisions required by the final

examining committee have been made.

Committee Members:

_____

Prof. Fuat Balcı

_____

Asst. Prof. Terry Eskenazi

_____

Asst. Prof. Çağlar Akçay

_____

Asst. Prof. Aslı Kılıç Özhan

_____

Prof. Seda Dural Çetinkaya

Date: _____

Statement of Authorship

This dissertation contains no material which has been accepted for any award or any other degree or diploma in any university or other institution. It is affirmed by the candidate that, to the best of his knowledge, the dissertation contains no material previously published or written by another person, except where due reference is made in the text of the thesis

The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

Signature Yalçın Akın Duyan

ABSTRACT OF THE DISSERTATION

ERROR MONITORING IN MAGNITUDE REPRESENTATIONS

by YALÇIN AKIN DUYAN

Dissertation Director:

Dr. Fuat Balcı

This dissertation aims to investigate error monitoring abilities for metric representations of number and space. Error monitoring, which is the ability to judge the accuracy of one's actions, have typically been categorical choice tasks. However, recent evidence in the timing domain suggests that error monitoring in humans also extends to continuous representations of magnitude, indicating that it has informationally richer foundations than previously thought. Considering the common psychometric properties of magnitude representations, I hypothesized that judgments of error direction and subjective confidence would reflect objective performance in other magnitude domains. Over three studies, error monitoring abilities was observed in 1) sequentially presented numerosities, 2) numerosity of simultaneously presented arrays and 3) line lengths. These results indicate that the ability to keep track of the direction and the degree of errors in their magnitude representations is a ubiquitous human trait.

ÖZET

Bu tez, insanların sayı ve uzam temsillerindeki hata izleme yetilerini araştırmayı amaçlamıştır. Hata izleme, yani bireylerin kendi davranışlarının doğruluğunu değerlendirebilme yetisi, genellikle kategorik karar verme görevleriyle incelenmiştir. Aralık zamanlama literatüründeki son bulgular, insanların hata izleme yetilerinin sürekli miktar temsilleri için de geçerli olabileceğini göstermiştir. Bu sonuçlar, insanlarda hata izleme becerilerinin bilgisel açıdan daha zengin temelleri olabileceğine işaret etmektedir. Miktar temsillerinin psikometrik özellikleri arasındaki benzerliklerden yola çıkılarak, insanların bu temsillerde yaptıkları hataların yönünü ve derecesine ilişkin değerlendirmelerin, nesnel performanslarını diğer alanlarda da yansıtabileceği öngörülmüştür. Yürütülen üç farklı çalışmada, insanların hata izleme yetilerinin 1) sıralı biçimde sunulan sayı tahminlerinde, 2) eşzamanlı olarak sunulan sayı tahminlerinde ve 3) uzamsal uzunluk tahminlerinde hata izleme yetilerine ilişkin kanıtlar sunulmuştur. Bu sonuçlar, insanların miktar temsillerindeki hatalarının yönünü ve derecesini değerlendirebilme yetilerinin genel-geçer bir insan özelliği olduğuna işaret etmektedir.

ACKNOWLEDGEMENT AND DEDICATION

First and foremost, I thank my supervisor Fuat Balcı for his support and guidance throughout my studies. I feel grateful and lucky to have had the opportunity to work with someone as creative and knowledgeable as him.

I also want to thank Aybike and Ezgi for their friendship and for all the good times.

I would like thank my dearest friends Hakan, Sezay, Serdar, Erkan, Volkan, Mehmet, İlker for their invaluable friendship.

I thank Öykü, for her constant support and affection in the better part of this journey; and making it so too.

I also want to thank my family, for supporting my decision and always appreciating what I do.

Last but not least, I want to thank my previous supervisors, Nurhan Er and Hakan Çetinkaya for their support and mentorship during my undergraduate and masters training.

It is always a good story when you are through with it.

Author's note: The first chapter has been published in Psychonomic Bulletin & Review. The second and third chapters have been published in Consciousness and Cognition.

Duyan, Y. A., & Balcı, F. (2018). Numerical error monitoring. *Psychonomic bulletin & review*, *25*(4), 1549-1555.

Duyan, Y. A., & Balcı, F. (2019). Metric error monitoring in the numerical estimates. *Consciousness and cognition*, *67*, 69-76.

Duyan, Y. A., & Balcı, F. (2020). Monitoring line length reproduction errors. *Consciousness and cognition*, *77*, 102831.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Every action is accompanied by a subjective degree of certainty about the veracity of that action. These actions can involve having to choose an appropriate option from a set of distinct alternatives, such as deciding what car to buy, or which road to take when commuting. On the other hand, some actions require a ballpark estimate of a certain magnitude like time, number, spatial length or distance, such as estimating whether a piece of furniture that you plan to buy is going to fit in your room. While it's been long known that confidence ratings follow the accuracy of categorical decisions to some degree (Baranski & Petrusic, 1998; Dougherty, 2001; Garrett, 1922; Johnson, 1939; Nelson & Narens, 1990; Vickers, 1979), error monitoring in magnitude representations is a relatively recent discovery. This dissertation investigates error monitoring abilities in the continuous domain, namely, representations of number and spatial length. Over a series of studies, it lays the behavioral groundwork for a potentially rich line of work that can provide a deeper understanding of the mechanisms that give rise to this metric error monitoring ability.

Models of Error monitoring in categorical decisions

Error monitoring abilities in humans have typically been studied using Two Alternative Forced Choice tasks (2AFC), where the observer is required to choose among two distinct options. There's a variety of theoretical models that aim to predict the differential response probabilities in forced-choice tasks, along with the time taken that to make that decision. The most prominent of these are a family of drift-diffusion models that characterize decision making as a noisy evidence accumulation process, where a decision is made when the accumulated evidence for a given option reaches a subjective threshold (Brown & Heathcote, 2005, 2008; Ratcliff, 1978;

Smith & Ratcliff, 2004; Teodorescu & Usher, 2013; Usher & McClelland, 2001). While these models have had a great deal of success in accounting for the response and reaction time (RT) patterns in decision-making behavior, they cannot readily account for the subjective confidence that accompanies these decisions (Van Zandt, 2000; Van Zandt & Maldonado-Molina, 2004; Vickers, 1979). It follows from the core assumptions of these models that every decision is made with absolute certainty. Hence, they cannot account for the fact that subjective confidence ratings closely follow the overall accuracy of decisions. This well-established finding indicates that even categorical decisions are made along a continuum.

Decision confidence has been suggested to emanate from the hierarchical processing of information from the lower motor and perceptual systems on a higher cognitive level. While it is expected to observe indicators of decision confidence in neural populations that allow for the first order perceptual decisions, it is possible that there is a higher-order processing mechanism that 'reads out' and translates this information into confidence ratings. In fact, a dissociation between conscious and nonconscious processing of errors have been demonstrated. For instance, Logan et al. (2010) asked skilled typists to type target words on the computer. On some trials, even when the participants typed the word correctly, the program inserted typing errors on the screen. Conversely, on some trials, their errors were automatically corrected. The results showed that post-error slowing occurred only after genuine errors and not when pseudo-errors were inserted, indicating that error monitoring is at least partially a nonconscious process.

Models of error monitoring have been mostly extensions of the aforementioned drift-diffusion models (Balakrishnan & Ratcliff, 1996; Moran, Teodorescu, & Usher, 2015; Pleskac & Busemeyer, 2011; Resulaj, Kiani, Wolpert, & Shadlen, 2009; Vickers, 1979; Vickers & Packer, 1982). These models also assume that decision making is the result of evidence accumulation

that is subject to internal and external noise. As mentioned above, the most crucial and curious aspect of decision confidence is its positive relation to decision accuracy. People tend to report higher confidence for the accurate decisions and lower confidence for the erroneous decisions termed the *positive resolution of confidence* (Baranski & Petrusic, 1998; Dougherty, 2001; Garrett, 1922; Johnson, 1939; Nelson & Narens, 1990; Vickers, 1979). The two most fundamental questions about decision confidence concern *what* information is utilized to form confidence and *when* this information is processed. Models of confidence can be distinguished based on the assumptions they make on these two questions. Single-stage models (Vickers, 1978), assumes that confidence is based on the mechanistic properties of a decision itself. They may assume that confidence is calculated heuristically by taking into account the duration of the decision process, or some other calculation based on the same evidence as the initial evidence. In contrast, dual-stage models assume that evidence accumulation continues after a decision has been made, and confidence relies on this extra information.

As mentioned above, some models assume that the decision and the confidence judgment are based on the information collected prior to the decision (Egan, Schulman & Greenberg, 1959; Kepecs, Uchida, Zariwala, & Mainen, 2008). These models assume that confidence is calculated by measuring the distance between the perceptual sample and the decision criterion. The Balance of Evidence (BOE; 1979) model posits that for any 2AFC decision, the evidence is accumulated separately for each alternative and a decision is made when one of them reaches a threshold. The difference (or distance) between the alternatives determine decision confidence.

Single-stage models can be further categorized by the type of information they assume used to calculate confidence. BOE provides a computational account, which assumes confidence is a monotonic function of the difference between the total evidence accumulated by each

accumulator. On the other hand, some models provide a heuristic account (Ratcliff, 1978; Volkman, 1934; Zakay & Tuvia, 1998), suggesting that confidence is calculated directly from the time it took to make that decision (i.e., decision time).

Dual-stage theories of confidence assume that evidence accumulation continues in the post-decisional stage (Pleskac & Busemeyer, 2011; Moran et al., 2014). This way, the inter-judgment time (the interval between decision and confidence response) provides more information, and the confidence response is based on both the choice-stage and the inter-judgment stage dynamics. In error trials, post-decision evidence will tend to contradict the initial decision. If the observer is sensitive the discrepancy between the pre- and post-decision evidence, a positive resolution will emerge.

As mentioned above, error monitoring and metacognition have generally been studied using two-alternative forced choice tasks (Gehring et al., 1993 but see Miltner, Braun & Coles 1997). However, perceptual decision-making domain is a world of absolute rights and wrongs that may not always map onto real life actions that require an approximate estimation of continuous magnitudes such as time, number and space. Utilization of these magnitude representations to study error monitoring posits a unique opportunity to discover the potentially rich informational basis of human metacognitive capacities, as these representations are relatively independent of the temporal dynamics that are involved in decision making.

The nature of magnitude representations

Magnitude estimations are an essential part of daily life, whether we want to estimate how much time we have spent on a certain task, or whether our car will fit an empty parking spot. The study of physical magnitudes and their corresponding mental representations is as old as the science of

psychology itself, and there are well-established characteristics of these representations that are observed across all types of magnitudes.

One such characteristic is that while humans are on average accurate estimators of magnitude, each estimation is subject to trial-to-trial variability. In other words, even when people are estimating (or reproducing) the same magnitude (e.g. a duration of 2 seconds, or a length of 11 centimeters) over and over, these estimates are rarely identical. This variability is well characterized by a response curve that is normally distributed around a mean which is equal to the target estimate, and a standard deviation that is determined by how precise an individual is in their representation of that specific magnitude. Hence, actions that are based on the estimation of a continuous magnitude can be conceptualized as actions that are made under uncertainty, with varying degrees of uncertainty across individuals.

Problem of optimality

Optimality, in a general sense, is about maximizing the return of action by incorporating the constraints, outcomes, and the uncertainties of the circumstances that is utilized to exert that action. Given that categorical decision and magnitude estimations are subject to internal sources of uncertainty, observing how close to optimality people act in these two contexts can give an idea of whether they can normatively incorporate their internal uncertainty into their actions (thus have a veridical representation of this knowledge).

In decision making, the problem of optimality primarily concerns the time taken to reach a decision. For example, if a task is too easy and requires a minimal amount of processing, then the decision-maker should spend as little time as possible on that task. On the other hand, if the task is too difficult to ever tackle, the decision-maker should also spend as little time as possible,

or even ignore the task altogether. On tasks that are intermediately difficult, one should find the optimal balance between speed and accuracy that would maximize the rate of reward per unit time (Bogacz et al., 2006). Recent studies show that a substantial proportion of humans are optimal decision-makers (Bogacz, Hu, Holmes & Cohen, 2010) and they can learn to be optimal over time (Balcı et al., 2011a).

For actions that require the estimation of a certain magnitude such as time, optimality concerns the incorporation of one's subjective uncertainty by aiming for a magnitude that would maximize the average rate of reward. For example, let's assume your commute to work takes an hour on average and the standard deviation of these commutes is 5 minutes. If you want to arrive at work as close as possible at the time where the workday starts and minimize the number of days that you arrive late, you should leave home about 70 minutes before clocking in to find the optimal balance between leaving home as late as possible and being late on a few days as possible. Using a similar task, a recent line of work (Balcı et al., 2011; Çavdaroğlu, Zeki & Balcı, 2014; Çavdaroğlu & Balcı, 2016; Gür & Balcı, 2018; but see Berkay, Freestone & Balcı, 2016) showed that humans and rodents maximize the rate of reward in their actions that involve timing or counting. These results suggest that humans and animals can normatively incorporate their subjective level of uncertainty in their actions, suggesting a representation of uncertainty itself. This uncertainty representation might simply be an aggregate or a summary statistics representation emerging from an error monitoring mechanism in magnitude representations.

Error monitoring in magnitude representations

Recently, Akdoğan and Balcı (2017) showed that humans can monitor the magnitude and the direction of their timing errors. In a series of studies using the temporal reproduction task, they

asked participants to reproduce two different target durations as accurately as possible. After each response, they were prompted to provide confidence ratings about the accuracy of their estimates and whether this estimate was longer or shorter than the target interval. They found that the confidence ratings and the reported direction of errors closely matched the participants' actual performance (see also Kononowicz & van Vassenhove, 2018; Doenyas, Mutluer, Genç & Balcı, 2019). Similarly, using a task where the participants reproduced the orientation of a low contrast grating Samaha and Postle (2017) observed that the absolute errors of reproductions diminished with increased confidence. This suggests that even when people are estimating a certain magnitude to the best of their ability on a single trial, information on the magnitude and the direction becomes available at some point before the confidence judgment is made.

While the positive relationship between confidence and accuracy is well-established, these results show that this metacognitive ability extends to the estimation of continuous magnitudes (i.e. interval timing). Importantly, however, whether this ability to monitor the magnitude and the direction of errors applies to other domains is unknown. Nevertheless, well-established evidence for a common magnitude processing (e.g., Walsh, 2003) and recent evidence for a task-independent confidence processing makes a plausible case for this account.

A common mechanism for magnitude processing

Walsh (2003) proposed that there's a ubiquitous brain mechanism for magnitude processing. Indeed, there's substantial neural and behavioral evidence for this assumption (Bueti & Walsh, 2009; Fabbri & Natale, 2009). There are specific mappings between time, space and number (Arzy, Adi-Japha, & Blanke, 2009; Arzy, Collette, Ionta, Fornari, & Blanke, 2009; Santiago, Lupiáñez, Perez, & Funes, 2007; Torralbo, Santiago, & Lupiáñez, 2006; Moyer & Landauer,

1967; Restle, 1970 (MTL); Dehaene, Bossini, & Giraux, 1993). In their seminal work (Dehaene, Bossini & Giraux, 1993) discovered that people respond to small numbers faster with a left key and respond to large numbers faster with a right key. Since then, similar mappings between space and time (Ishihara, Keller, Rossetti, & Prinz, 2008) and time and number (Kiesel & Vierck, 2009) have also been discovered. Moreover, magnitude information from one domain seems to affect the magnitude estimations for other domains, depending on the reliability of the information from the interfering domain. Weber's law is valid for a vast number of continuous dimensions. DeWind and Brannon (2012) found that the Weber fractions on a numerosity comparison task correlated with the Weber fractions on a line length comparison task. Moreover, work with rats and pigeons showed that these species of animals can readily apply previously learned temporal associations to a similar numerical task (Meck & Church, 1983). These results provide converging evidence for a common magnitude processing system.

Error monitoring across different domains

There is also some evidence of similar metacognitive performance across different tasks within and separate domains that suggest confidence might also be processed by a common mechanism (Bornstein and Zickafoose, 1999; Heereman, Walter and Heerkeren, 2015) in a task-independent fashion. Gardelle and Mamassian (2014) assessed whether there is a common currency for confidence across tasks using spatial discrimination and orientation discrimination tasks. Participants performed a pair of trials in succession, consisting of either two trials of the same task or one of each task. They then indicated on which trial they were more confident. The psychometric functions of confidence were identical for between- and within-task comparisons. Similar results were also obtained between vision and audition suggesting that this costless translation of confidence extends across modalities (Gardelle, Corre and Mamassian, 2016).

Moreover, the effects of stimulus variance on confidence were found to be similar on a numerical and a visual task, although the former task is cognitive while the latter is perceptual (Navajas et al., 2017). Finally, metacognitive accuracy and bias appear to be consistent within individuals across different tasks and domains and time (Ais, Zylberberg, Barttfeld & Sigman, 2016).

The primary incentive for the studies presented below is to determine whether this recently discovered parametric error monitoring ability in magnitude estimation extends to other domains as implicated by previous work from a wide range of areas. These studies provide the behavioral foundations for a potentially rich line of work, which can be extended to other domains in humans and animals and provide a more precise understanding of the neural mechanisms that gives rise to this error monitoring ability. The study of metric error monitoring would also provide us with more suitable empirical tools to better understand how error processing in the brain is parametrically related to corrective actions.

CHAPTER I

Numerical Error Monitoring

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

Introduction

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, & Balcı, 2014; Çavdaroğlu & Balcı, 2016) has addressed the importance of the above mentioned 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 & Balcı, 2016; Çavdaroğlu, Zeki, & Balcı, 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 Balcı (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 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 Supplement Chapter 1, 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 (β = .217, *SE* = .026, $p < .001$, R^2 = .26 for T11; β = .109, *SE* = .01, $p < .001$, R^2 = .189 for T19), indicating that confidence judgments in general followed objective performance (see Table 1.1). 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. 1.1, 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 Supplement Chapter 1 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, $t(28) = 1.041$, $p = .31$, $BF_{10} = 3.095$.

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, BF10 = 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, $t(29) = 1.264$, $p = .217$, $BF_{10} = 2.465$.

	Estimate	*Standard Error*	*df*	*t value*	*p*	*CI*	*AIC*	*Log Likelihood*
Experiment 1								
T11	*.217*	*.026*	*3452*	*8.443*	*< .001*	*.167-.267*	*11904*	*-5947*
T19	*.109*	*.01*	*2915*	*11.157*	*< .001*	*.09-.129*	*10683*	*-5336.5*
Experiment 2								
T7	*.483*	*.08*	*1615*	*6.017*	*< .001*	*.326-641*	*5525.7*	*-2757.9*
T11	*.323*	*.063*	*1809*	*5.131*	*< .001*	*.199-.446*	*6361.5*	*-3175.7*

Table 1.1 *Summary results of the linear mixed-effects models. The full model table and diagnostic plots are provided in Supplement Chapter 1, Figures S3.1, S3.2, S3.3, S3.4*

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 & Balcı, 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], $BF_{10} = 3.45$. 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 Supplement Chapter 1, 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 (β = .483, SE = .08, p < .001, R^2 = .262 for T7; β = .323, SE = .063, p < .001, R^2 = .178 for T11; see Table 1.1).

The mean standardized slopes were .412 (CI [.309, .516]) and .354 (CI [.253, .455]) for T7 and T11, respectively (see Fig. 1.1, 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 Supplement Chapter 1, 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.

