Monitoring line length reproduction errors

Yalçın Akın Duyan, Fuat Balçı*

Department of Psychology & Research Center for Translational Medicine, Koç University, Rumelifeneri Yolu, Saryer, Istanbul 34450, Turkey

ARTICLE INFO

Keywords:
Error monitoring
Line reproduction
Magnitude representations
Confidence judgments

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.

1. 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 & Balçı, 2017), numerical estimates (Duyan & Balçi, 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.

* Corresponding author.
E-mail address: fbalci@ku.edu.tr (F. Balci).

https://doi.org/10.1016/j.concog.2019.102831
Received 10 January 2019; Received in revised form 6 September 2019; Accepted 28 September 2019
Available online 04 November 2019
1053-8100/ © 2019 Elsevier Inc. All rights reserved.
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.

2. Experiment 1

2.1. Participants

Twenty undergraduate students from Koç 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 includes 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.

2.2. 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).

2.3. 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).

2.4. 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 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 deviation 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 Supplementary online materials (SOM, 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.

2.5. Results

2.5.1. 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, $F(1.5, 28.49) = 7.284, p = .005, \eta^2 = 0.277; [M \text{ (short)} = 2.211, M \text{ (medium)} = 2.012, M \text{ (long)} = 2.1)]$. 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, $\Delta BIC = 5.4, f^2_{\text{fixed}} = 0.129$). Hence the formula for our best-fit model was: Error Directionality Judgment $- 1 + \text{Target Category} * z\text{-Score}_{\text{reproduction}} + (1 | \text{Participant}) + (z\text{-Score}_{\text{reproduction}} | \text{Participant})$. The model resulted in a significant effect of $z$-score, $\beta = 0.059, SE = 0.013, p < .001$, 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 overshoot the target increased with longer targets ($\hat{\beta}_{\text{long-medium}} = 0.129, SE = 0.017, p_{\text{bonferroni}} < 0.001; \hat{\beta}_{\text{medium-short}} = 0.069, SE = 0.017, p_{\text{bonferroni}} < 0.001; \hat{\beta}_{\text{short-long}} = 0.198, SE = 0.017, p_{\text{bonferroni}} < 0.001$). 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 ($\beta = 0.08, SE = 0.16, p < .001$) and long targets ($\beta = 0.072, SE = 0.16, p < .001$) but not in the short target ($\beta = 0.026, SE = 0.16, p = .095$).

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 $- 1 + \text{Target Category} + \text{Absolute z-Score}_{\text{reproduction}} + (1 | \text{Participant}) + (\text{Absolute z-Score}_{\text{reproduction}} | \text{Participant})$. 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 ($\hat{\beta}_{\text{long-short}} = -0.214, SE = 0.024, p_{\text{bonferroni}} < 0.001; \hat{\beta}_{\text{medium-short}} = -0.202, SE = 0.024, p_{\text{bonferroni}} < 0.001$).

### 2.5.2. Composite measure analysis

For each participant, we first categorized first 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 overshoot 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, $\Delta BIC = 3.4, f^2_{\text{fixed}} = 0.01$). The formula for the best fitting model was Confidence-Directionality Judgment $- 1 + \text{Target Category} * z\text{-Score}_{\text{reproduction}} + (1 | \text{Participant}) + (z\text{-Score}_{\text{reproduction}} | \text{Participant})$. 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 ($F(2,4545.4) = 10.2, p < .001; \hat{\beta}_{\text{long-short}} + \text{z-Score} = 0.276, SE = 0.07; p < .001; \hat{\beta}_{\text{medium-short}} + \text{z-Score} = 0.27, SE = 0.07; p < .001$). 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, $F(2,4554) = 58.9, p < .001; \hat{\beta}_{\text{medium-long}} = -0.49, SE = 0.07, p_{\text{bonferroni}} < 0.001; \hat{\beta}_{\text{short-medium}} = -0.25, SE = 0.07, p_{\text{bonferroni}} = 0.001; \hat{\beta}_{\text{short-long}} = -0.74, SE = 0.07, p_{\text{bonferroni}} < 0.001$) (for a summary of the results, see Table 1 and Fig. 2). The details of model outputs for Experiment 1 are presented in the Supplemental Online Material S1.