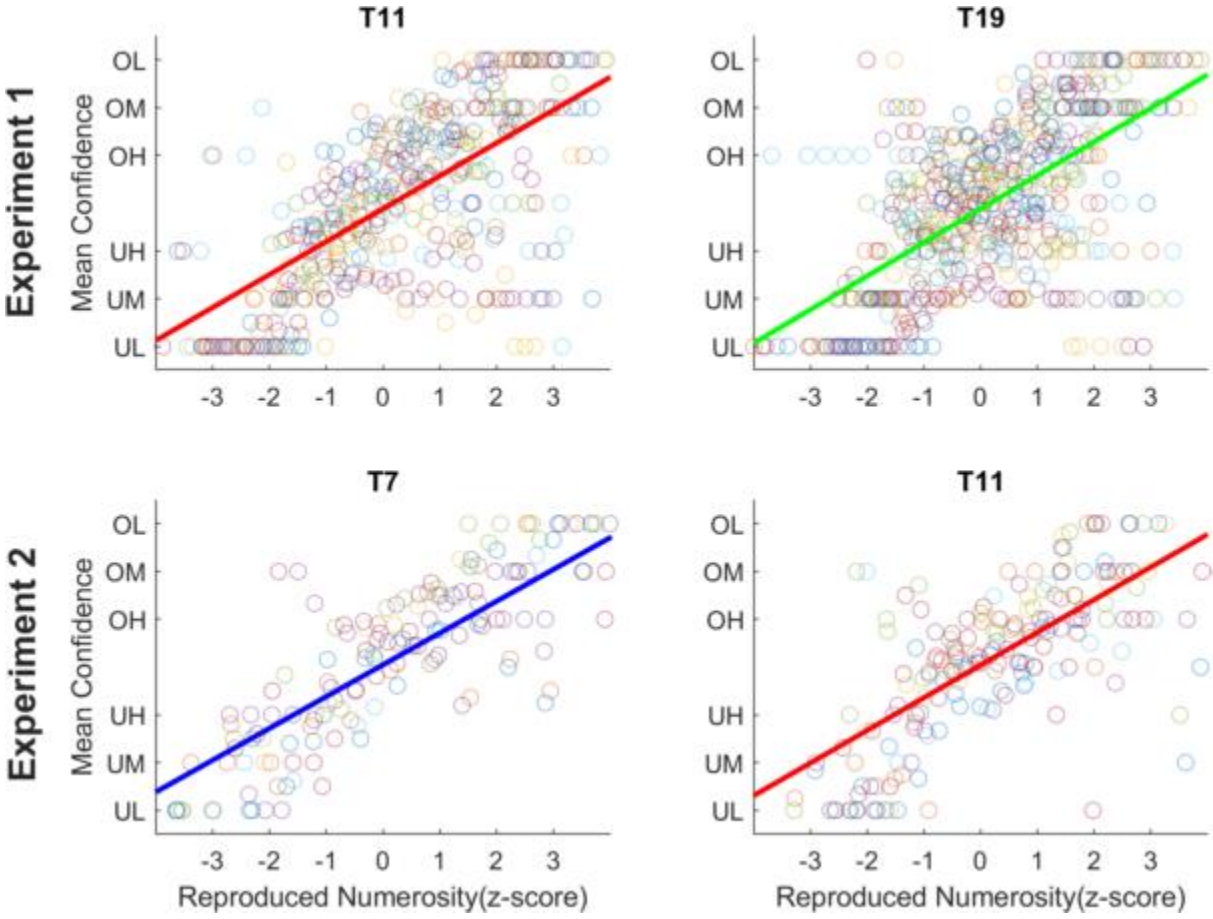

Fig. 1.1 Relationship between numerosity estimates and signed confidence ratings. Average signed confidence ratings (-3: UL, -2: UM, -1: UH, 1: OH, 2: OM, 3: OL) as a function of z-score transformed numerosity estimates (including on target responses) for Experiment 1 (top panel) and Experiment 2 (bottom 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. The plots show the linear fits of mean confidence ratings on the z-transformed numerical reproductions from all participants T7 = Target 7, T11 = Target 11, T19 = Target 19. Note that T11 was included both in Experiment 1 and Experiment 2.

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], $BF_{10} = 2.367$ 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 Supplement Chapter 1, 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 (Balcı et al., 2011; Çavdaroğlu & Balcı, 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 & Balcı, 2017; Rattat & Droit-Volet, 2012). To this end and as intended, the current study differs from Akdoğan and Balcı (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 Balcı'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 Supplement Chapter 1, 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 arises 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 & Balcı, 2017) show that quantitative information in the domain of magnitude representations is accessible to the error monitoring mechanism.

CHAPTER II

Metric Error Monitoring in the Numerical Estimates

Abstract

Recent studies have shown that participants can keep track of the magnitude and direction of their errors while reproducing target intervals (Akdoğan & Balcı, 2017) and producing numerosities with sequentially presented auditory stimuli (Duyan & Balcı, 2018). Although the latter work demonstrated that error judgments were driven by the number rather than the total duration of sequential stimulus presentations, the number and duration of stimuli are inevitably correlated in sequential presentations. This correlation empirically limits the purity of the characterization of "numerical error monitoring". The current work expanded the scope of numerical error monitoring as a form of "metric error monitoring" to numerical estimation based on simultaneously presented array of stimuli to control for temporal correlates. Our results show that numerical error monitoring ability applies to magnitude estimation in these more controlled experimental scenarios underlining its ubiquitous nature.

Keywords: Numerical estimation, Metric error monitoring, Mathematical cognition**,** Magnitude representation

Introduction

Error monitoring refers to the ability of humans to keep track of their errors in their decisions, choices, and judgments without guidance of objective feedback. Although this ability has been investigated primarily in behavioral paradigms that require binary choice based on some stimulus property (Fleming, Dolan, & Frith, 2012), at least three recent studies have shown that error monitoring relies on richer metric information that cannot be captured by two alternative forced choice behavior. Specifically, these studies revealed that humans can keep track of the direction and magnitude of errors in their reproduction of target time intervals (Akdoğan & Balcı, 2017 - see also Kononowicz, Roger, & van Wassenhove, 2017) and numerosity of sequentially presented stimuli with random ISIs (Duyan & Balcı, 2018). Although, based on a number of analyses Duyan and Balcı (2018) demonstrated that error judgments were driven by the number rather than the total duration of sequential stimulus presentations in their counting task, in sequential presentations the number and duration of stimuli are inevitably correlated empirically limiting the characterization of "numerical error monitoring". Critically for the study of numerical error monitoring, humans are also known to be able to estimate (on average) the numerosity of simultaneously presented array of items (Gebuis & Reynvoet, 2012). In order to reach at a more complete and cleaner (without the time correlate) characterization of "numerical error monitoring" ability, the current study investigated if metric error monitoring generalizes to the estimates of the numerosity of simultaneously presented array of items.

Recently, a number of studies provided evidence for a metric error monitoring mechanism in magnitude representations (e.g. Akdoğan & Balcı, 2017; Kononowicz et al., 2017, Duyan & Balcı, 2018). In the timing domain, Akdoğan and Balcı (2017) showed that in a temporal reproduction task using durations ranging between 1.5 and 6 s, participant's confidence ratings

and error-directionality judgments reflected the amount and the direction (over- or under-reproduction) of errors on a trial-to-trial basis. Moreover, for longer durations, people are also aware of their general biases in time perception (Brocas, Carrillo, & Tarraso, 2018). For number representations, a few studies hinted at a possible numerical error monitoring mechanism (e.g., Gelman & Gallistel, 1978; Kibbe & Feigenson, 2015). For instance, Vo, Li, Kornell, Pouget, and Cantlon (2014) found that children made more high-risk bets after correct decisions and on easier trials in a numerical discrimination task.

The counting ability of humans has been widely investigated in the literature. One important conclusion that can be derived from these studies is that numerosity estimates based on sequentially presented signals and simultaneously presented array of items are subject to the similar cognitive procedures and/or constraints pointing at an overlap between the representations (namely Approximate Number System; Dehaene, 2011; Gallistel & Gelman, 1992, 2000) that result from these two different presentation formats characterizing an abstract number sense (e.g., Barth, Kanwisher, & Spelke, 2003; but see Tokita & Ishiguchi, 2012). This suggests that numerical error monitoring can be studied based on the numerosity estimates about simultaneous array of items without the temporal correlates (compared to sequential presentations).

In support of a common abstract numerical representational system activated independent of sensory modalities and presentation format (simultaneous vs. sequential), Barth et al. (2003) showed that there was no performance cost for comparing numerosities between modalities and formats. This behavioral observation suggested that the quantitative comparisons were made based on a common and abstract mental metric activated by different experimental settings. Corroborating this conclusion, Arrighi, Togoli, and Burr (2014) showed that adaptation (i.e., the effect of viewed numerosity on the enumeration of the subsequently presented stimuli)

generalized not only between sensory modalities but also presentation formats. Importantly, the degree of this cross modality and presentation format adaptation was as strong as those observed in within modalities and formats.

Support for a common numerical representational system on behavioral and psychophysical bases were coupled also with neural evidence. For instance, Nieder, Diester, and Tudusciuc (2006) showed that a group of neurons in the intraparietal sulcus of monkeys exhibited numerosity selectivity irrespective of sequential vs. simultaneous presentation of to-be-enumerated items. The observations were corroborated by human neuroimaging work. For instance, Dormal, Andres, Dormal, and Pesenti (2010) tested humans with simultaneously vs. sequentially presented stimuli and their conjunction analysis showed that right intraparietal sulcus and precentral gyrus were commonly activated with a very similar activation pattern during numerical judgments made based on both presentation formats.

These different lines of evidence that point at a common numerical representational system both necessitate the study of numerical error monitoring previously reported with sequential presentations also with simultaneous array of stimuli but also enables the characterization of this ability independent of the temporal correlate of numerosity faced during the sequential presentation format. To this end and different from our previous study, we used a numerical estimation task rather than a production task. Despite similar performance between different task structures, Crollen, Castronovo, and Seron (2011) showed that the perception, production and reproduction of numerosities can result in different patterns of performance. In general, participants over-estimate numerosities when asked to produce symbolically presented numbers (Castronovo and Seron, 2007; Cordes, Gelman, Gallistel, & Whalen, 2001; Whalen, Gallistel, & Gelman, 1999). However, they under-estimate the number of simultaneously

presented items on a display (e.g. an array of dots), especially for numerosities over 20 (Castronovo & Seron, 2007). Consequently, the investigation of numerical error monitoring with the numerical estimation task will complement previous findings regarding the numerical-representational-dependency rather than peculiar task-dependency of the metric error monitoring ability.

Method

Participants

18 undergraduate students from Koç University participated in the experiment for course credit. All participants gave informed consent prior to testing. 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-in. monitor. Experiments were controlled via Matlab (Mathworks, Natick, MA) using the Psychophysics toolbox (Brainard, 1997) on an iMac. Raw data and the analysis code can be accessed at the Open Science Framework (osf.io/xae6k).

Stimuli

The numerical stimuli were arrays of grey dots on a black background, generated using a modified version of the program described in Gebuis and Reynvoet (2011, 2012). Overall, the numbers ranged between 4 and 29. The target numbers (i.e., 4, 7, 11, 16, 22 and 29) were presented on half of the trials (40 each). The rest of the numbers were presented on the rest of the trials as filler numerosities (12 each). The program developed by Gebuis and Reynvoet (2011) manipulates three visual properties of the dot array: the convex hull, average diameter of the dots

and density. A total of 480 arrays were generated for the task. During the experiment, all participants were presented the same pre-generated stimuli set in a random order.

Procedure

A random dot array was presented for 500 ms randomly from the stimulus set. A question mark immediately appeared on the screen, and the participants typed in their numerosity estimate using the number pad on the keyboard. Participants were prompted to provide a confidence rating by pressing Q (low), W (medium) or E (high) keys to indicate (100-milliseconds after their initial response) how close they thought their estimate captured the actual number of dots in the array. They were then (after 100-milliseconds) asked whether they undershot or overshot the target by pressing A or D keys, respectively. The intertrial interval (ITI) varied between 1.5 and 2.5 s (uniformly-distributed) (see Fig. 2.1).

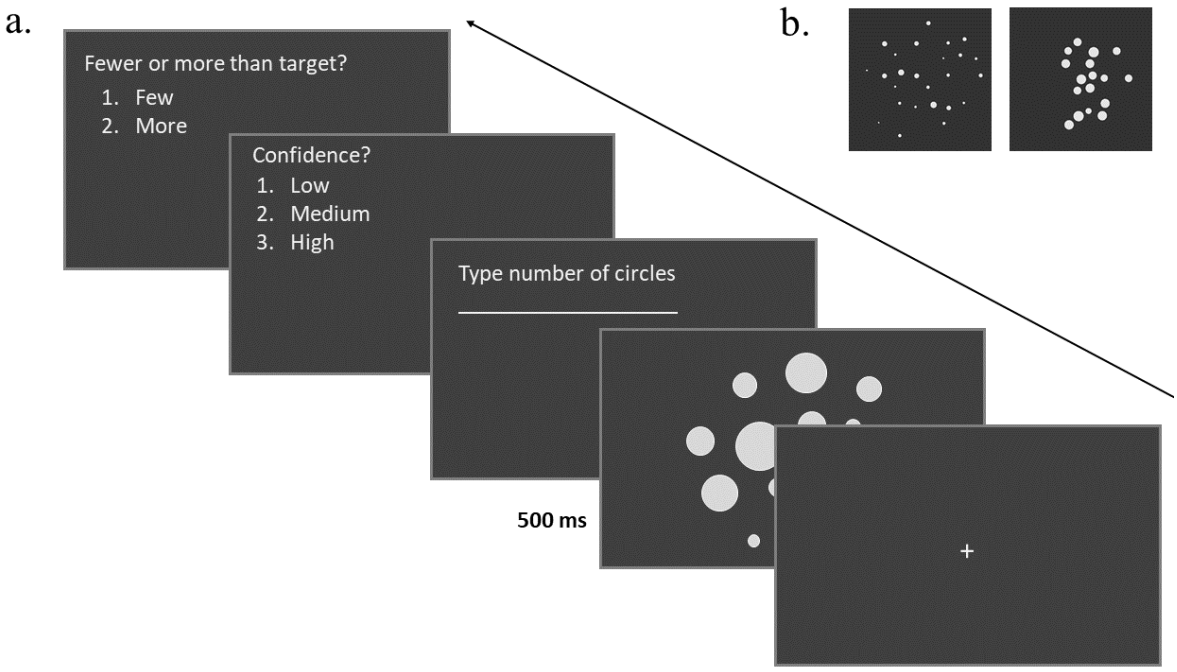

Fig 2.1 An illustration of the experimental task (a) and sample arrays used in the experiment (b).

Data analysis

In order to provide the experience of a wide range of numerosities to the participants, we included set sizes between 4 and 29 with the prior knowledge that they would have performed differently for set sizes that are small. In order to base our analysis on those set sizes that were processed similarly (e.g., not via subitizing), we made piecewise linear fits to the average proportion of on-target responses for different set sizes. The fit that maximized the difference between the slopes of two lines led to an inflection point at 10 (pre-inflection point slope=−0.125; post-inflection point slope=−0.011) suggesting that our participants likely adopted different enumeration strategies for set sizes below 10. Compared to other numerosities (target or filler), participants committed substantially lower frequency of errors in their numerosity estimates for counts lower than 10. This is likely due to subitizing and chunking (Gobet et al., 2001; Mandler & Shebo, 1982). These counts were however important during testing as they contributed to the wider range of numerosities faced by the participants to prevent them from honing into specific targets.

For each participant, we excluded trials where a numerical estimate was 3 mean absolute deviations (MAD) for that participant's mean reproduction for a given target numerosity. Participants' and directionality judgments on their estimates yielded six confidence direction pairs: Under(U)-Low(L), Under(U)-Medium(M), Under(U)-High(H), Over(O)-High(H), Over(O)-Medium(M), Over(O)-Low (L). These response categories were numerically coded as −3, −2, −1, 1, 2 and 3, respectively. Confidence ratings reflected how close the participants thought their estimates were to the actual numerosity, regardless of whether they over- or underestimated the target. Moreover, the directionality judgments (i.e over/under) should reflect under- or overestimations in their responses. Hence, if humans can correctly judge the precision

of their numerical estimates, trials with low confidence ratings should on average be further away from the target, on an individual basis. In the ideal case, the mean estimate for a specific numerosity should be the lowest on trials with UL judgments and the highest on trials with OL judgments. Consequently, the regression of these confidence-direction pairs on the numerosity estimates should yield a positive slope. A higher slope will reflect a better error monitoring ability.

Table 2.1 Mixed effects model results for target numerosities only.

	Estimate	*SE*	*tStat*	*DF*	*pValue*	*Lower*	*Upper*
Fixed Effects							
Intercept	-.197	.221	-.894	2788	.371	-.63	.235
Response	.435	.057	7.687	2788	<.001	.324	.545
Random Effects Group: Participant (18 Levels)							
	Std	*Lower*	*upper*				
Intercept	.923	.66	1.292				
Response	.182	.104	.321				
AIC: 11634 *BIC*: 11664 *Log Likelihood*: -5812.2							

To capture the whole data with a single model, we calculated the z-scores from each participant's responses for each target numerosity. Thus, the z-scores reflect the participant's amount of deviation from his/her mean estimate for a specific numerosity. This way, we could include all the target numerosities in a single model. To explore whether participants' signed confidence judgments followed their estimation performance, we regressed the confidence judgments on the z-scores from all trials by constructing a linear mixed effects model. Because

we expect the error monitoring performance to differ among individuals (hence, different slopes), we included the participants as a random factor on the slope and the intercept. Effect sizes (d) were calculated following Westfall, Kenny, and Judd (2014).

Results

In the first model, we included the responses for the target numerosities that were above the aforementioned inflection point (i.e. 11, 16, 22, 29). The linear mixed effects model revealed a statistically significant positive slope for the effect of subjective numerosity estimates on signed confidence ratings, $\beta = 0.435$, $SE = 0.057$, $p < .001$, $R^2 = 0.22$, $d = 0.204$ pointing at an error monitoring ability for subjective numerical representations (Table 2.1).

In the second model, in addition to the targets included in the previous model, we also included the responses for the filler numerosities. The model yielded similar results, revealing a significant positive slope for the z-transformed numerosity estimates, $\beta = 0.447$, $SE = 0.057$, $p < .001$, $R^2 = 0.23$, $d = 0.209$ (see Table 2.2, Fig. 2.2).

The same set of analyses were also conducted for targets that were higher than 7 (the first target after 4, which falls within the subitizing range). The same results held (see Supplement Chapter 2, Table S1 & S2).

As mentioned above, the confidence-directionality pairs did not include zero, because the participants had no way of reporting that their response was exactly on target. While this manner of coding is intuitive given the rationale of our approach, we should note that it runs the risk of estimating a higher slope for the fixed effects factor when the less-more judgments are made accurately, compared to when they are coded from 1 to 6. Thus, we also fitted the same mixed effects models with the latter manner of coding these judgment pairs, and the results (despite a lower slope) did not change: $\beta = 0.333$, $SE = 0.044$, $p < .001$, $R^2 = 0.218$, $d = 0.2$ for the first

model; $\beta = 0.343$, $SE = 0.044$, $p < .001$, $R^2 = 0.228$, $d = 0.203$ for the second model (See Supplement Chapter 2, Table S3 & S4).

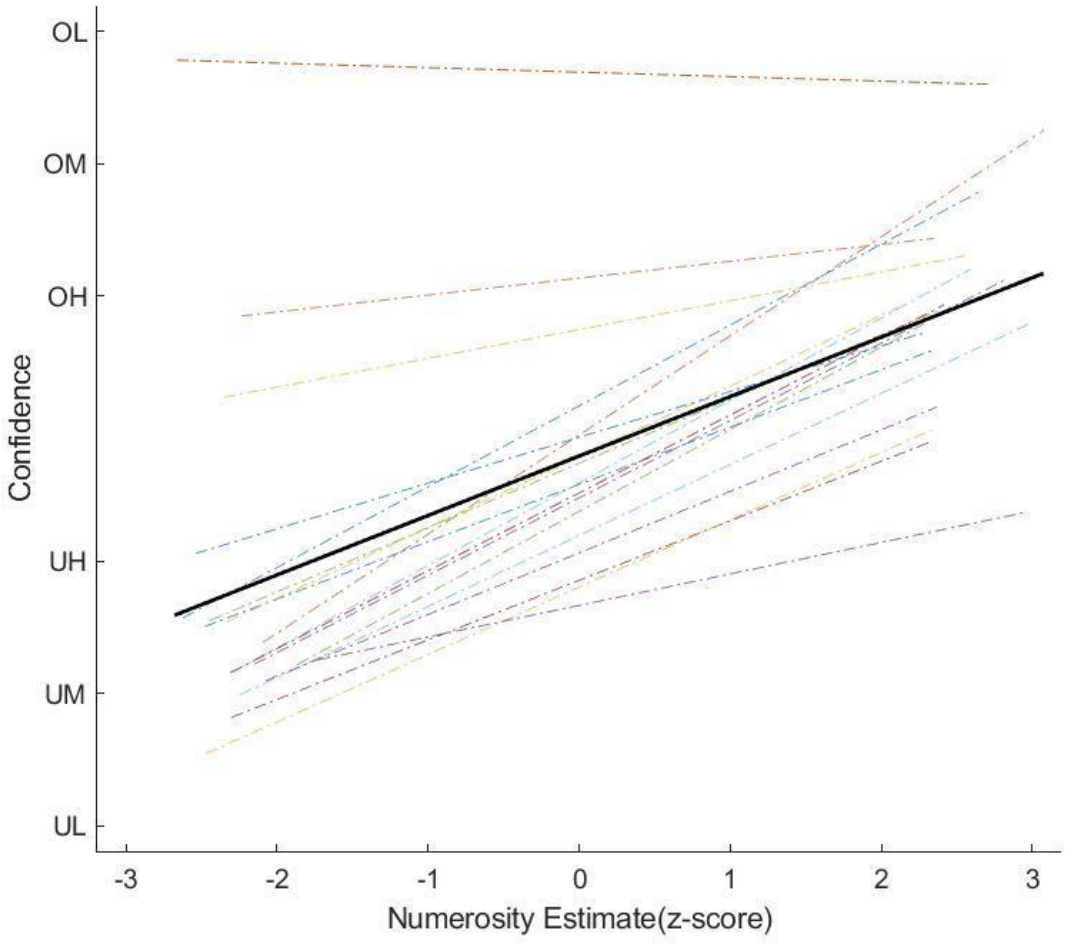

Fig 2.2 Each dashed line represents the fits for individual participants gathered from the linear mixed effects analysis. The solid line shows the estimate for the average effect from the same model. Note that each slope for individual participants (except for one participant with a slope of -.034) has a positive slope. The individual lines differ in length, because they show the predicted values for each participant's z-transformed response range. This figure presents the data gathered for all numerosities after the inflection point (10 - see Data Analysis for details).

Table 2.2 Mixed effects model results for targets and distractors.

	Estimate	*SE*	*tStat*	*DF*	*pValue*	*Lower*	*Upper*
Fixed Effects							
Intercept	-.207	.223	-.928	6135	.353	-.645	.23
Response	.447	.057	7.906	6135	<.001	.337	.558
Random Effects Group: subject (18 Levels)							
	Std	*Lower*	*upper*				
Intercept	.941	.676	.1.311				
Response	.216	.144	.325				
AIC: 25543	*BIC*: 25576 *Log Likelihood*: -12766						

Finally, it is possible that participants tended to increase the variability in their numerical estimates in order to perform better in the second order judgments. To see whether estimation and error monitoring performances were related, we calculated the correlation between individual slopes obtained from the first mixed effects model and the CVs of the numerical estimates (calculated by normalizing each estimate by the corresponding target to obtain a single CV for each participant) for the target numerosities that were included in the model (i.e 11, 16, 22 and 29). There was no correlation between these two variables, $r = -0.246$, $n = 18$, $p = .325$. Similarly, there was also no correlation between the individual slopes from the second mixed effects model and the CVs of estimates for the corresponding numerosities (10:29), $r = -108$, $n = 18$, $p = .671$.

Estimation performance

Several studies have shown that participants underestimate numerosities that are presented simultaneously and non-symbolically (Castronovo & Seron, 2007). In order to see whether this was the case in our experiment, we fit a power function curve to each participant's estimates on the target numerosities (See supplement Chapter 2, Fig S2). The mean exponent of the power function curves was significantly smaller than 1, (M = 0.725, SD = 0.101, $t(17) = -11.565$, p < .001, CI = [0.675, 0.775], d = 2.723, showing that in line with previous work, participants generally underestimated the target numerosities as the magnitude of the numerosity increased.

Discussion

Recent work from our research group has shown that humans can keep track of errors in their temporal reproductions (Akdoğan & Balcı, 2017; see also Kononowicz et al., 2017) and numerical productions (Duyan & Balcı, 2018) pointing at the metric properties of error monitoring as it relates to quantitative representations. However, the characterization of numerical error monitoring has been made solely based on numerical production of target numerosities by terminating sequentially presented stimuli. These limited the study of metric error monitoring to (re)production tasks as well as making the results subject to temporal correlates of numerosities observed by the participants. The current work was designed to overcome these empirical limitations and extended the scope of study of metric error monitoring to estimation (as opposed to production) tasks by testing participants with simultaneously presented array of stimuli. Corroborating our previous results in Duyan and Balcı (2018), the findings of the current study clearly suggest that humans can closely keep track of the direction and magnitude of errors in their numerical estimates. These results combined with our earlier

work suggest that metric error monitoring might apply to many magnitude-based representations irrespective of the way through which they are accessed in the task.

Our results overall set the foundations of a new branch of research what we refer to as "metric error monitoring" that encompasses time and numerosities that have been assumed to be underlain by similar representational systems (e.g., Martin, Wiener, & Van Wassenhove, 2017; Walsh, 2003). This very ability and its information content can have important implications regarding the nature of quantity representations. For instance, these findings are another clear indication for the preservation of metric information in quantity representations (Montemayor & Balci, 2007). Corroborating our conclusions in relation to error-monitoring for magnitude based representations on the neurophysiological grounds, the metric features of errors in the spatial domain was also reported in the error-related event related potential components (i.e., error related negativity [ERN] and error positivity [Pe] - Vocat, Pourtois, & Vuilleumier, 2011). The question of how these components relate to metric error-monitoring ability requires further investigation.

Although our findings constitute very clear demonstration of numerical error monitoring, they do not speak specifically to the mechanisms through which this can be achieved. To this end, the model proposed by Akdoğan and Balcı (2017) to account for temporal error monitoring can be adopted to explain the numerical error monitoring performance as well. According to the generalized (to magnitude based representations) version of this modeling approach, the quantity estimates can rely on one of many simultaneously active generative processes of the same kind; although one of these processes underlies the quantity estimation or the action guided by the quantity estimation, the output of the remaining processes could be used for retrospective comparisons for judging the veridicality of the original estimate. Note that the same computation

can also be achieved by the comparison of the original estimate with a random sample drawn from the long-term memory of the same target (after repurposing the computational assumptions of the Scalar Timing Theory - Gibbon, Church, & Meck, 1984). Such memory-based error-monitoring mechanism would however apply only to those cases where there is one target or targets are far enough from each other to activate only their own memory.

An alternative explanation for numerical error monitoring ability in our task is that each item in the array might activate an "object index" and the total activation of these indices can underlie the numerosity estimate for the corresponding array. In this case, when error-related questioning takes place following the original numerical estimation, the current activation level of these indices can be used to judge the veridacility of the original estimate. Such computations might be possible through the comparison of the current activation of different neurons tuned for different numerosities (e.g., in intraparietal sulcus) for the presented array (Nieder, 2016).

Although the set was presented as an array of items and thus there was no temporal correlate of the set size, it is still possible that participants used response time as a proxy for their confidence ratings. In fact, such a strategy appears to be in place for error monitoring in two alternative forced choice tasks (e.g., Ratcliff, Smith, Brown, & McKoon, 2016). Although such an auxiliary strategy could account for confidence level (i.e., slower RT lower confidence), it would fall short of providing information regarding the directionality of the errors, which we show can be predicted by participants.

We find the extension of metric error-monitoring studies to estimation tasks crucial and complementary to our previous work. The fundamental difference between perception and production tasks is that in perception tasks participants are required to convert nonsymbolic values to symbolic numbers, whereas in production tasks they are required to provide a non-

symbolic response for a target numerosity. This discrepancy between different modes of numerical perception has been theorized to depend on the logarithmic nature of the mental number line, with the assumption that different numerical representations (i.e. symbolic or non-symbolic) are transcoded within the brain. It is further assumed that non-symbolic numerical representations are less precise than symbolic representations, as they are logarithmically compressed along the mental number line (e.g., Dehaene, Izard, Spelke, & Pica, 2008; Crollen et al., 2011; Crollen & Seron, 2012). Combined with our earlier findings, the results of this study show that metric error monitoring performance does not depend uniquely on the nature of transformations required by the task features.

Given earlier studies that imply numerical error-monitoring in children, future studies can focus on the developmental trajectory of metric error monitoring based on the procedures used in the current study. Another related branch of further study regards the trainability of metric error monitoring. For instance, it remains to be unknown if expertise in counting and timing (e.g., via musical training) benefit the error monitoring performance in these quantitative domains.

The autonomy of number processing can be a crucial component of numerical error monitoring. For instance, using a variant of the numerical Stroop task (pitting numerical vs. physical distances against each other) and based on behavioral and neuroimaging data Tang, Critchley, Glaser, Dolan, and Butterworth (2006) showed that the processing requirement of numerical information was higher compared to physical magnitudes, which importantly was not modulated by task relevance of the numerical information. Authors showed an enhanced activation in the parietal lobe while processing numerical distances (irrespective of the task relevance of the numerical information) suggesting that numerical and physical distances were processed differently in the brain (at least in the context of the numerical Stroop task). In the light

of these results providing strong support for the automaticity of numerical processing, the initial subjective estimate might be dependent on autonomous processing, and subjective confidence might rely on the controlled processing of this very same information. Indeed, in perceptual decision making tasks, metacognitive accuracy increases when participants are asked to prioritize speed (Baranski & Petrusic, 1994; Moran, Teodorescu & Usher, 2015; Pleskac & Busemeyer, 2010), suggesting that accumulation of evidence continues even after a decision is made (e.g., Akdoğan & Balcı, 2017). While we cannot claim that decision making under pressure and autonomous processing of magnitudes are analogous, future studies can inspect how metacognitive accuracy (i.e. regression slope in this case) changes when magnitude estimations are made under speed pressure or when the target stimuli is presented for longer durations.

Furthermore, Critchley, Tang, Glaser, Butterworth, and Dolan (2005) found enhanced activity in the dorsal and rostral anterior cingulate cortex during the numerical Stroop task. Coupling the analysis of pupillary response with the analysis of fMRI data, Critchley et al. suggested that ACC could serve integrate error processing and autonomic arousal signals. This later study is indicative of possible autonomic arousal-related and neural signaling correlates of numerical error monitoring. The elucidation of this important integrative functional network possibly underlying numerical error monitoring would benefit from further integrated pupillometry and ERP studies (during numerical estimation) focusing on error-related signals (e.g., Gehring, Goss, Coles, Meyer, & Donchin, 1993) that are source localized to ACC.

Conclusions

The results of the current study show that human participants can keep track of the direction and magnitude of metric errors in their estimates of numerosities based on simultaneously presented array of items in which numerosities are not coupled with temporal correlates. These findings

along with the results of other recent work (Akdoğan & Balcı, 2017; Duyan & Balcı, 2018; Kononowicz et al., 2017) form a strong empirical basis for the metric error-monitoring ability. The neurocognitive mechanisms that underlie metric error monitoring and their relationship with those underlie error-monitoring in the two-alternative forced choice behavior remain to be elucidated.

CHAPTER III

Monitoring Line Length Reproduction Errors

Abstract

Previous work revealed that humans can keep track of the direction and degree of errors in their temporal and numerical reproductions/estimations. Given the behavioral and psychophysical commonalities to various magnitudes and the implication of an overlapping neuroanatomical locus for their representation, we hypothesized that participants would capture the direction of errors and confidence ratings would track the magnitude of errors in line-length reproductions. In two experiments, participants reproduced various target lengths as accurately as possible, and reported the direction of their errors and provided confidence ratings for their reproductions. The isolated analysis of these two second-order judgments showed that participants can correctly report the direction of errors in their line-length reproductions and subjective confidence decreases as the magnitude of errors increases. These results show that humans can robustly keep track of the direction of errors in their line-length reproductions and their subjective confidence corroborates the magnitude of these errors.

Keywords Error monitoring, Line reproduction, Magnitude representations, Confidence judgments

Introduction

Error monitoring is the ability to assess one's own accuracy on a given task in the absence of feedback regarding objective performance. This ability has been typically studied using tasks that require categorical judgments like the two-alternative forced choice task (2AFC) in perceptual decision making and recognition memory domains. These studies showed that error monitoring judgments closely follow the objective performance in humans (Fleming Stephen, Dolan Raymond, & Frith, 2012) however the tasks utilized also imposed paradigmatic constraints on which aspects of error monitoring can be addressed and cannot encompass error monitoring regarding metric estimates about quantities.

Humans and other animals can estimate magnitudes such as time, number and spatial distances accurately on average (e.g., Gallistel, 1990). However, magnitude representations inherently contain uncertainty, which results in trial-to-trial variability in the corresponding quantity estimates. Recent work showed that confidence ratings also reflect the amount of deviation from the target in the estimation of continuous magnitudes such time (Akdoğan & Balcı, 2017), numerical estimates (Duyan & Balcı, 2018, 2019), and grating orientation (Samaha & Postle, 2017). Given the plethora of findings that point to a common system for the representation of magnitudes (Martin, Wiener, & van Wassenhove, 2017; Walsh, 2003), with evidence from studies on cross-modal transfer (e.g., (Balci & Gallistel, 2006)), cross-dimension interference (Henik & Tzelgov, 1982), and neurophysiological findings that relate a common locus (i.e., intraparietal sulcus, e.g., (Bueti & Walsh, 2009)) to various magnitude-based judgments, we hypothesized that confidence ratings and error directionality judgments would also reflect objective performance in the estimation of spatial attributes such as distance. In order

to test this hypothesis, we measured the error directionality judgement and confidence rating of participants in a length reproduction task.

Estimates of spatial distances are crucial in both the planning of simple motor actions, such as deciding how much the related muscles should be contracted when jumping over a hurdle, and the planning of more complex action sequences, such as estimating the time it would take to get to a destination, whether your car will fit an empty spot for parking or whether a new piece of furniture you're planning to buy at the store would fit a vacant space in your apartment. Errors in the estimation of distances might have trivial or dire consequences depending on the context. For instance, overestimating the empty space in your apartment when buying furniture will require an extra trip to the store; underestimating a distance and braking late may result in a car accident. Practically, being aware of general biases (i.e., thinking that you over/underestimate distances), or the magnitude of deviation from the target in a specific estimate would be useful in situations alike (e.g., parking slowly when not certain about the space available; delaying a purchase). Theoretically, the ability to monitor the direction and the degree of errors in trial-to-trial estimates of magnitudes would indicate that humans' error-monitoring system has a resolution that goes above and beyond the categorical judgments regarding the veridicality of the first order binary decisions.

Previous work has shown that humans can estimate lengths accurately (Stevens & Galanter, 1957; Verillo, 1983) with psychophysical patterns similar to other magnitude estimations such as time and numerosity (Petzschner, Glasauer, & Stephan, 2015). Moreover, confidence judgments closely follow the objective performance in sensory discrimination tasks involving judgments of spatial distances (Baranski & Petrusic, 1994, 1999). Recently, Akdoğan and Balcı (2017) reported more direct evidence for metric performance monitoring ability in

interval timing (see also (Kononowicz, Roger, & van Wassenhove, 2018)). Using a temporal reproduction task, they asked the participants to reproduce a target duration as accurately as possible. On each trial, they also obtained confidence ratings and directionality judgments on the accuracy of the reproduced duration. Over four experiments with different durations and different block designs, they found that these confidence ratings and error-directionality judgments (as a composite measure) closely followed the participants' objective timing performance (i.e., parametric and directional errors). Using a numerical analogue of this task, Duyan & Balcı (2018, 2019) showed that such quantitative error-monitoring ability also extends to the numerical estimates based on counts of sequential events as well as based on counts of a simultaneously presented array of circles. Moreover, Samaha and Postle (2017) asked participants to reproduce the orientation of a briefly presented low contrast grating and obtained error monitoring ratings on each trial. Similarly, they observed an inverse relationship between confidence ratings and absolute errors, such that the confidence ratings increased with decreasing mean absolute error (i.e. closer to the actual orientation). These results point to a general ability for the monitoring of errors in tasks that require metric estimations.

In the current study, over two experiments, we examined whether error directionality judgments reflect the direction and confidence judgments reflect the magnitude of the deviations from mean reproductions in the line reproduction task. We presented the participants with three different target lengths for a brief period and on each trial asked them to reproduce the target length as accurately as possible. We subsequently prompted them to provide confidence ratings on and the directionality of error judgments (e.g., shorter or longer than target) in their first-order task performance.

Experiment 1

Method

Participants

Twenty undergraduate students from Koc University participated in Experiment 1. Power analyses were done using the SimR package (Green & MacLeod, 2016; Green, MacLeod, & Alday, 2016). The SimR package estimates statistical power for linear mixed models by randomly sampling a portion of the dataset along with a grouping variable (here, the number of participants). It then fits the same model from the sample data and outputs the proportion of fits that returned a significant effect. Using the data from our study on numerical error monitoring (Duyan & Balcı, 2019), the simulations showed that for the composite analyses, a sample of five participants was sufficient to achieve a power of 0.8. For the isolated linear mixed effects that only include confidence ratings, the simulations showed that 15 participants were sufficient, however, for comparability purposes, we kept the sample size more similar to our previous studies. All participants provided signed informed consent prior to the experiment and received course credit or monetary incentive in return. The study was approved by the local ethics committee at Koç University.

Procedure

Participants were tested in a dimly lit room, seated approximately 50 cm from a 22 in. iMac screen with a 60 Hz. refresh rate. Stimulus presentation and data recording were controlled via Matlab (Mathworks, Natick, MA) using the Psychophysics toolbox (Brainard, 1997) on an iMac. Participants gave their responses by pressing buttons on a mechanical keyboard (Zalman ZM-K500).

Task

There were three target lengths (4.5, 7.5, and 11.5 cm), which were randomly ordered across trials. Each trial began with the presentation of a target line at a random location confined within 80% of the sides of the screen for 100 ms. The entire screen was immediately masked with white noise for 150 ms. Then, the reproduction line, which was initially 0.25 cm, was presented at the center of the screen. The participants had to adjust the length of the reproduced line via left and right arrow keys on the keyboard to shorten or elongate it, respectively. Each key press changed the length of the line in 0.40 cm increments or decrements. Thus, the experimental program did not allow the participants to be exactly on target in any of the trials. The participants then had to press the space key to confirm their response when they thought the reproduced line was as close as possible to the target in that trial. Participants could also skip the trial by pressing 'P' if they had not seen the target, which then would be presented again after a random number of trials. After they confirmed their response, participants were prompted to provide a confidence rating on a 1–3 scale (1 for low, 2 for medium and 3 for high confidence) to indicate how close they thought the reproduced line was to the actual target length. Participants were instructed to try to use the full confidence rating scale. They were then asked if they thought they under-reproduced or over-reproduced the target length in that trial (see Fig. 1). Participants completed a total of 240 trials, where each target was presented in 80 trials (randomly ordered).

Data analysis

We analyzed the data in two different ways. As part of the first approach, we conducted an isolated analysis of the error directionality judgments and confidence ratings. We fit separate linear mixed effects models predicting directionality and confidence judgments from the

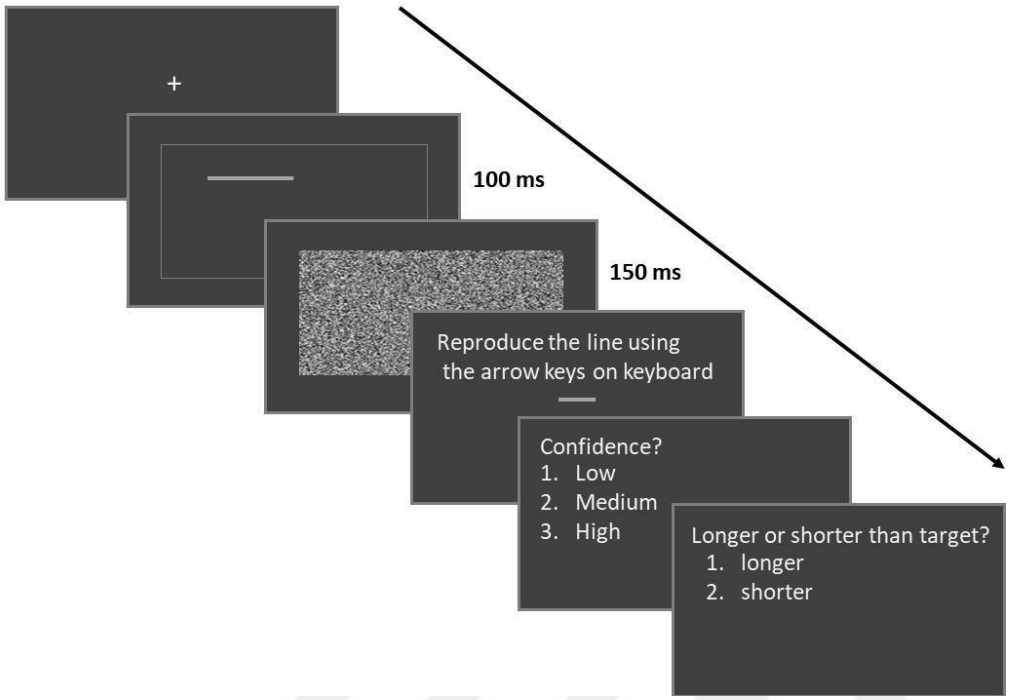

Fig 3.1 An illustration of the experimental task

participants' reproduction performance across different targets. As part of the second approach, we combined these two orthogonal judgments to generate a composite measure to represent the error directionality judgments along with confidence ratings. Note that although in this composite measure confidence ratings are treated as proxies for the error magnitude judgments (an assumption tested by the approach outlined above), in reality, participants were asked to rate their confidence.