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 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 SOM 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

Table 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.

<table>
<thead>
<tr>
<th>Experiment 1</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>β</td>
<td>Standard error</td>
<td>df</td>
<td>t value</td>
<td>p</td>
</tr>
<tr>
<td>Direction judgment</td>
<td>0.059</td>
<td>0.012</td>
<td>20</td>
<td>4.89</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Confidence rating*</td>
<td>−0.047</td>
<td>0.018</td>
<td>20.9</td>
<td>−2.52</td>
<td>= 0.020</td>
</tr>
<tr>
<td>Composite score</td>
<td>0.245</td>
<td>0.05</td>
<td>19.9</td>
<td>3.859</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment 2</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>β</td>
<td>Standard error</td>
<td>df</td>
<td>t value</td>
<td>p</td>
</tr>
<tr>
<td>Direction judgment</td>
<td>0.07</td>
<td>0.012</td>
<td>21</td>
<td>5.87</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Confidence rating*</td>
<td>−0.065</td>
<td>0.021</td>
<td>21.5</td>
<td>−3.17</td>
<td>= 0.005</td>
</tr>
<tr>
<td>Composite score</td>
<td>0.283</td>
<td>0.052</td>
<td>20</td>
<td>5.44</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Fig 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.
that participants could never be exactly accurate.

3. Experiment 2

3.1. Method

3.1.1. Participants

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

3.1.2. 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.

3.2. Results

3.2.1. 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 \(\sim 1 + \text{Target Category} + z-\text{Score}_{\text{reproduction}} + (1 \mid \text{Participant}) + (z-\text{Score}_{\text{reproduction}} \mid \text{Participant})\). 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 (\(F(2,4758.1) = 89.6, p < .001; \beta_{\text{long-medium}} = 0.102, SE = 0.16, p < .001; \beta_{\text{medium-short}} = 0.116, SE = 0.16, p < .001; \beta_{\text{long-short}} = 0.218, SE = 0.34, p < .001\)). As for confidence ratings, the best-fit model was again Model 3 (vs. Model 4 as the second best fit model, \(\Delta BIC = 15.3, f^2_{\text{fixed}} = 0.618\)). 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 \(\sim 1 + \text{Target Category} + \text{Absolute z-score}_{\text{reproduction}} + (1 \mid \text{Participant}) + (\text{Absolute z-score}_{\text{reproduction}} \mid \text{Participant})\). The results showed that subjective confidence overall followed the degree of absolute errors (\(F(1, 21.5) = 10; \beta = 0.065, SE = 0.02, p = .005\)), and overall confidence level varied across targets (\(F (2,4757.1) = 28.2, p < .001; \beta_{\text{short-medium}} = 0.136, SE = 0.02, p < .001; \beta_{\text{short-long}} = 0.123, SE = 0.02, p < .001\)). Overall confidence was higher for shortest target length compared to the other target lengths.

3.2.2. 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, \(\Delta BIC = 16, f^2_{\text{fixed}} = 0.149\)) 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 SOM S4 for the results of the analogous analysis for confidence ratings).

4. 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, Genç, & 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 Balçi (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 Balçi (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, & Balçi, 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 (Balcí et al., 2011; Gorea & Sagi, 2000; Gorea, Caetta, & Sagi, 2005; Rahev & 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 judgments 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
confident 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 can also be 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.

5. Author note

This study will be a part of Yalçın Akın Duyan's PhD thesis.

Funding

This work was supported by Turkish Academy of Sciences (TÜBA GEBİP 2015) to FB.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.concog.2019.102831.

References