We excluded those trials where the participants pressed the space button to confirm their response without adjusting the reproduction line (1.1% of the trials), as the participants would likely have reported low confidence ratings on these trials, biasing the data in favor of our hypothesis. With the same logic, we also excluded the trials where a participant's reproduction response was three mean absolute deviations (MAD) above or below the participant's mean reproduction for a given target (3% of the trials).

For all linear mixed effect analyses, we fitted four different models: first, as the null model (Model 1), we only included an intercept as a fixed effect and also included the participants as a random factor on the intercept. For the second model, we included z-transformed reproductions (absolute z-scores in models where we predict confidence ratings, as it would show the deviance from a participant's mean reproduction) as a fixed effect and included the participants as uncorrelated random effects on the intercept and the slope of these reproductions (Model 2). For the third model, we added the target category as a factor to the previous model, without an interaction term (Model 3). Finally, for the last model, we also included an interaction term between the target category and the corresponding z-scores. In the paper, we report the results gathered from the model that best explains the data based on the BIC scores. Model comparison tables are included in Supplement Chapter 3, Table 3.1 and 3.2. All models were computed using GAMLj module (retrieved from https://www.jamovi.org) in jamovi (The jamovi project, 2019). All Bayesian analyses were done using JASP software (JASP Team, 2018), which uses the Cauchy distribution as priors.

Results

Isolated analysis of confidence ratings and error directionality Judgments

Overall, the mean reproduced lengths were 4.816 cm (*SE* = 0.194) for the short target, 6.635 (*SE* = 0.187) cm for the medium target and 9.906 (*SE* = 0.222) cm for the long target. The participants tended to over-reproduce the short target and under-reproduce the long target, in accordance with the Vierordt's law, which is a common effect observed in temporal reproduction tasks (Lejeune & Wearden, 2009; see Hollingworth, 1910 for size judgments). A one-way repeated measures analysis of variance (ANOVA) revealed that mean confidence (rated 1–3) significantly varied across different lengths, $F(1.5,\ 28.49) = 7.284$, $p = .005$, $\eta^2 = 0.277$; [*M*

(short) = 2.211, *M*(medium) = 2.012, *M*(long) = 2]). Similarly, a one-way repeated measures Bayesian analysis of variance (BANOVA) showed strong evidence for the main effect of length on confidence (BF_{10} = 17.324). Post-hoc tests showed that this effect was due to mean confidence for the short target being higher than the medium (*MD* = 0.2, *SE* = 0.05, *p* = .002, BF_{10} = 45.475) and the long targets (*MD* = 0.212, *SE* = 0.078, *p* = .041, BF_{10} = 3.955).

In order to assess the overall accuracy of *error directionality judgments*, we first transformed each subject's length reproductions for a given target into z-scores. Consequently, we fit a linear mixed effect model predicting error directionality judgments from z-transformed responses across different targets as well as their interactions. We also included participants as uncorrelated effects on the intercept and the z-score slope. Among the four alternative models, Model 4 (which contained the interaction term between z-scores and target length) was the best fit model (i.e., Model 4 vs. Model 3 as the second best model, ΔBIC = 5.4, f^2_{fixed} = 0.129). Hence the formula for our best-fit model was: *Error Directionality Judgment* ~ 1 + Target Category * z-$\text{Score}_{\text{reproduction}}$ + (1 | Participant) + (z-$\text{Score}_{\text{reproduction}}$ | Participant). The model resulted in a significant effect of z-score, β = 0.059, *SE* = 0.013, *p* < .001, showing that error directionality judgments tracked the direction of errors in length reproductions. The model also yielded an effect of target category on the directionality judgments ($F(2,4558) = 73.20$, $p < .001$), showing that the participants' probability of reporting that they overshot the target increased with longer lengths ($\beta_{\text{long-medium}}$ = 0.129, *SE* = 0.017, $p_{\text{bonferroni}} < 0.001$; $\beta_{\text{medium-short}}$ = 0.069, *SE* = 0.017, $p_{\text{bonferroni}} < 0.001$; $\beta_{\text{long-short}}$ = 0.198, SE = 0.017, $p_{\text{bonferroni}} < 0.001$). There was also a significant interaction between target category and reproduction, meaning that the effect of reproduction on error directionality judgments was different across targets ($F(2,4558.4) = 6.12$, $p < .01$). Specifically, simple effects analyses showed that error directionality judgments followed

reproduction performance in the medium (β = 0.08, *SE* = 0.16, $p < .001$) and long targets (β = 0.072, *SE* = 0.16, $p < .001$) but not in the short target (β = 0.026, *SE* = 0.16, $p = .095$).

Second, in order to assess how closely the confidence judgments tracked the participants' deviation from their own mean reproductions, we compared the four alternative models. The model that best fit the data was Model 3 without the interaction term (vs. Model 4 as the second best fit model, $\Delta BIC = 12.9$, $f^2_{\text{fixed}} = 0.16$). Consequently, we fitted linear mixed models predicting confidence from absolute values of z-transformed reproductions as a fixed effect on confidence across different targets and included the participants as a random effect on the slope and intercept. The formula for the best fit model was *Confidence* ~ 1 + Target Category + Absolute z-Score$_{\text{reproduction}}$ + (1 | Participant) + (Absolute z-Score$_{\text{reproduction}}$ | Participant). The model yielded a significant effect of absolute errors on confidence, ($F(1,20.9) = 6.37$, $\beta = -0.047$, *SE* = 0.019, $p = .02$). The model also showed a significant effect of target on overall confidence ($F(2,4578) = 49.03$, $p < .001$). Post-hoc comparisons showed that the overall confidence significantly lower for the medium and long targets than for the shortest target ($\beta_{\text{long-short}} = -0.214$, *SE* = 0.024, $p_{\text{bonferroni}} < 0.001$; $\beta_{\text{medium-short}} = -0.202$, *SE* = 0.024, $p_{\text{bonferroni}} < 0.001$).

Composite measure analysis

For each participant, we first categorized confidence ratings and directionality judgments in each trial under six groups: under-low (UL), under-medium (UM), under-high (UH), over-high (OH), over-medium (OM), over-low (OL). These confidence-directionality judgment pairs were numerically coded as −3, −2, −1, 1, 2 and 3, respectively (Note that these were treated as an ordinal variable in the analysis). In this measure, confidence ratings were treated as a proxy for participants' judgment on how closely they reproduced the target length, an assumption that was

confirmed by our analysis of the relationship between confidence ratings and absolute magnitude of error (see above). The composite measure combined this variable with the directionality judgments (i.e. over/ under judgment) on whether they overshot or undershot the target. In order to assess "spatial error-monitoring" performance based on the composite measure, we fit mixed linear models to predict the six confidence-directionality judgment pairs (as an ordinal variable) from the participants' reproduction performance across different target lengths. Consequently, significantly positive slopes would indicate an ability to monitor the degree (again using confidence ratings as a proxy) and the directionality of errors in reproduced lengths. Again, we also included participant's as uncorrelated random effects on the intercept and the slope. Model comparison statistics showed that Model 4 was the best-fit model (compared Model 3 as the second best fit model, $\mathbf{\Delta}BIC = 3.4$, $f^2_{\text{fixed}} = 0.01$). The formula for the best fitting model was *Confidence-Directionality Judgment* ~ 1 + Target Category * z-Score$_{\text{reproduction}}$ + (1 | Participant) + (z-Score$_{\text{reproduction}}$ | Participant). In line with our hypothesis, the model showed a significant effect of reproduction performance on confidence-directionality judgments ($F(1,19.9) = 24.9$, $\beta = 0.245$, $SE = 0.028$, $p < .001$). The model also revealed significant interactions between target categories and z-transformed reproductions ($F(2,4554.5) = 10.2$, $p < .001$; $\beta_{(\text{long-short})}$ * z-score = 0.276, $SE = 0.07$; $p < .001$; $\beta_{(\text{medium-short})}$ * z-score = 0.27, $SE = 0.07$; $p < .001$). Simple effects analyses showed that these confidence-directionality judgments followed the magnitude and the direction of errors in the medium ($\beta = 0.339$, $SE = 0.49$, $p_{\text{bonferroni}} < 0.001$) and long targets ($\beta = 0.333$, $SE = 0.49$, $p_{\text{bonferroni}} < 0.001$), but not in the short target ($\beta = 0.063$, $SE = 0.49$, $p_{\text{bonferroni}} = 0.204$). There was also a main effect of target category on the composite score, $F(2,4554) = 58.9$, $p < .001$; $\beta_{\text{medium-long}} = -0.49$, $SE = 0.07$, $p_{bonferroni} < 0.001$; $\beta_{\text{short-medium}} = -0.25$, $SE = 0.07$, $p_{bonferroni} = 0.001$; $\beta_{\text{short-long}} = -0.74$, $SE = 0.07$, $p_{\text{bonferroni}} < 0.001$) (for a summary of the results,

see Table 3.1 and Fig. 3.2). The details of model outputs for Experiment 1 are presented in the Supplemental Online Material S1.

	$\beta_{z\text{-}score}$	*Standard Error*	*df*	*t value*	*p*	*CI (%95)* Lower Upper
Experiment 1						
Model						
Direction judgment	*.059*	*.012*	*20*	*4.89*	*< .001*	*.036; .102*
Confidence rating*	*-.047*	*.018*	*20.9*	*-2.52*	*=.020*	*-.083; -.011*
Composite score	*.245*	*.05*	*19.9*	*3.859*	*< .001*	*.149; .341*
Experiment 2						
Model						
Direction judgment	.07	.012	21	5.87	<.001	.047; .093
Confidence rating*	*-.065*	*.021*	*21.5*	*-3.17*	*=.005*	*-.106; -.025*
Composite score	.283	.052	20	5.44	<.001	.181-.384

Table 3.1 Main results of the mixed effect models in both experiments. *Note that for the confidence model, we used absolute z-scores to predict confidence ratings as they would reflect the amount of deviation from the target, irrespective of the direction of errors. Hence, a negative slope indicates that confidence ratings decreased with higher deviations from the target.

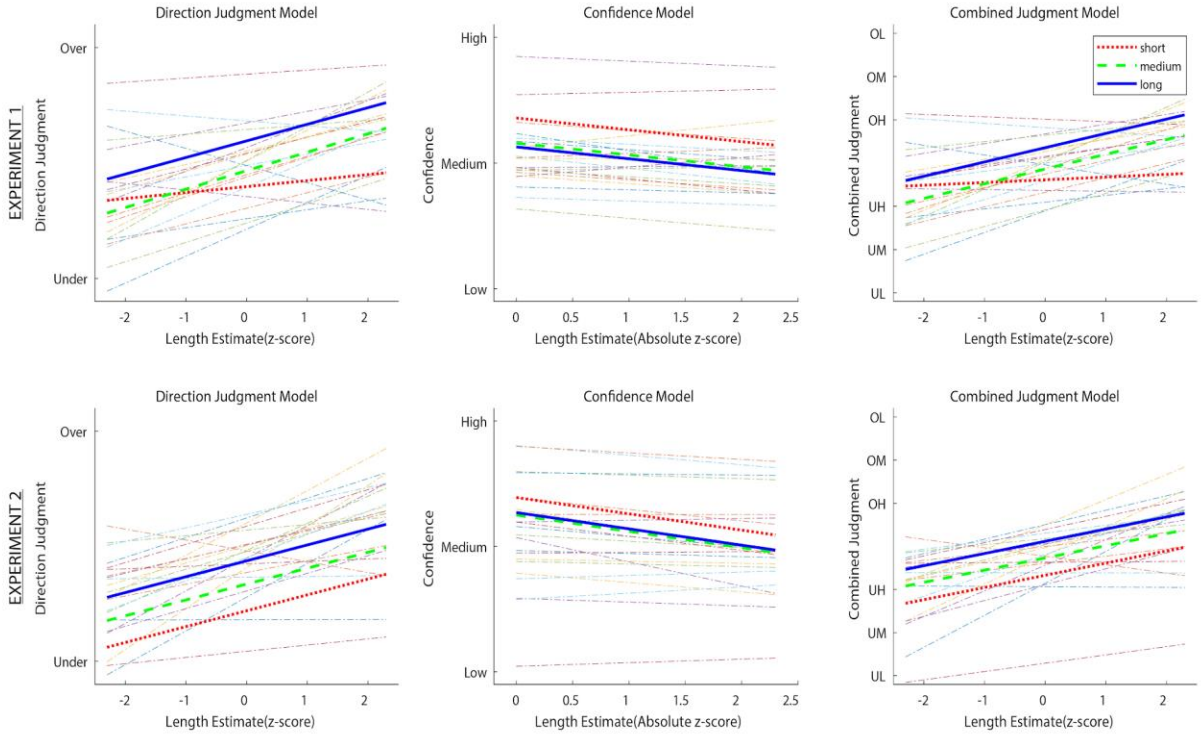

Fig 3.2 Results of the mixed effects model fits. We fit linear mixed effects models to see the effect of z-transformed estimates (across different target lengths) on directionality judgments, confidence ratings and their composite scores, respectively. We included individual participants as uncorrelated random effects on the slope and the intercept. The regression lines from the models are depicted with a bold line. Colored-dashed lines show the group-level (i.e., participants) estimates across all target lengths.

Finally, a potential risk underlying these results is that the participants could increase the variability in their responses to increase their performance in the second order judgments. To see if there was a relationship between estimation and error monitoring performance, we calculated the correlations between the coefficients of variation (CVs) of the length estimates and the individual slopes obtained from the mixed effects model. An individual's CV is calculated by

dividing the standard deviation of their response distribution by its mean. In this sense, CV is a measure of the overall precision of one's subjective estimates where a lower CV indicates better overall performance (i.e., higher precision). The individual slopes obtained from the mixed effects model that predicts confidence from absolute deviations is indicative of how close a participants' confidence ratings track their deviation from their own means, as it shows how much confidence judgments account for the variability in an individual's responses. Hence, a significant negative correlation between individual CVs and slopes would raise the possibility that the participants increased the variability of their line reproductions by intentionally over- or undershooting the target to perform better in their error monitoring judgments. In order to test this possibility, we calculated a pooled CV for each participant's reproduction scores for all targets and investigated its relationship with individual slopes. Refuting this possibility, we found that there was no significant correlation between these measures $r = 0.338$, $p = .145$, $BF_{10} = 0.748$ (see Supplement Chapter 3, S4 for the results of the analogous analysis for confidence ratings).

Given the lack of metric error-monitoring ability for the shortest length (for both error-directionality judgments and composite measures) in the first experiment, we found it necessary to test if this observation was specific to the shortest anchor of the test set or to the absolute test line length itself. Thus, we tested another group of participants with a test set, the shortest value of which was equal to the middle value of the test set in Experiment 1. The second reason behind running the second experiment was to replicate the first study but by giving participants the opportunity to be exactly on target. In Experiment 1, the incrementations were set such that participants could never be exactly accurate.

Experiment 2

Method

Participants

Twenty-one undergraduate students from Koç University participated in the experiment.

Procedure

The procedure and the task were very similar to Experiment 1. The target lengths were 7.5, 11.5 and 16.5 cm and each adjustment elongated the reproduction line by 0.25 cm increments; which allowed participants to reproduce the target line to its exact length. Trials where the participants did not adjust the target line (1.3% of all trials) and where the reproduced line was 3 MADs above or below a participant's mean reproduction for that target (3.1% of all trials) were excluded from all analyses.

Results

Isolated analysis of confidence ratings and error directionality judgments

The mean reproduced lengths were 7.013 (*SE* = 0.201) cm, 9.797 (*SE* = 0.307), and 12.855 (*SE* = 0.387) for the short, medium and long targets, respectively. Similar to Experiment 1, a repeated measures ANOVA showed there was an effect of target length on overall confidence, $F(1.196, 23.925) = 4.548$, $p = .017$, $\eta^2 = 0.185$. Also, one-way repeated measures BANOVA found moderate evidence ($BF_{10} = 3.03$) for such an effect. Post hoc tests revealed that mean confidence for the short target was higher than the medium target ($MD = 0.132$, $SE = 0.046$, $p = .027$, $BF_{10} = 5.556$).

Different from Experiment 1, the model that best fit the data was Model 3 (vs. Model 4 as the second best fit model, $\Delta BIC = 16$, $f^2_{\text{fixed}} = 0.103$). The best fitting model was *Error Directionality Judgment* ~ 1 + Target Category + z-Score$_{\text{reproduction}}$ + (1 | Participant) + (z-

$Score_{reproduction}$ | Participant). The model yielded significant effect of z-transformed reproductions ($F(1,21) = 34.5$, $\beta = 0.07$, $SE = 0.012$, $p < .001$). The model also showed that participants' probability of reporting that they overshot the target increased with longer targets ($F(2,4758.1) = 89.6$, $p < .001$; $\beta_{long\text{-}medium} = 0.102$, $SE = 0.016$, $p < .001$; $\beta_{medium\text{-}short} = 0.116$, $SE = 0.016$, $p < .001$; $\beta_{long\text{-}short} = 0.218$, SE = 0.34, $p < .001$). As for confidence ratings, the best-fit model was again Model 3 (vs. Model 4 as the second best fit model, ΔBIC=15.3, $f^2_{fixed} = 0.618$). Consequently, we included the absolute values of z-transformed reproductions and target category as fixed effects on confidence, and including participants as uncorrelated random effects on the slope and the intercept. The formula for the best fit model was: *Confidence* ~ 1 + Target Category + Absolute z-$Score_{reproduction}$ + (1 | Participant) + (Absolute z-$Score_{reproduction}$ | Participant). The results showed that subjective confidence overall followed the degree of absolute errors ($F(1, 21.5) = 10$; $\beta = 0.065$, $SE = 0.002$, $p = .005$), and overall confidence level varied across targets ($F(2,4757.1) = 28.2$, $p < .001$; $\beta_{short\text{-}medium} = 0.136$, $SE = 0.02$, $p < .001$; $\beta_{short\text{-}long} = 0.123$, SE=0.02, $p < .001$). Overall confidence was higher for shortest target length compared to the other target lengths.

Composite measure analysis

In order to assess spatial error monitoring performance based on the composite measure (using confidence ratings as a proxy for error magnitude judgements as confirmed by the analysis above), we fitted the same linear mixed model (Model 3 as the best fit model compared to Model 4 as the second best fit model, $\Delta BIC = 16$, $f^2_{fixed} = 0.149$) that predicts signed error monitoring from z-transformed length reproductions, including participants as independent random effect terms on the intercept and the slope. Similarly, the model showed a statistically significant positive slopes for reproduced lengths ($F(1,20) = 29.6$, $\beta = 0.283$, $SE = 0.052$, $p < .001$). There

was a significant effect of target category on the composite score ($F(2,4756.1) = 74$, $p < .001$; $\beta_{\text{Short-Long}} = -0.788$, $SE = 0.065$, $p < .001$; $\beta_{\text{Medium-Long}} = -0.39$, $SE = 0.065$, $p < .001$; $\beta_{\text{Short-Medium}} = -0.4$, $SE = 0.065$, $p < .001$) (for a summary of the results, see Table 1 and Fig. 2). The details of model outputs for Experiment 2 are presented in the Supplemental Online Material S2. Finally, there was no significant correlation between individual slopes and pooled CVs, $r = 0.08$, $p = .737$, $BF_{10} = 0.292$ (see Supplement Chapter 3, S4 for the results of the analogous analysis for confidence ratings).

Discussion

In the current study, we have asked participants to reproduce the length of briefly presented lines and obtained confidence ratings and error-directionality judgments on their trial-to-trial line length reproduction performance. Our findings showed that in the absence of explicit feedback, humans can monitor the direction of errors in their reproductions of lengths (for nearly all targets) and their confidence ratings can track the magnitude of their errors for all target lengths. Together with our previous findings which indicate that humans can better than chance guess the direction and match their confidence to the degree of errors (based on composite measure) in their temporal (Akdoğan & Balcı, 2017; see also Doenyas, Mutluer, Genc, & Balcı, 2019; Kononowicz et al., 2018) and numerical (Duyan & Balcı, 2018) reproductions and numerical estimations (Duyan & Balcı, 2019), the results of Samaha and Postle (2017) that point toward a similar ability in relation to orientation judgment errors, and given converging evidence for a general magnitude representation system in the brain (Bueti & Walsh, 2009; Walsh, 2003), we surmise that this error monitoring ability would most likely extend to other metric domains.

This new line of studies conducted in the domains of time, numerosity, and spatial distances suggest that error monitoring regarding the direction of metric errors and confidence

matching to the magnitude of errors are likely shared metacognitive features of magnitude representations. One of the questions that arise from these findings relates to the information processing basis of this generalized metric error monitoring ability. To this end, one possibility is that the generative process that underlies magnitude estimations might contain more information value than what is manifested in the form of reproduction or estimate. To this end, Akdoğan and Balcı (2017) proposed that multiple generative processes are realized during the task however and partially depending on the task representation (i.e., requirement to provide point estimates), only one of these generative processes is manifested in the form of the behavioral output. In such an in-parallel information processing scheme, participants can compare their estimate to the other estimate(s) that would have been made if they had relied on another process/other processes in a retrospective fashion and this comparison can inform the agent regarding the direction and magnitude of their errors.

Duyan and Balcı (2018) proposed another comparison strategy in which participants would compare their current estimate of magnitude with a random sample from their long-term memory representation for that target. In fact, such a comparison strategy forms the basis of the decision stage of the Scalar Timing Theory to guide the first order timing performance (Gibbon, Church, & Meck, 1984). But note that this comparison strategy would work with single target tasks and for tasks that contain multiple targets it would require participants to know the corresponding memory representation for the currently targeted quantity for which the estimate has been made. This can be done based on the likelihood comparisons of the current reading under different memory representations that correspond to different previously experienced targets. Similar theoretical approaches can be applied to the line length reproduction/spatial metric error monitoring task used in the current study. In any case, future studies are needed to

fully test for a generalized metric error monitoring system based on the formal analysis of patterns across error monitoring performances in different magnitude domains.

Earlier studies show that scalar variability (as the manifestation of representational uncertainty) is observed in the processing of different magnitudes such as time, number, and length (e.g., Droit-Volet, Clement & Fayol, 2007). In light of this information, one of the potential reasons behind the lack of the metric error monitoring ability for the shortest length in Experiment 1 could be the stronger manifestation of non-scalar/and thus presumably non-representational variability in the reproductions for the shortest target length (akin to generalized Weber's Law; Treisman, 1964). In order to address this possibility, we tested if CVs for different target lengths differed within each experiment. We did not find a significant difference in Experiment 2 ($F(2,40) = 2.602$, $p = .09$) but there was a statistically significant overall difference between the CVs of different line lengths in Experiment 1 ($F(2,38) = 4.806$, $p = .014$, $\eta^2 = 0.202$, $BF_{10} = 3.703$). Post-hoc analysis among the CV's for the three target lengths (after Bonferroni correction) showed that the CV for the shortest line length was higher than the CV for the longest target length ($p < .05$). Thus, in line with our speculation, the effect of non-scalar (non-representational) sources of variability appears to have stronger manifestation than the scalar sources of representational uncertainty for the shortest target length of Experiment 1. If metric error monitoring (including confidence ratings as a proxy for error magnitude rating) relies primarily on representational scalar uncertainty, the stronger manifestation of non-scalar sources of variability could indeed limit the metric error monitoring performance (for a similar argument in animal decision-making see (Berkay, Freestone, & Balcı, 2016)). Future studies are needed to fully address this possibility. Despite the higher CVs for short target lengths, we also observed that overall confidence was higher for the short target compared to medium and long targets in

Experiment 1, indicating that the participants made more 'high confidence' judgments for this target (possibly due to the dependence of subjective confidence level on the absolute variability). As a result, because high confidence judgments are coded as either −1 or 1, this would constrict the range of predicted values around zero, such that the reproduced values could not account for the lower variability in error monitoring judgments (e.g., based on composite measure). This could be another reason for not observing metric error monitoring ability for the shortest target length in Experiment 1 (however note that confidence was higher also for the shortest target length of Experiment 2 for which we observed metric error monitoring).

As outlined above, partially different findings gathered from two experiments could be due to the fact that the absolute error was higher for the longer target lengths in the second experiment simply due to the scalar property. Furthermore, given the intermixing of different target lengths during testing, it is possible that participants adopted a single criterion as an aggregate representation (Balcı et al., 2011; Gorea & Sagi, 2000; Gorea, Caetta, & Sagi, 2005; Rahnev & Denison, 2018). These two facts would result in lower subjective error estimates for the aggregate representation of the target lengths in Experiment 1 compared to the aggregate representation of the target lengths in Experiment 2. This resultant divergence between the two experiments in terms of the levels of subjective uncertainty could have resulted in the relatively and partially weaker association between the participants' error judgments and the objective errors in Experiment 1.

We analyzed the set of second order judgments collected in this study based on two different approaches. In the first approach, we first assessed if participants could report whether their reproductions were shorter or longer than the target (error directionality judgments) and then we characterized the relationship between the confidence ratings and absolute magnitude of

metric errors. Both the error directionality judgements and confidence ratings nicely matched the actual direction (except for the shortest target length of Experiment 1) and the magnitude of errors, respectively. Particularly, the second finding suggested that confidence ratings could be utilized as a proxy for the error magnitude judgments (although participants were asked to provide confidence ratings). Thus, in the second approach, we combined error directionality judgments and confidence ratings to calculate a composite variable (akin to our earlier work on metric error monitoring in timing and counting). The results of our analysis of the composite variable supported the existence of metric error monitoring ability. That being said, assessing the direction and magnitude of errors by asking participants to guess where their reproduction fell in relation to the target (e.g., on a continuum) would constitute a more direct way of investigating metric error monitoring and thus should be used in future work. Given the lack of a relationship between error directionality judgments and objective performance in the shortest target in Experiment 1, we can infer that error directionality judgments and confidence ratings might rely on different processes and that the composite error-directionality ratings may in part be inherited from the error directionality judgments.

In the current study, participants reproduced the targeted lengths with multiple keypresses that were self-paced and they could shorten the reproduced length in addition to lengthening it in an individual trial. Given these potentially informative features of behavioral testing, we also studied the pattern of key pressing that led to the ultimate line reproduction in order to investigate if the observed patterns could be informative regarding the underlying error-related information processing. To this end, we looked at the time it took participants to confirm their line reproduction response (i.e., the delay between the last line reproduction-related keypress and the keypress to finalize/confirm the reproduction), as well as the frequency of the switches between

the key presses used to lengthen and shorten the reproduced line length. We did not observe any meaningful patterns in either of these measures suggesting that the error judgment was indeed made in a retrospective fashion at least in the current task representation. Thus, this very observation suggests that the metric error information relied on during interrogation for performance monitoring was likely not used to adaptively guide the line reproduction earlier in the same trial.

Although it is difficult to attribute our findings to a meta-cognitive ability per se, we believe that our results would have important implications for theories of metacognition and theories on the magnitude representation system and would constitute a fruitful empirical ground for their theoretical integration in future research. Future studies can also focus on neuroscientific approaches in order to reach an overarching understanding of error-monitoring and confidence ratings in different tasks; to this end event-related potentials (ERP) such as error-related negativity (ERN) and error-positivity (Pe) are logical targets (e.g., see Di Gregorio et al., 2018; Falkenstein, Hoormann, Christ, & Hohnsbein, 2000; Taylor, Stern, & Gehring, 2007). Although these ERP components are traditionally looked at to investigate all-or-none effects as in the case of two-alternative forced-choice scenarios, our findings suggest that they can be modulated in a graded fashion as a function of the magnitude of errors. This prediction is also supported by a number of studies in which participants could observe the degree of their errors in sensorimotor tasks.

For instance, a recent EEG study investigating the ERP components of participants observing their avatar's behavior through virtual reality showed that ERN but not Pe was modulated in a graded fashion (as opposed to exhibiting an all-or-none character) as a function of the magnitude of the observed errors in the avatar's action in relation to a target (Spinelli, Tieri,

Pavone, & Aglioti, 2018). In an earlier study, Vocat, Pourtois, and Vuilleumier (2011) tested participants in a ballistic pointing to a target task with optical prisms to induce gross motor errors that participants can adapt to over time as well as without optical prisms. They observed that both ERN and Pe amplitude were parametrically modulated with the magnitude of deviations of the pointing endpoints from the visual target (see also Luft, Takase, & Bhattacharya, 2014; Torrecillos, Albouy, Brochier, & Malfait, 2014 regarding the graded modulation of relevant ERP signals). These electrophysiological findings point at the possibility that errors in magnitude estimations can also be encoded via the same cognitive architecture that underlies binary error-related ERP components. The extension of electrophysiological investigations to metric error monitoring (in the absence of feedback) would contribute to our understanding of the neural information-processing correlates of this ability. A recent study by Kononowicz et al. (2018) have taken an empirical step in this direction and showed that temporal error monitoring performance was predicted by the brain oscillatory signals that emerge early during the timing of the event to be judged. It remains to be answered if similar dynamics would also emerge in the case of numerical and spatial error monitoring despite differences in the way they are experienced.

GENERAL DISCUSSION

Error monitoring refers to the ability to judge one's own objective performance in the absence of explicit feedback. Human error monitoring abilities have been shown in various domains such as memory and perceptual decision making. However, until recently, this ability was studied almost exclusively with categorical choice tasks. Recent evidence (Akdoğan & Balcı, 2017) from interval timing domain suggests that humans can also accurately judge their objective performance in their timing estimates. This finding suggests that human error monitoring abilities extend beyond categorical decisions and that humans also have metacognitive access to their errors in magnitude representations. Given this finding, we investigated humans' error monitoring abilities in magnitude representations. Overall, our results point at a ubiquitous error monitoring ability in humans that span across multiple domains, in line with the theoretical accounts that suggest a common representational system for magnitudes (Walsh, 2003). Moreover, it also suggests that metacognitive skills in humans have a richer informational basis than models of error monitoring suggest.

In the first study, we tested error monitoring in number reproductions with a sequential non-verbal counting task. Our results showed that humans can monitor the direction and the degree of errors in their numerosity estimations, over and beyond its RT correlates. On the other hand, we observed no relationship between the precision of numerosity reproductions and error monitoring performance. These results pointed toward a general error monitoring mechanism in magnitude representations that is independent of first order performance.

In the second study, we employed a numerosity estimation based on a simultaneously presented array of dots. This way, we eliminated the potential effects of time in counting, and tested the participants on a wider range of target numerosities. The results provided further

evidence that humans can track the direction and the degree of their errors in numerical estimates. Importantly, these studies on numerical error monitoring constitute two different modes of numerical processing involving the symbolic and non-symbolic representations, namely, reproduction and estimation. In the reproduction task (Chapter 1), participants are required to provide a non-symbolic response for a symbolic target, whereas in the estimation task (Chapter 2) they are required to provide a symbolic estimate for a non-symbolic numerosity. While these two modes of representation are differ in terms of their precision, our results suggests that numerical error monitoring is independent of the direction of this conversion.

In the third study, we found that metric error monitoring ability also extends to the reproductions of spatial length. In addition, we analyzed the relationship between the second order judgments and line reproduction performance based on two different approaches. In the first approach, the results showed that the error-directionality judgments closely matched the actual direction of errors, except for the shortest target in Experiment 1. The second approach revealed that subjective confidence increased as the absolute magnitude of errors decreased. These results suggest that error monitoring abilities are robust to different types of judgments of performance.

Future directions

The most studied biomarker of errors in the human brain is the *error related negativity* (ERN), an event-related potential that is marked by a negative deflection that is observed 80-150 ms following errors (Gehring et al., 1993). The ERN emerges from the anterior cingulate cortex (ACC) and occurs immediately after an erroneous motor action and before external feedback (Gehring et al. 1993). Although ERN has been predominantly studied with all-or-none responses,

recent findings suggest that it can be modulated in a graded fashion (Spinelli et al., 2018; Vocat et al. 2011). Future studies can investigate the ERP components of error monitoring to observe whether errors of magnitude estimations elicit a graded ERN response (although for negative results see Kononowicz & van Wassenhove, 2019).

Lastly, while it is evident that humans can represent and use uncertainty estimates in memory, learning and cognition, the neural principles that give rise to these computations are unclear. However, there's apparently a neurological dissociation between domains, specifically between memory and perception (Fleming et al., 2014; Baird et al., 2013; McCurdy et al., 2013). For instance, lesions to the anterior prefrontal cortex selectively disrupt metacognitive accuracy in perceptual decisions, while leaving metacognitive skills in memory intact (Fleming et al., 2014). Gray matter volume in the anterior prefrontal cortex (aPFC) for the individual differences in metacognitive performance in visual discrimination, whereas the grey matter volume in medial parietal cortex accounts for the individual differences in the accuracy of confidence judgments in a recognition memory task (Fleming et al., 2014). Considering this evidence, it has been proposed that different subregions of the aPFC might be involved in the online monitoring of perceptual processes and the information stored in memory. In support of this claim, Baird et al. (2013) found that greater connectivity between various subcortical areas and lateral aPFC and medial aPFC differentially predicted metacognitive accuracy for perceptual decisions and memory retrieval, respectively. Error monitoring in sequentially and momentarily presented numerosities might allow for a dissociation between these distinct circuitries within the same domain, as they tap into different modes of memory. Simultaneously presented arrays require the retrieval of numerousity representation from sensory (iconic) memory (Sperling, 1983), while sequentially presented requires a continuous update of the same representation in working

memory (Baddeley, 1983). Hence, metacognitive circutries underlying perceptual and mnemonic processing in lateral aPFC and medial aPFC can be differentiated within the same domain.

In conclusion, the presented line of work offers a novel approach to observe the behavioral correlates of error monitoring and potentially discover its underlying neural basis by utilizing the continuous and variable nature of magnitude representations. This approach is novel in a similar sense to the transition from accounting for the speed and accuracy of a binary decision to accounting for the confidence in that decision. Only now, we are setting out to account for confidence and error judgments in magnitude representations, where there are varying degrees of accuracy instead of correct or incorrect decisions. This way, behavioral and neural correlates of error monitoring can be determined in a more parametric fashion. The proposed line of work lays the groundwork for a new approach understanding of neural error monitoring (and signaling) that are observed ubiquitously across a wide range of functions from low-level motor learning to the more cognitively taxing multisensory decision making.

References

Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition, 146*, 377–386. https://doi.org/10.1016/j.cognition.2015.10.006

Akdoğan, B., & Balcı, 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

Arrighi, R., Togoli, I., & Burr, D. C. (2014). A generalized sense of number. *Proceedings of the Royal Society. 281*: e16161. https://doi.org/10.1098/rspb.2014.1791

Arzy, S., Adi-Japha, E., & Blanke, O. (2009). The mental timeline. An analogue of the mental number line in the mapping of life events. Consciousness and Cognition, 18, 781–785, https://doi.org/10.1016/j.concog.2009.05.007.

Arzy, S., Collette, S., Ionta, S., Fornari, E., & Blanke, O. (2009). Subjective mental time: The functional architecture of projecting the self to past and future. European Journal of Neuroscience, 30, 2009–2017, https://doi.org/10.1111/j.1460-9568.2009.06974.x.

Baddeley, A. D. (1983). Working memory. *Philosophical Transactions of the Royal Society of* London. B, Biological Sciences, 302(1110), 311-324. https://doi.org/10.1098/rstb.1983.0057

Baird, B., Smallwood, J., Gorgolewski, K. J., & Margulies, D. S. (2013). Medial and Lateral Networks in Anterior Prefrontal Cortex Support Metacognitive Ability for Memory and Perception. *Journal of Neuroscience*, *33*(42), 16657–16665. https://doi.org/10.1523/JNEUROSCI.0786-13.2013

Balakrishnan, J. D., & Ratcliff, R. (1996). Testing models of decision making using confidence ratings in classification. *Journal of Experimental Psychology. Human Perception and Performance*, *22*(3), 615–633.

Balci, F., & Gallistel, C. R. (2006). Cross-domain transfer of quantitative discriminations: Is it all a matter of proportion?. *Psychonomic Bulletin & Review*, *13*(*4*), 636-642. https://doi.org/10.3758/BF03193974.

Balcı, F., Simen, P., Niyogi, R., Saxe, A., Hughes, J., Holmes, P., & Cohen, J. D. (2011a). Acquisition of decision-making criteria: Reward rate ultimately beats accuracy. *Attention, Perception, & Psychophysics, 73(2)*, 640-657. https://doi.org/10.3758/s13414-010-0049-7

Balcı, F., Freestone, D., Simen, P., deSouza, L., Cohen, J. D., & Holmes, P. (2011b). Optimal Temporal Risk Assessment. *Frontiers in Integrative Neuroscience, 5(September)*, 56. https://doi.org/10.3389/fnint.2011.00056.

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & psychophysics*, *55*(*4*), 412-428. https://doi.org/10.3758/BF03205299

Baranski, J. V, & Petrusic, W. M. (1998). Probing the locus of confidence judgments: experiments on the time to determine confidence. *Journal of Experimental Psychology. Human Perception and Performance*, *24*(3), 929–945. https://doi.org/10.1037/0096-1523.24.3.929

Baranski, J. V., & Petrusic, W. M. (1999). Realism of confidence in sensory discrimination. *Perception & Psychophysics*, *61*(*7*), 1369-1383. https://doi.org/10.3758/BF03206187

Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition, 86(3)*, 201-21. https://doi.org/10.1016/S0010-0277(02)00178-6

Berkay, D., Freestone, D., & Balcı, F. (2016). Mice and rats fail to integrate exogenous timing noise into their time-based decisions. Animal cognition, 19(6), 1215-1225. https://doi.org/10.1007/s10071-016-1033-y

Bornstein, B. H., & Zickafoose, D. J. (1999). "I know I know it, I know I saw it": The stability of the confidence-accurate relationship across domains. *Journal of Experimental Psychology: Applied*, *5*(1), 76–88.

Brainard, D. H. (1997). The Psychophysics toolbox. Spatial Vision, *10*, 433–436. http://dx.doi.org/10.1163/156856897X00357.

Brocas, I., Carrillo, J. D., & Tarraso, J. (2018). Self Awareness of Biases in Time Perception. *Journal of Economic Behavior and Organization, 148*, 1-19. https://doi.org/10.1016/j.jebo.2018.02.001

Brown, S. D., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review*, *112*(1), 117–128. https://doi.org/10.1037/0033-295X.112.1.117

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Bueti, D., & Walsh, V. (2009). The parietal cortex and the representation of time, space, number and other magnitudes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(*1525*), 1831-1840. https://doi.org/10.1098/rstb.2009.0028

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

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

Critchley, H. D., Tang, J., Glaser, D., Butterworth, B., & Dolan, R. J. (2005). Anterior cingulate activity during error and autonomic response. *Neuroimage*, 27(*4*), 885-895. https://doi.org/10.1016/j.neuroimage.2005.05.047

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(1)*, 39. https://doi.org/10.1027/1618-3169/a000064

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.

Çavdaroğlu, B., & Balcı, 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.

de Gardelle, V., & Mamassian, P. (2014). Does Confidence Use a Common Currency Across Two Visual Tasks? *Psychological Science*, *25*(6), 1286–1288. https://doi.org/10.1177/0956797614528956

de Gardelle, V., Le Corre, F., & Mamassian, P. (2016). Confidence as a Common Currency between Vision and Audition. *PLOS ONE*, *11*(1), e0147901. https://doi.org/10.1371/journal.pone.0147901

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and numerical magnitude. Journal of Experimental Psychology. General, 122,371–396

Dehaene, S., Izard, V., Spelke, E., & 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

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

DeWind, N. K., & Brannon, E. M. (2012). Malleability of the approximate number system: effects of feedback and training. *Frontiers in Human Neuroscience*, *6*(April), 1–10. https://doi.org/10.3389/fnhum.2012.00068

Dougherty, M. R. P. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *Journal of Experimental Psychology: General*, 130, 579–599.

Dormal, V., Andres, M., Dormal, G., & Pesenti, M. (2010). Mode-dependent and mode-independent representations of numerosity in the right intraparietal sulcus. *Neuroimage*, *52*, 1677–1686. https://doi.org/10.1016/j.neuroimage.2010.04.254

Duyan, Y. A., & Balcı, F. (2018). Numerical error monitoring. *Psychonomic Bulletin & Review*, *25*(*4*), 1549-1555. https://doi.org/10.3758/s13423-018-1506-x

Duyan, Y. A., & Balcı, F. (2019). Metric error monitoring in the numerical estimates. *Consciousness and Cognition*, *67*, 69-76. https://doi.org/10.1016/j.concog.2018.11.011

Egan, J. P. (1958). Recognition memory and the operating characteristic (Tech. Rep. No. AFCRC-TN-58–51). Bloomington, IN: Hearing and Communication Laboratory, Indiana University

Fabbri,M., & Natale, V. (2009).Does the ATOM(A Theory Of Magnitude)model represent an advance in psychological research? In A.M.Columbus (Ed.), Advances in Psychological Research, Volume 62.(pp. 83–111) New York: Nova Science Publishers, Inc.

Falkenstein, M., Hoormann, J., Christ, S., & Hohnsbein, J. (2000). ERP components on reaction errors and their functional significance: a tutorial. *Biological Psychology*, *51*(*2-3*), 87-107. https://doi.org/10.1016/S0301-0511(99)00031-9

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*(July), 1–9. https://doi.org/10.3389/fnhum.2014.00443

Fraisse, P. (1963). *Psychology of Time*. New York: Harper & Row.

Freestone, D., & Balci, F. (2017). The Biological Basis of Economic Choice. In V. Tucci (Ed.), *Handbook of Neurobehavioral Genetics and Phenotyping* (pp. 143-178) John Wiley & Sons, Inc., Hoboken, NJ, USA. https://doi.org/10.1002/9781118540770.ch7.

Gallistel, C. R. (1990). *The Organization of Learning*. Cambridge, MA: MIT press.
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
Gallistel, C.R., & Gelman, R. (2000). Non-verbal numerical cognition: from reals to integers. *Trends in Cognitive Science, 4(2)*, 59-65. https://doi.org/10.1016/S1364-6613(99)01424-2
Garrett, H. E. (1922). A study of the relation of accuracy to speed. *Archives of Psychology*, 56, 1–105.
Gehring, W. J., Goss, B., Coles, M. G., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological science*, 4(*6*), 385-390. https://doi.org/10.1111/j.1467-9280.1993.tb00586.x

Gelman, R. & Gallistel, C. (1978). *The Child's Understanding of Number.* Cambridge, MA. Harvard University Press.
Gibbon, J., Church, R. M., & Meck, W. H. (1984). Scalar timing in memory. *Annals of the New York Academy of Sciences, 423*, 52-77. https://doi.org/10.1111/j.1749-6632.1984.tb23417.x
Gobet, F., Lane, P. C., Croker, S., Cheng, P. C., Jones, G., Oliver, I., & Pine, J. M. (2001). *Chunking mechanisms in human learning*. Trends in cognitive sciences, 5(*6*), 236-243. https://doi.org/10.1016/S1364-6613(00)01662-4
Gorea, A., & Sagi, D. (2000). Failure to handle more than one internal representation in visual detection tasks. *Proceedings of the National Academy of Sciences of the United States of America, 97(22)*, 12380–4. https://doi.org/10.1073/pnas.97.22.12380
Gorea, A., Caetta, F., & Sagi, D. (2005). Criteria interactions across visual attributes. *Vision research*, *45*(19), 2523-2532. http://dx.doi.org/10.1016/j.visres.2005.03.018
Gür, E., & Balcı, F. (2017). Mice optimize timed decisions about probabilistic outcomes under deadlines. Animal cognition, 20(3), 473-484. https://doi.org/10.1007/s10071-017-1073-y
Green, P., MacLeod, C. J. and Alday, P. (2016a). Package 'simr', Available at: https://cran.r-project.org/web/packages/simr/simr.pdf.
Green, P., & MacLeod, C. J. (2016b). SIMR: an R package for power analysis of generalized linear mixed models by simulation. Methods in Ecology and Evolution, 7(4), 493-498. https://doi.org/10.1111/2041-210X.12504
Di Gregorio, F., Maier, M. E., & Steinhauser, M. (2018). Errors can elicit an error positivity in the absence of an error negativity: Evidence for independent systems of human error monitoring. *NeuroImage*, *172*, 427-436. https://doi.org/10.1016/j.neuroimage.2018.01.081
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.
Henik, A., & Tzelgov, J. (1982). Is three greater than five: The relation between physical and semantic size in comparison tasks. *Memory & Cognition*, *10*, 389–395. https://doi.org/10.3758/BF03202431

Hollingworth, H.L. (1910). The Central Tendency of Judgment. *Journal of Philosophy, Psychology & Scientific Methods, 7,* 461-469. http://dx.doi.org/10.2307/2012819

JASP Team (2018). JASP (Version 0.9)[Computer software].

Johnson, D. M. (1939). Confidence and speed in the two-category judgment. *Archives of Psychology,* 34, 1–53.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231. https://doi.org/10.1038/nature07200

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

Kiesel, A., & Vierck, E. (2009). SNARC-like congruency based on number magnitude and response duration. Journal of Experimental Psychology. Learning, Memory, and Cognition, 35, 275–279, https://doi.org/10.1037/a0013737.

Kononowicz, T. W., Roger, C., & van Wassenhove, V. (2017). Temporal metacognition as the decoding of self-generated brain dynamics. *bioRxiv*, 206086. https://doi.org/10.1101/206086

Lejeune, H., & Wearden, J. H. (2009). Vierordt's The Experimental Study of the Time Sense (1868) and its legacy. *European Journal of Cognitive Psychology*, *21*(6), 941-960. https://doi.org/10.1080/09541440802453006

Logan, G. D., & Crump, M. J. C. (2010). Cognitive Illusions of Authorship Reveal Hierarchical Error Detection in Skilled Typists. *Science, 330*(6004), 683–686. https://doi.org/10.1126/science.1190483

Luft, C. D. B., Takase, E., & Bhattacharya, J. (2014). Processing graded feedback: electrophysiological correlates of learning from small and large errors. *Journal of Cognitive Neuroscience*, *26*(*5*), 1180-1193. https://doi.org/10.1162/jocn_a_00543

Mandler, G., & Shebo, B. J. (1982). Subitizing: an analysis of its component processes. *Journal of Experimental Psychology: General*, 111(*1*), 1. https://doi.org/10.1037//0096-3445.111.1.1

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

Martin, B., Wiener, M. & Van Wassenhove, V. (2017). A Bayesian perspective on accumulation in the magnitude system. *Scientific Reports, 7(1),* 630. https://doi.org/10.1101/101568

Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes*, *9*(3), 320–334. https://doi.org/10.1037/0097-7403.9.3.320

Montemayor, C., & Balci, F. (2007). Compositionality in language and arithmetic. *Journal of Theoretical and Philosophical Psychology, 27(1),* 53-72. http://doi.org/10.1037/h0091281

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, *78*, 99–147. https://doi.org/10.1016/j.cogpsych.2015.01.002

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical in- equality. Nature, 215, 1519–1520, https://doi.org/10.1038/2151519a0.

Navajas, J., Hindocha, C., Foda, H., Keramati, M., & Latham, P. E. (2017). The idiosyncratic nature of confidence. *bioRxiv*, 1–30. https://doi.org/10.1101/102269

Nelson, T. O.,& Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, 26, 125–141.

Nieder, A., Diester, I., & Tudusciuc, O. (2006). Temporal and spatial enumeration processes in the primate. *Science, 313*, 1431–1435. http://doi.org/10.1126/science.1130308

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

Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A Bayesian perspective on magnitude estimation. Trends in Cognitive Sciences, *19*(*5*), 285-293. https://doi.org/10.1016/j.tics.2015.03.002

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. http://doi.org/10.1037/a0019737

Rahnev, D. & Denison, R.N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences, 41*, e223. https://doi.org/10.1017/s0140525x18000936

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences, 20(4)*, 260-281. http://doi.org/10.1016/j.tics.2016.01.007

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.

Restle, F. (1970). Speed of adding and comparing numbers. *Journal of Experimental Psychology*, 83, 274–278, https://doi.org/10.1037/h0028573.

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(7261), 263–266. https://doi.org/10.1038/nature08275

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

Samaha, J., & Postle, B. R. (2017). Correlated individual differences suggest a common mechanism underlying metacognition in visual perception and visual short-term memory. *Proceedings of the Royal Society B: Biological Sciences*, *284*(*1867*), 20172035. https://doi.org/10.1098/rspb.2017.2035

Santiago, J., Lupiáñez, J., Perez, E., & Funes, M. J. (2007). Time (also) flies from left to right. Psychonomic Bulletin & Review, 14, 512–516. https://doi.org/10.3758/BF03194099

Smets, K., Sasanguie, D., Szuecs, D., & Reynvoet, B. (2015). The effect of different methods to construct non-symbolic stimuli in numerosity estimation and comparison. Journal of Cognitive Psychology, 27(3), 310 - 325. https://doi.org/10.1080/20445911.2014.996568

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168. https://doi.org/10.1016/j.tins.2004.01.006

Sperling, G. (1983). Why we need iconic memory. *Behavioral and Brain Sciences*, *6*(1), 37-39. https://doi.org/10.1017/S0140525X00014564

Spinelli, G., Tieri, G., Pavone, E. F., & Aglioti, S. M. (2018). Wronger than wrong: Graded mapping of the errors of an avatar in the performance monitoring system of the onlooker. *NeuroImage*, *167*, 1-10. https://doi.org/10.1016/j.neuroimage.2017.11.019

Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, *54*(*6*), 377. http://dx.doi.org/10.1037/h0043680

Tang, J., Critchley, H. D., Glaser, D. E., Dolan, R. J., & Butterworth, B. (2006). Imaging informational conflict: A functional magnetic resonance imaging study of numerical Stroop. *Journal of cognitive neuroscience*, 18(*12*), 2049-2062. https://doi.org/10.1162/jocn.2006.18.12.2049

Taylor, S. F., Stern, E. R., & Gehring, W. J. (2007). Neural systems for error monitoring: recent findings and theoretical perspectives. *The Neuroscientist*, *13*(*2*), 160-172. https://doi.org/10.1177/1073858406298184

Teodorescu, A. R., & Usher, M. (2013). Disentangling decision models: from independence to competition. *Psychological Review*, *120*(1), 1–38. https://doi.org/10.1037/a0030776

Tokita, M., & Ishiguchi, A. (2012). Behavioral evidence for format-dependent processes in approximate numerosity representation. *Psychonomic Bulletin & Review*, *19*, 285–293. http://doi.org/10.3758/s13423-011-0206-6

Torralbo, A., Santiago, J., & Lupiáñez, J. (2006). Flexible conceptual projection of time onto spatial frames of reference. Cognitive Science, 30, 749–757. https://doi.org/10.1207/s15516709cog0000_67

Torrecillos, F., Albouy, P., Brochier, T., & Malfait, N. (2014). Does the processing of sensory and reward-prediction errors involve common neural resources? Evidence from a frontocentral negative potential modulated by movement execution errors. *Journal of Neuroscience*, *34*(*14*), 4845-4856. https://doi.org/10.1523/JNEUROSCI.4390-13.2014

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, *7*(3), 424–465. https://doi.org/10.3758/BF03214357

Van Zandt, T., & Maldonado-Molina, M. M. (2004). Response reversals in recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 1147–1166. https://doi.org/10.1037/0278-7393.30.6.1147

Vickers, D. (1979). *Decision Processes in Visual Perception*. Retrieved from http://books.google.com/books?hl=en&lr=&id=LXA-AwAAQBAJ&pgis=1

Vickers, D., & Packer, J. (1982). Effects of alternating set for speed or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. *Acta Psychologica*, *50*(2), 179–197. https://doi.org/10.1016/0001-6918(82)90006-3

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

Vocat, R., Pourtois, G., & Vuilleumier, P. (2011). Parametric modulation of error-related ERP components by the magnitude of visuo-motor mismatch. *Neuropsychologia*, *49*(3), 360-367. https://doi.org/10.1016/j.neuropsychologia.2010.12.027

Volkman, J. (1934). The relation of the time of judgment to the certainty of judgment. *Psychological Bulletin,* 31, 672–673.

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.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, 143(*5*), 2020. https://doi.org/10.1037/xge0000014

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. http://doi.org/10.3389/fpsyg.2014.00851.

Zakay, D., & Tuvia, R. (1998). Choice latency times as determinants of post-decisional confidence. *Acta Psychologica*, 98(1), 103–115. https://doi.org/10.1016/S0001-6918(97)00037-1

Supplementary Materials – Chapter I

1. Distribution of Confidence-Directionality (Signed Confidence) Ratings

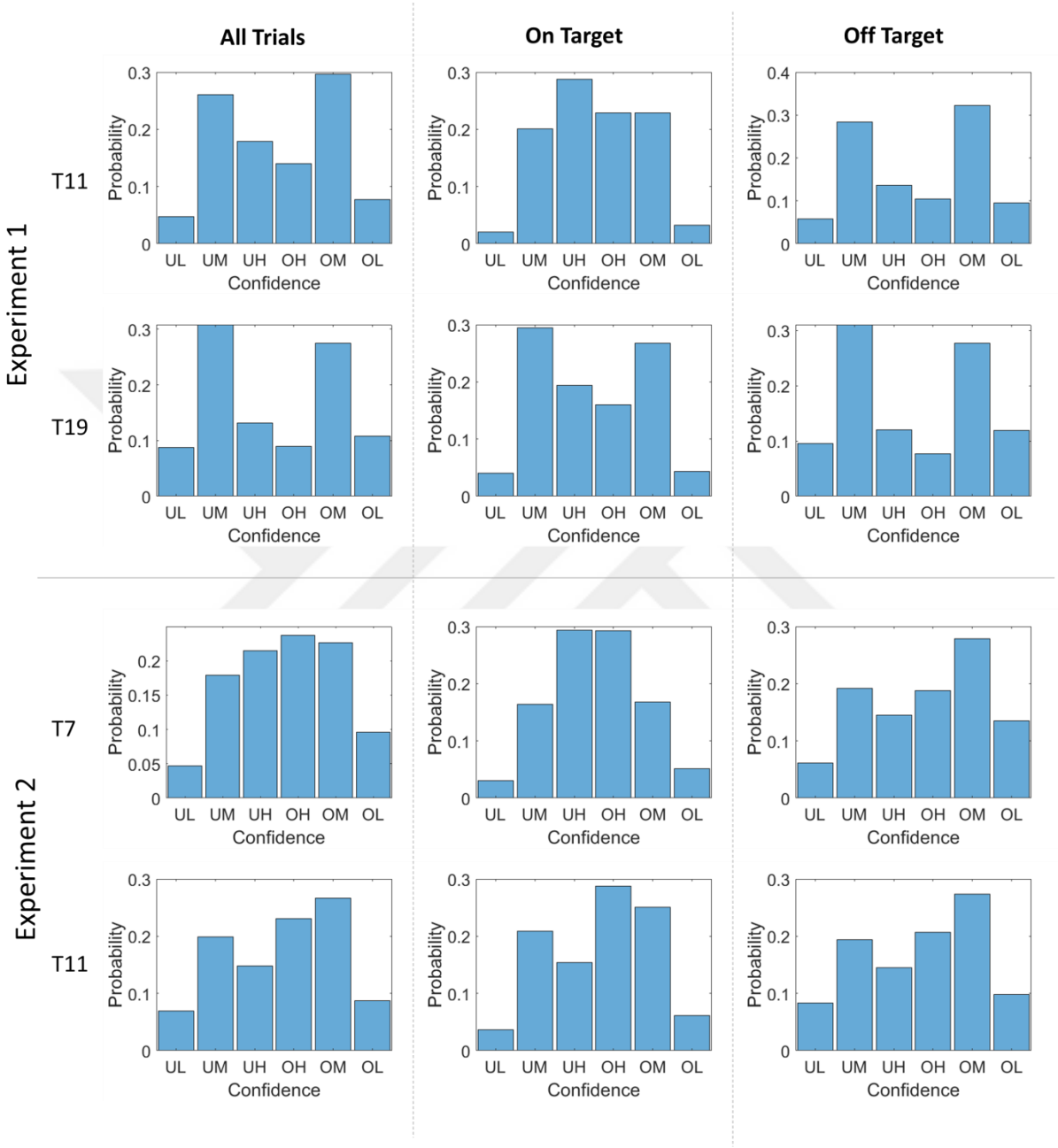

Figure S1.1.1 Distribution of Confidence-Directionality ratings for all trials and for trials where the participants' responses were on target and off target in both experiments.

2. Model Summaries for Individual Fits

	T11			T19		
	Beta (Standardized)	R^2 (Adjusted)	*p* value	Beta (Standardized)	R^2 (Adjusted)	*p* value
ID						
1	0.625	0.382	**<.001**	0.591	0.339	**<.001**
2	0.293	0.07	**0.024**	0.22	0.037	**0.043**
3	0.406	0.156	**<.001**	0.211	0.035	**0.036**
4	0.406	0.159	**<.001**	0.423	0.172	**<.001**
5	0	0.045	0.756	0.122	0.004	0.246
6	0.356	0.119	**<.001**	0.225	0.041	**0.023**
7	0.634	0.397	**<.001**	0.619	0.38	**<.001**
8	0.59	0.343	**<.001**	0.618	0.376	**<.001**
9	0.305	0.087	**<.001**	0.075	-0.004	0.456
10	0.456	0.201	**<.001**	0.11	0.001	0.289
11	-0.096	0.002	0.269	0.164	0.018	0.09
12	0.49	0.234	**<.001**	0.4	0.151	**<.001**
13	0.327	0.097	**0.001**	0.188	0.026	0.053
14	0.314	0.093	**<.001**	0.258	0.058	**0.007**
15	0.192	0.029	**0.029**	0.259	0.057	**0.011**
16	0.482	0.217	**<.001**	0.146	0.007	0.225
17	0.306	0.087	**<.001**	0.326	0.098	**0.001**
18	0.264	0.062	**0.003**	0.262	0.059	**0.008**
19	0.176	0.025	**0.029**	0.099	-0.002	0.354
20	0.26	0.062	**<.001**	0.231	0.046	**0.007**
21	0.265	0.063	**0.003**	0.075	-0.007	0.498
22	0	0.006	**<.001**	0.25	0.054	**0.009**
23	0.727	0.524	**<.001**	0.556	0.302	**<.001**
24	-0.083	-0.002	0.385	-0.092	-0.004	0.413
25	0.248	0.054	**0.006**	0.388	0.142	**<.001**
26	0.433	0.181	**<.001**	0.294	0.078	**0.002**
27	-0.039	-0.007	0.666	0.033	-0.006	0.701
28	0.135	0.01	0.142	0.425	0.171	**<.001**
29	0.862	0.735	**<.001**	0.173	0.015	0.156

Table S1.2.1 Standardized coefficients, R^2 and *p* values for individual regression fits of signed confidence ratings on reproduced numerosities in Experiment 1.

	T7			**T11**		
	Beta (Standardized)	R^2 (Adjusted)	*p* value	Beta (Standardized)	R^2 (Adjusted)	*p* value
ID						
1	0.481	0.225	**<.001**	0.156	0.403	**<.001**
2	0.535	0.276	**<.001**	0.458	0.682	**<.001**
3	0.605	0.361	**<.001**	0.185	0.437	**<.001**
4	0.398	0.153	**<.001**	0.121	0.358	**<.001**
5	0.426	0.176	**<.001**	-0.004	0.054	0.524
6	n.a.	n.a.	n.a.	0.42	0.655	**<.001**
7	0.197	0.032	**0.015**	0.009	0.127	0.122
8	0.23	0.047	**0.005**	0.095	0.321	**0.001**
9	0.262	0.058	**0.012**	0.137	0.379	**<.001**
10	0.658	0.428	**<.001**	0.147	0.399	**0.001**
11	0.175	0.025	**0.026**	0.107	0.337	**<.001**
12	0.609	0.363	**<.001**	0.161	0.41	**<.001**
13	n.a.	n.a.	n.a.	0.002	0.113	0.27
14	0.258	0.058	**0.008**	0.021	0.164	**0.04**
15	0.526	0.269	**<.001**	0.217	0.472	**<.001**

Table S1.2.2 Standardized coefficients, R^2 and *p* values for individual regression fits of signed confidence ratings on reproduced numerosities, in Experiment 2. In T7, the regression slopes could not be computed for participants 6 and 13, because their reproductions were on target for a significant portion of their trials.

3. Summary tables and diagnostic plots for the mixed-effect models

Fixed effects coefficients (95% CIs):

Name	*Coefficient*	*SE*	*tStat*	*DF*	*p*	*Lower*	*Upper*
(Intercept)	.994	.252	3.935	3452	< .001	.499	1.489
nrBeeps	.217	.026	8.443	3452	<.001	.167	.267

Random effects covariance parameters (95% CIs):
Group: subject (29 Levels)

Name 1	*Name 2*	*std*	*Lower*	*Upper*
Intercept	Intercept	1.121	.718	1.749
nrBeeps	nrBeeps	0.123	0.083	0.182

AIC: 11904 *BIC:* 11935 *Log Likelihood:* -5947 *Deviance:* 11894

Table S1.3.1 Summary table for the mixed effects model for T11 in Experiment 1

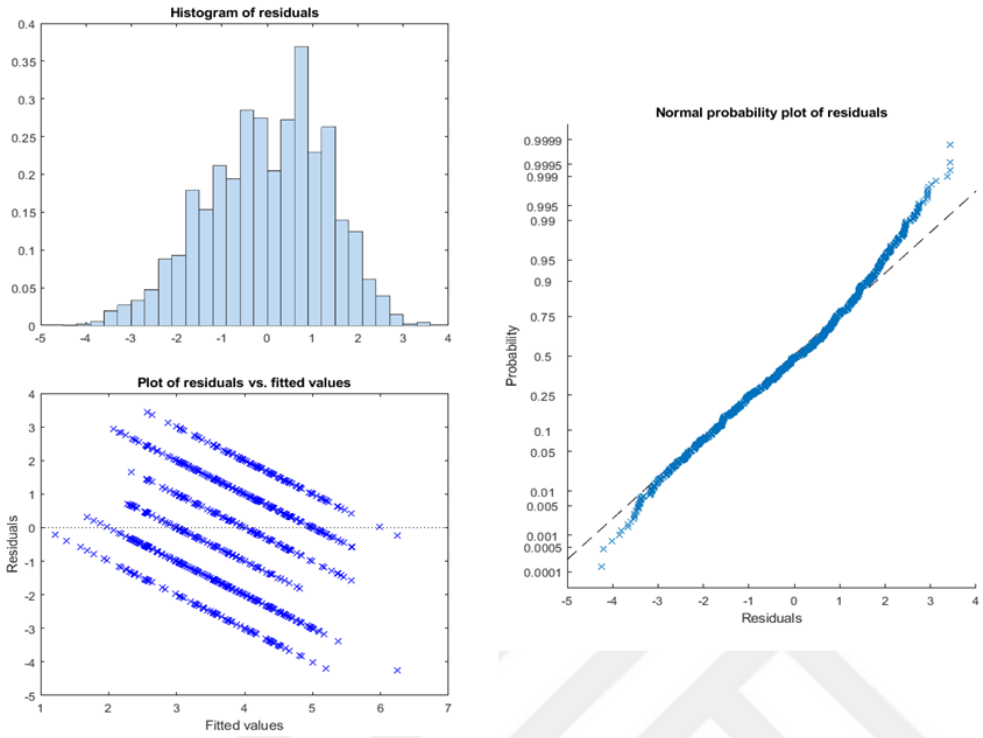

Figure S1.3.1 Diagnostic plots for the mixed-effect model for T11 in Experiment 1

Fixed effects coefficients (95% CIs):

Name	*Coefficient*	*SE*	*tStat*	*DF*	*p*	*Lower*	*Upper*
(Intercept)	1.292	.18	7.205	2915	< .001	.941	1.645
nrBeeps	.109	.01	11.157	2915	<.001	.09	.129

Random effects covariance parameters (95% CIs):
Group: subject (29 Levels)

Name 1	*Name 2*	*std*	*Lower*	*Upper*
Intercept	Intercept	.52	.29	.935
nrBeeps	nrBeeps	.035	0.023	0.054

AIC: 10683 *BIC:* 10713 *Log Likelihood:* -5336.5 *Deviance:* 10673

Table S1.3.2 Summary table for the mixed effects model for T19in Experiment 1.

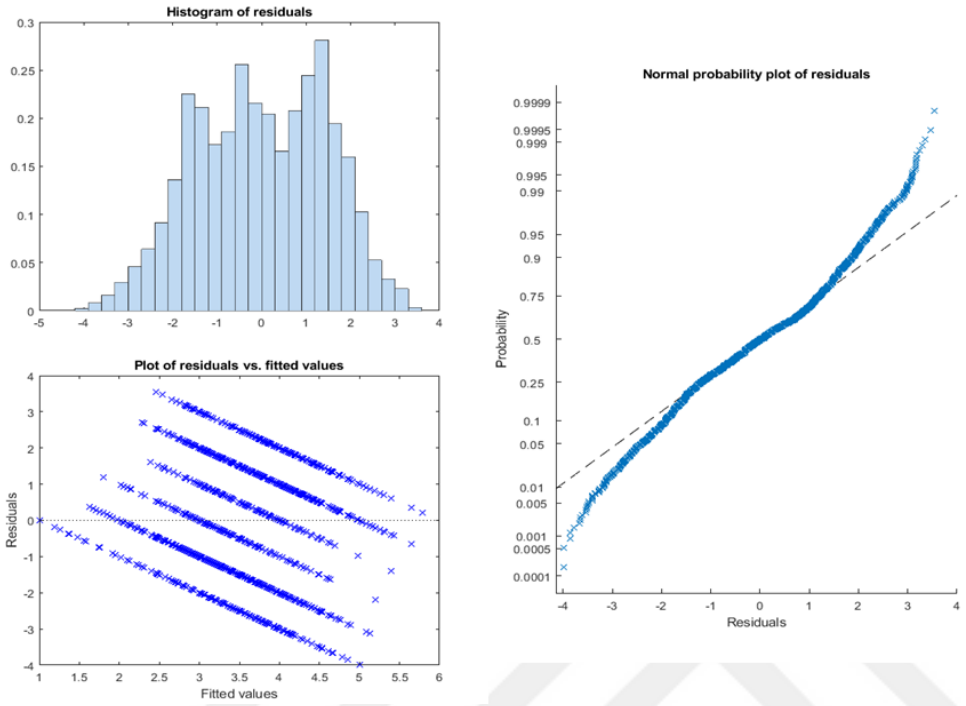

Figure S1.3.2 Diagnostic plots for the mixed-effect model for T19 in Experiment 1.

Fixed effects coefficients (95% CIs):

Name	*Coefficient*	*SE*	*tStat*	*DF*	*p*	*Lower*	*Upper*
(Intercept)	.396	.54	.732	1615	.464	-.664	1.455
nrBeeps	.483	.08	6.017	1615	<.001	.326	.641

Random effects covariance parameters (95% CIs):
Group: subject (29 Levels)

Name 1	*Name 2*	*std*	*Lower*	*Upper*
Intercept	Intercept	1.852	1.204	2.848
nrBeeps	nrBeeps	.283	.186	.432

AIC: 5525.7 *BIC:* 5552.7 *Log Likelihood:* -2757.9 *Deviance:* 5515.7

Table 1.3.3 Summary table for the mixed effects model for T7 in Experiment 2.

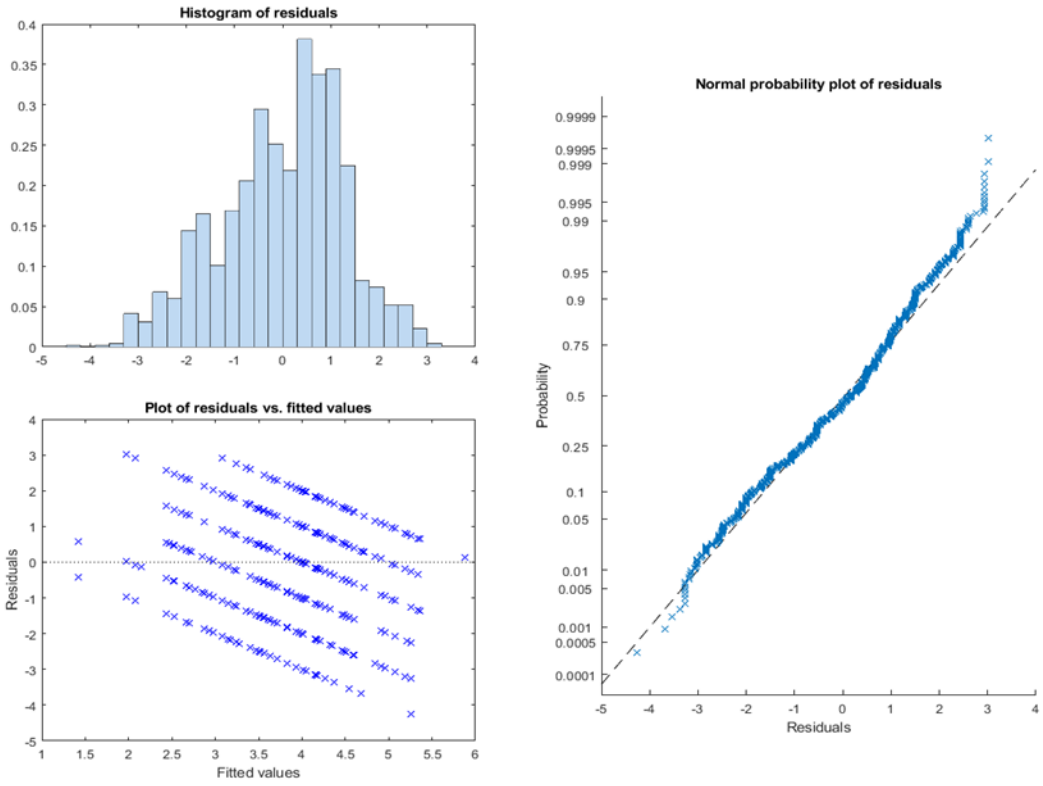

Figure S1.3.3. Diagnostic plots for the mixed-effect model for T7 in Experiment 2.

Fixed effects coefficients (95% CIs):							
Name	*Coefficient*	*SE*	*tStat*	*DF*	*p*	*Lower*	*Upper*
(Intercept)	.293	.66	.444	1809	.657	-1	1.588
nrBeeps	.323	.063	5.131	1809	<.001	.199	.446
Random effects covariance parameters (95% CIs): Group: subject (29 Levels)							
Name 1	*Name 2*	*std*	*Lower*	*Upper*			
Intercept	Intercept	2.403	1.521	3.8			
nrBeeps	nrBeeps	.23	.146	.361			
AIC: 6361.5 *BIC:* 6389 *Log Likelihood:* -3175.7 *Deviance:* 6351.5							

Table S1.3.4. Summary table for the mixed effects model for T11 in Experiment 2.

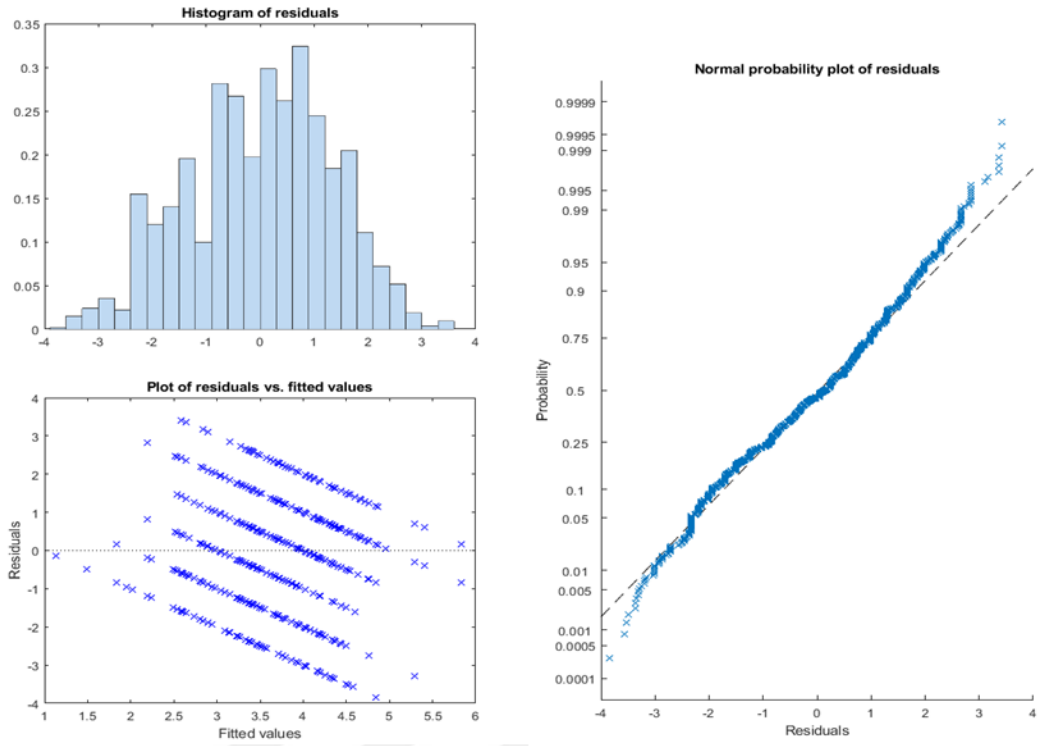

Figure S1.3.4. Diagnostic plots for the mixed-effect model for T11 in Experiment 2.

4. Estimation performance across trials

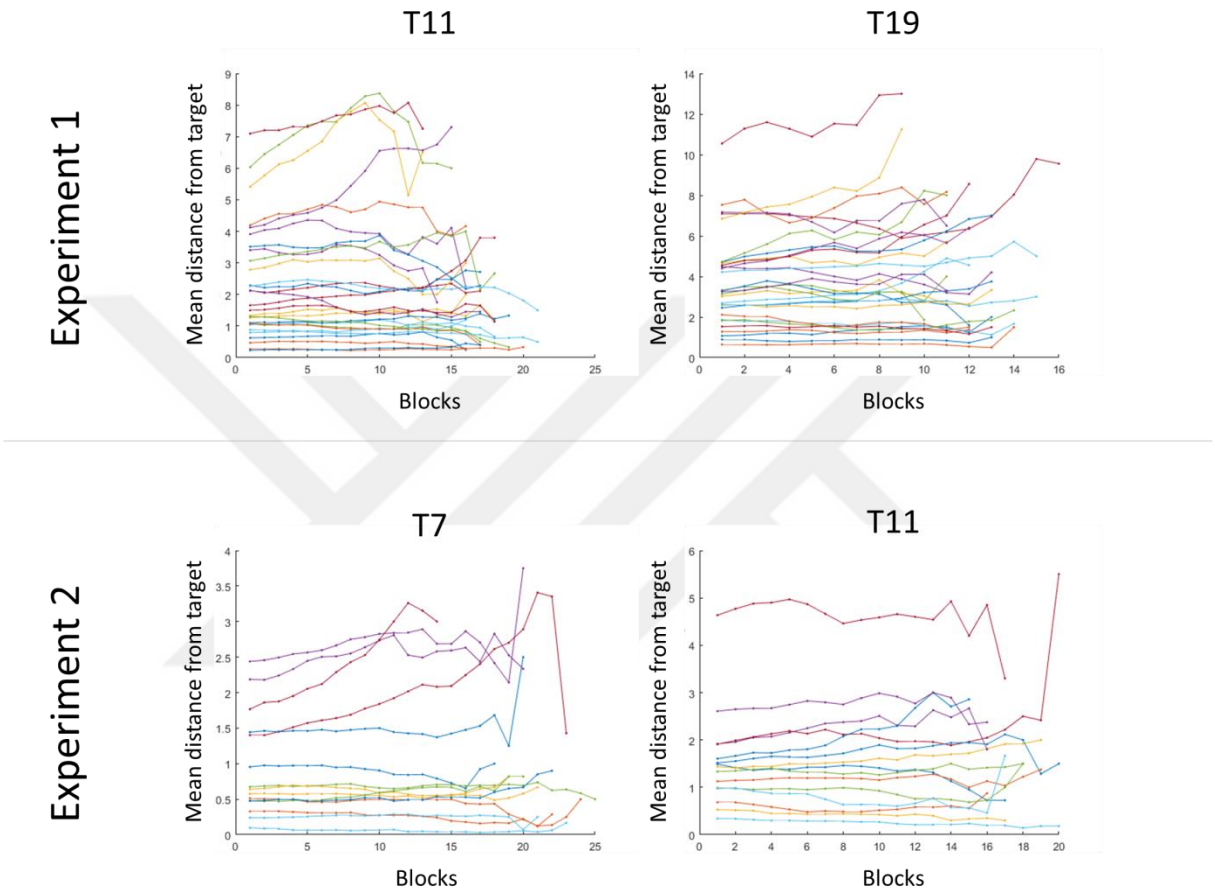

Figure S1.4.1 Change in numerical estimation performance across blocks of 10 trials. Each line shows a participant's mean absolute error across blocks. To explore whether the participants' performance improved across trials, we regressed mean distances for each block on block numbers. We then compared the resulting slopes to zero. There were no significant negative slopes, indicating that the participants' overall performance did not increase across blocks. Note that the last bin might contain more than 10 trials depending on the total number of trials of a participant.

5. Beta coefficients, Numerical CVs and CVs of RTs

	T11			T19		
ID	Beta NrBeep	CV(number)	CV(time)	Beta NrBeep	CV(number)	CV(time)
1	0.625	0.085	0.104	0.591	0.085	0.099
2	0.293	0.077	0.108	0.220	0.093	0.102
3	0.406	0.156	0.169	0.211	0.180	0.174
4	0.406	0.218	0.231	0.423	0.210	0.214
5	0.000	0.080	0.090	0.122	0.114	0.116
6	0.356	0.125	0.139	0.225	0.162	0.167
7	0.634	0.173	0.181	0.619	0.412	0.436
8	0.590	0.305	0.321	0.618	0.303	0.310
9	0.305	0.208	0.220	0.075	0.209	0.219
10	0.456	0.252	0.254	0.110	0.168	0.177
11	-0.096	0.229	0.241	0.164	0.218	0.224
12	0.490	0.278	0.294	0.400	0.199	0.202
13	0.327	0.102	0.112	0.188	0.113	0.126
14	0.314	0.172	0.177	0.258	0.177	0.174
15	0.192	0.168	0.175	0.259	0.180	0.187
16	0.482	0.045	0.079	0.146	0.048	0.072
17	0.306	0.145	0.158	0.326	0.157	0.160
18	0.264	0.232	0.245	0.262	0.208	0.210
19	0.176	0.201	0.204	0.099	0.164	0.169
20	0.260	0.240	0.257	0.231	0.257	0.266
21	0.265	0.145	0.148	0.075	0.157	0.158
22	0.000	0.143	0.153	0.250	0.163	0.157
23	0.727	0.112	0.122	0.556	0.136	0.150
24	-0.083	0.340	0.354	-0.092	0.483	0.494
25	0.248	0.162	0.172	0.388	0.157	0.164
26	0.433	0.122	0.126	0.294	0.109	0.116
27	-0.039	0.115	0.122	0.033	0.130	0.140
28	0.135	0.154	0.160	0.425	0.112	0.121
29	0.862	0.046	0.073	0.173	0.069	0.087

Table S1.5.1 Numerical and temporal CVs and standardized coefficients for reproduced numerosities in Experiment 1.

	T7			T11		
ID	Beta NrBeep	CV(number)	CV(time)	Beta NrBeep	CV(number)	CV(time)
1	0.481	0.129	0.150	0.403	0.128	0.141
2	0.535	0.079	0.109	0.682	0.092	0.096
3	0.605	0.112	0.141	0.437	0.116	0.136
4	0.398	0.245	0.265	0.358	0.192	0.198
5	0.426	0.133	0.155	0.054	0.152	0.170
6	n.a	0.059	0.081	0.655	0.053	0.083
7	0.197	0.226	0.253	0.127	0.242	0.255
8	0.230	0.148	0.161	0.321	0.151	0.154
9	0.262	0.125	0.145	0.379	0.149	0.162
10	0.658	0.119	0.132	0.399	0.080	0.101
11	0.175	0.277	0.296	0.337	0.243	0.255
12	0.609	0.108	0.127	0.410	0.109	0.127
13	n.a.	0.042	0.084	0.113	0.113	0.127
14	0.258	0.367	0.416	0.164	0.579	0.643
15	0.526	0.111	0.135	0.472	0.180	0.207

Table S1.5.2 Numerical and temporal CVs and standardized coefficients for reproduced numerosities in Experiment 2.

6. Correlations between CVs and Slopes

	Experiment 1		Experiment 2	
	T11	*T19*	*T7*	*T11*
r	-.233	.037	-.663	-.469
***p* value**	.224	.848	0.014	.07

Table S1.6.1 Coefficients and corresponding p values for the correlations between individual regression slopes for and numerical CVs.

Supplementary Materials – Chapter II

Table S2.1. *Table for mixed effects for targets only (for target numerosities > 7)*

	Estimate	*SE*	*tStat*	*DF*	*pValue*	*Lower*	*Upper*
Fixed Effects							
Intercept	*-.232*	*.205*	*-1.133*	*3455*	*.257*	*-.633*	*.169*
Response	*.411*	*.051*	*8.045*	*3455*	*<.001*	*.311*	*.511*
Random Effects *Group: subject (18 Levels)*							
	Std	*Lower*	*upper*				
Intercept	*.858*	*.614*	*1.199*				
Response	*.168*	*.098*	*.291*				
AIC: 13959 BIC: 13989 *Log Likelihood: -6974.4*							

Table S2.2. *Table for mixed effects for targets and distractors (for numerosities >7)*

	Estimate	*SE*	*tStat*	*DF*	*pValue*	*Lower*	*Upper*
Fixed Effects							
Intercept	*-.229*	*.212*	*-1.08*	*7211*	*.28*	*-.645*	*.187*
Response	*.419*	*.052*	*8.094*	*7211*	*<.001*	*.317*	*.52*
Random Effects *Group: subject (18 Levels)*							
	Std	*Lower*	*upper*				
Intercept	*.9*	*.644*	*1.246*				
Response	*.197*	*.131*	*.296*				
AIC: 29343	*BIC: 29377 Log Likelihood: -14666*						

Table S2.3. *Table for mixed effects for targets only (for target numerosities >7), when confidence-directionality judgment pairs are coded 1-6.*

	Estimate	*SE*	*tStat*	*DF*	*pValue*	*Lower*	*Upper*
Fixed Effects							
Intercept	*3.361*	*.173*	*19.407*	2788	*<.001*	*3.021*	*3.7*
Response	*.333*	*.044*	*7.595*	*2788*	*<.001*	*.247*	*.42*
Random Effects *Group: subject (18 Levels)*							
	Std	*Lower*	*upper*				
Intercept	*.725*	*.518*	*1.014*				
Response	*.14*	*.079*	*.255*				
AIC: 10289 *BIC:* 10319 *Log Likelihood: -5139*							

Table S2.4. *Table for mixed effects for targets and distractors (for target numerosities >7), when confidence-directionality judgment pairs are coded 1-6.*

	Estimate	*SE*	*tStat*	*DF*	*pValue*	*Lower*	*Upper*
Fixed Effects							
Intercept	*3.351*	*.176*	*19.094*	*6135*	*<.001*	*3.007*	*3.695*
Response	*.343*	*.0447*	*7.742*	*6135*	*<.001*	*.256*	*.429*
Random Effects *Group: subject (18 Levels)*							
	Std	*Lower*	*upper*				
Intercept	*.74*	*.532*	*1.03*				
Response	*.168*	*.111*	*.253*				
AIC: 22564	*BIC: 22597*	*Log Likelihood: -11277*					

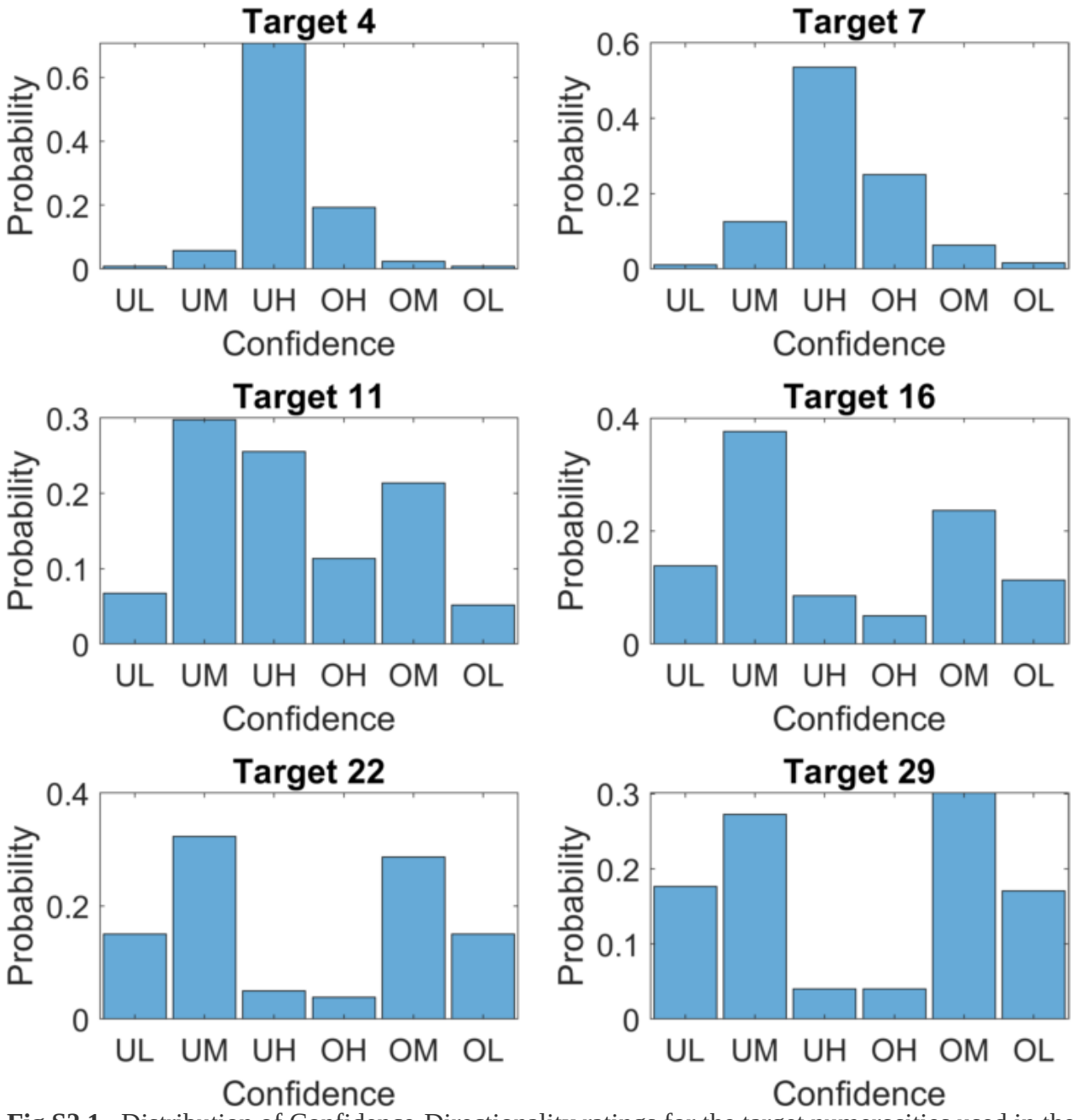

Fig S2.1. Distribution of Confidence-Directionality ratings for the target numerosities used in the experiment. Note the change in the pattern of ratings (from unimodal to bimodal) after Target numerosity 7.

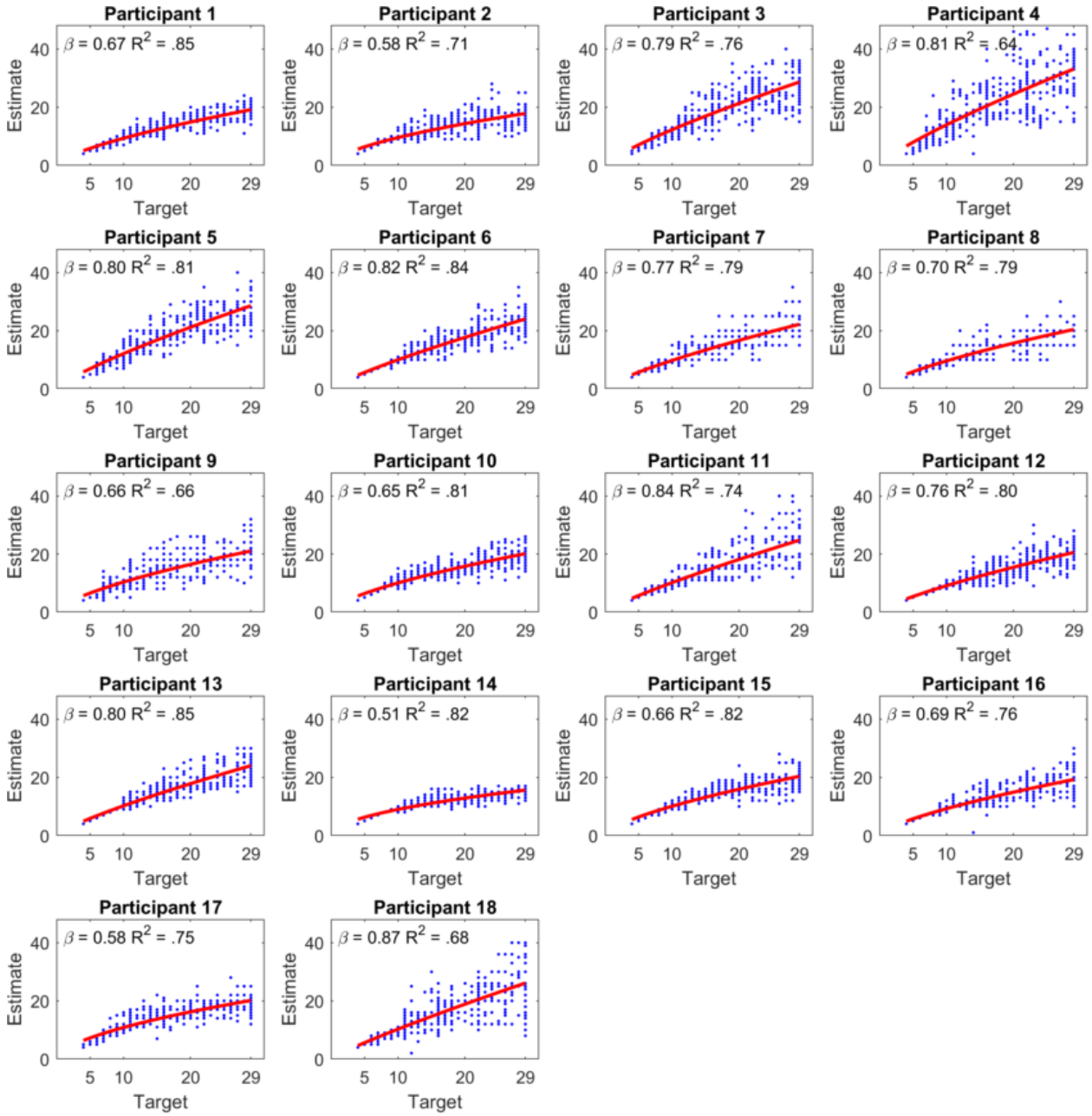

Fig S2.2. Power function fits to the numerosity estimations for each participant.

Supplementary Materials – Chapter III

S3. Experiment 1: Model Tables

Model Info

Info	
Estimate	Linear mixed model fit by ML
Call	direction ~ 1 + targetCategory + zScore + zScore:targetCategory+(1|id)+(0+ zScore|id)
AIC	6024.1690
BIC	6082.0694
LogLikel.	-3003.0845
R-squared Marginal	0.0432
R-squared Conditional	0.1524

Table S3.1.1

Fixed Effect Omnibus tests

	F	Num df	Den df	p
targetCategory	73.20	2	4558.0	< .001
zScore	23.90	1	20.0	< .001
targetCategory ✻ zScore	6.12	2	4558.4	0.002

Table S3.1.2

Fixed Effects Parameter Estimates

Names	Effect	Estimate	SE	95% Confidence Interval		df	t	p
				Lower	Upper			
(Intercept)	(Intercept)	1.4875	0.0362	1.4164	1.5585	20.0	41.05	< .001
Short	2 - 1	0.0693	0.0166	0.0367	0.1019	4558.0	4.17	< .001
targetCategory2	3 - 1	0.1983	0.0166	0.1657	0.2309	4558.0	11.92	< .001
zScore	zScore	0.0594	0.0122	0.0356	0.0833	20.0	4.89	< .001
targetCategory1 ✻ zScore	2 - 1 ✻ zScore	0.0540	0.0167	0.0212	0.0868	4558.5	3.23	0.001
targetCategory2 ✻ zScore	3 - 1 ✻ zScore	0.0466	0.0167	0.0137	0.0794	4558.4	2.78	0.005

Table s3.1.3

Model Info

Info	
Estimate	Linear mixed model fit by ML
Call	confidence ~ 1 + targetCategory + absZ+(1|id)+(0+ absZ|id)
AIC	9487.8130
BIC	9532.8467
LogLikel.	-4736.9065
R-squared Marginal	0.0194
R-squared Conditional	0.1552

Table S3.1.4

Fixed Effect Omnibus tests

	F	Num df	Den df	p
targetCategory	49.10	2	4560.0	< .001
absZ	6.37	1	20.9	0.020

Note. Satterthwaite method for degrees of freedom

Table S3.1.5

Fixed Effects Parameter Estimates

Names	Effect	Estimate	SE	95% Confidence Interval		df	t	p
				Lower	Upper			
(Intercept)	(Intercept)	2.1117	0.0620	1.9902	2.2333	21.6	34.06	< .001
targetCategory1	2 - 1	-0.2023	0.0243	-0.2499	-0.1547	4560.6	-8.33	< .001
targetCategory2	3 - 1	-0.2138	0.0243	-0.2614	-0.1662	4559.3	-8.80	< .001
absZ	absZ	-0.0468	0.0186	-0.0832	-0.0105	20.9	-2.52	0.020

Table S3.1.6

Model Info

Info	
Estimate	Linear mixed model fit by ML
Call	signedConfidence ~ 1 + targetCategory + zScore + targetCategory:zScore+(1|id)+(0+ zScore|id)
AIC	19180.0119
BIC	19237.9122
LogLikel.	-9581.0059
R-squared Marginal	0.0402
R-squared Conditional	0.1265

Table S3.1.7

Fixed Effect Omnibus tests

	F	Num df	Den df	p
targetCategory	59.0	2	4558.0	< .001
zScore	24.9	1	19.9	< .001
targetCategory ✻ zScore	10.2	2	4558.4	< .001

Note. Satterthwaite method for degrees of freedom

Table S3.1.8

Fixed Effects Parameter Estimates

Names	Effect	Estimate	SE	95% Confidence Interval		df	t	p
				Lower	Upper			
(Intercept)	(Intercept)	-0.056	0.132	-0.316	0.203	20.0	-0.42	0.675
targetCategory1	2 - 1	0.2492	0.069	0.113	0.386	4558	3.58	< .001
targetCategory2	3 - 1	0.7427	0.069	0.606	0.879	4558	10.67	< .001
zScore	zScore	0.2449	0.049	0.149	0.341	19.9	4.99	< .001
targetCategory1 ✻ zScore	2 - 1 ✻ zScore	0.2762	0.070	0.139	0.414	4558	3.95	< .001
targetCategory2 ✻ zScore	3 - 1 ✻ zScore	0.2703	0.070	0.133	0.408	4558	3.86	< .001

Table S3.1.9

S2. Experiment 2: Model Tables

Model Info

Info	
Estimate	Linear mixed model fit by ML
Call	direction ~ 1 + targetCategory + zScore+(1|id)+(0+ zScore|id)
AIC	6279.3024
BIC	6324.6370
LogLikel.	-3132.6512
R-squared Marginal	0.0517
R-squared Conditional	0.1400

Table S3.2.1

Fixed Effect Omnibus tests

	F	Num df	Den df	p
targetCategory	89.6	2	4758.1	< .001
zScore	34.5	1	21.0	< .001

Note. Satterthwaite method for degrees of freedom

Table S3.2.2

Fixed Effects Parameter Estimates

Names	Effect	Estimate	SE	95% Confidence Interval		df	t	p
				Lower	Upper			
(Intercept)	(Intercept)	1.3306	0.0328	1.2664	1.3948	24.9	40.62	< .001
targetCategory1	2 - 1	0.1157	0.0163	0.0838	0.1476	4758.1	7.11	< .001
targetCategory2	3 - 1	0.2178	0.0163	0.1859	0.2498	4758.1	13.37	< .001
zScore	zScore	0.0699	0.0119	0.0466	0.0933	21.0	5.87	< .001

Table S3.2.3

Model Info

Info	
Estimate	Linear mixed model fit by ML
Call	confidence ~ 1 + absZ + targetCategory+(1 | id)+(0+absZ | id)
AIC	8267.3245
BIC	8312.6591
LogLikel.	-4126.6622
R-squared Marginal	0.0100
R-squared Conditional	0.3881

Table S3.2.4

Fixed Effect Omnibus tests

	F	Num df	Den df	p
targetCategory	28.2	2	4757.1	< .001
absZ	10.0	1	21.5	0.005

Note. Satterthwaite method for degrees of freedom

Table S3.2.5

Fixed Effects Parameter Estimates

				95% Confidence Interval				
Names	Effect	Estimate	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	2.1537	0.0993	1.959	2.3483	20.1	21.69	< .001
targetCategory1	2 - 1	-0.1358	0.0200	-0.175	-0.0966	4757.6	-6.80	< .001
targetCategory2	3 - 1	-0.1230	0.0200	-0.162	-0.0838	4756.5	-6.16	< .001
absZ	absZ	-0.0652	0.0206	-0.106	-0.0248	21.5	-3.17	0.005

Table S3.2.6

Model Info

	Info
Estimate	Linear mixed model fit by ML
Call	signedConfidence ~ 1 + targetCategory + zScore+(1 | id)+(0+zScore | id)
AIC	19537.3830
BIC	19582.7176
LogLikel.	-9761.6915
R-squared Marginal	0.0452
R-squared Conditional	0.1690

Table S3.2.7

Fixed Effect Omnibus tests

	F	Num df	Den df	p
zScore	29.6	1	20.0	< .001
targetCategory	74.0	2	4756.1	< .001

Note. Satterthwaite method for degrees of freedom

Table S3.2.8

Fixed Effects Parameter Estimates

Names	Effect	Estimate	SE	95% Confidence Interval		df	t	p
				Lower	Upper			
(Intercept)	(Intercept)	-0.274	0.1541	-0.575	0.0284	20.0	-1.78	0.091
zScore	zScore	0.283	0.0520	0.181	0.3847	20.0	5.44	< .001
targetCategory1	2 - 1	0.402	0.0647	0.275	0.5283	4756.1	6.21	< .001
targetCategory2	3 - 1	0.788	0.0647	0.661	0.9144	4756.1	12.16	< .001

Table S3.2.9

S3. Model Comparison Outputs

Model	**BIC**	**ΔBIC**	**R^2**
Direction Judgment			
direction ~ 1 + (1 | id)	6277.514		0.101
direction ~ 1 + zScore + (1 + zScore | id) + (1 | id)	6204.3135	-73.20	0.1228
direction ~ 1 + zScore + targetCategory + (1 + zScore | id) + (1 | id)	6087.4298	-116.8	0.1502
direction ~ 1 + zScore*targetCategory + (1 + zScore | id) + (1 | id)	6082.0694	-5.36	0.1524
Confidence Rating			
confidence ~ 1 + (1 | id)	9603.766		0.134
confidence ~ 1 + absolute z-score + (1 + absZ | id) + (1 | id)	9613.1321 5	9.37	0.1369
confidence ~ 1 + targetCategory + absZ + (1 + absZ | id) + (1 | id)	9532.8467	-80.29	0.1552
confidence ~ 1 + targetCategory*absZ + (1 + absZ | id) + (1 | id)	9545.7819	12.94	0.0201
Composite Score			
signedConfidence ~ 1 + (1 | id) + (1 + zScore | id)	19409.125 7		0.0788
signedConfidence ~ 1 + zScore + (1 + zScore | id) + (1 | id)	19340.4	-68.73	0.1
signedConfidence ~ 1 + zScore + targetCategory + (1 + zScore | id) + (1 |id)	19241.308 6	-99.09	0.1226
signedConfidence ~ 1 + zScore*targetCategory + (1 + zScore | id) + (1 | id)	19237.912 2	-3.40	0.1265

Table S3.3.1. Model comparisons for Experiment 1

Model	**BIC**	**ΔBIC**	**R^2**
Direction Judgment			
direction ~ 1 + (1 | id)	6589.384		0.0799
direction ~ 1 + zScore + (1 + zScore | id) + (1 | id)	6483.5109	-105.8	0.1076
direction ~ 1 + zScore + targetCategory + (1 + zScore | id) + (1 | id)	6324.637	-158.8	0.14
direction ~ 1 + zScore*targetCategory + (1 + zScore | id) + (1 | id)	6340.6696	16.03	0.1402
Confidence Rating			
confidence ~ 1 + (1 | id)	8364.852		0.353
confidence ~ 1 + absZ + (1 + absZ | id) + (1 | id)	8351.78319	-13.07	0.37967
confidence ~ 1 + targetCategory + absZ + (1 + absZ | id) + (1 | id)	8312.6591	-39.12	0.3881
confidence ~ 1 + targetCategory*absZ + (1 + absZ | id) + (1 | id)	8327.9571	15.30	0.3889
Composite Score			
signedConfidence ~ 1 + (1 | id) + (1 + zScore | id)	19828.297		0.114
signedConfidence ~ 1 + zScore + (1 + zScore | id) + (1 | id)	19711.5344	-116.7	0.1432
signedConfidence ~ 1 + zScore + targetCategory + (1 + zScore | id) + (1 | id)	19582.7176	-128.8	0.169
signedConfidence ~ 1 + zScore*targetCategory + (1 + zScore | id) + (1 | id)	19599.0251	16.31	0.1691

Table S3.3.2. Model comparisons for Experiment 2

S4. Correlations Between Individual Slopes and Coefficients of Variation (CV)

There was no significant correlation between individual slopes from the confidence model and CVs pooled in either experiment, $r = .338$, $p = .14$, $BF_{10} = .748$ in Experiment 1, and $r = .08$, $p = .737$, $BF_{10} = .292$ in Experiment 2.