The Cognitive Architecture of Temporal Metacognition

by

Tutku Öztel

A Dissertation Submitted to the

Graduate School of Social Sciences and Humanities

in Partial Fulfillment of the Requirements for

the Degree of

Doctor of Philosophy in Psychology

KOÇ ÜNİVERSİTESİ

18th September, 2024

The Cognitive Architecture of Temporal Metacognition

Koç University

Graduate School of Social Sciences and Humanities

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

Tutku Öztel

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 (supervisor): ____

Asst. Prof. Terry Eskenazi: ____

Asst. Prof. Yalçın Akın Duyan: ____

Asst. Prof. Efe Soyman: ____

Asst. Prof. Sertaç Üstün: ____

Dedication

I am dedicating my PhD dissertation to you, Doruk. I hope you will accept this as a wedding gift.

ABSTRACT

The Cognitive Architecture of Temporal Metacognition

Tutku Öztel
Doctor of Philosophy
18th September, 2024

Metacognition, which is the introspective ability to reason about one's own cognition, has long been investigated in perceptual and memory domains. Recent studies have investigated this ability within metric domains which include time, space and numerosity. While these studies demonstrate the human capability of keeping track of one's own cognition as it relates to these metric domains, no study thus far has investigated the cognitive architecture of this generalized metric error monitoring ability. The present dissertation investigates the different influences on the temporal error monitoring ability as a special case of metric metacognition, which includes keeping track of one's own temporal error magnitude (i.e., the extend of under or over estimating a target duration) and direction (i.e., under or overshooting the target duration) under two categories including (1) the exogenous and (2) endogenous factors. Study 1 (Chapter 2) investigates the effect of these two kinds of influences in a complementary way by testing the endogenous and exogenous feedback on global ability of temporal error monitoring. As a potential exogenous factor, Study 2 (Chapter 3) investigates whether the diminishing effect of working memory (WM) load documented previously for the perceptual metacognition applies to the temporal error monitoring ability. Finally, Study 3 (Chapter 4) investigates different endogenous factors (e.g., motor action, error ownership and belief of error owning) on the temporal error monitoring ability.

Overall, results of the first study demonstrated no consistent effect of external feedback on trial-by-trial (online) temporal error monitoring ability while also pointing to the distinction between the online and global temporal error monitoring ability (i.e., reporting the average bias in the overall timing performance). Further, the second study pointed to no diminishing effect of WM load on online temporal error monitoring ability. Finally, the third study demonstrated the differential effect of motor action and belief of owning the temporal error on confidence and error directionality judgments as the two components of temporal error monitoring ability. Altogether, the present dissertation (1) demonstrated the distinction between online and global forms of temporal awareness (2) highlighted differential characteristics of temporal awareness with respect to the perceptual metacognition and (3) discovered the distinct characteristics of error magnitude and direction monitoring as two components of temporal error monitoring ability.

ÖZETÇE

Zamansal Üstbilişin Bilişsel Mimarisi

Tutku Öztel

18 Eylül, 2024

Üstbiliş, kişinin kendi bilişsel süreçleri üzerine düşünme yetisi, literatürde baskın olarak algısal ve bellek alanlarında araştırılmaktadır. Son yıllarda yapılan çalışmalar, bu yetiyi zaman, mekan ve çokluk gibi metrik alanlarda incelemiştir. Bu çalışmalar, insanın kendi bilişsel süreçlerini bu metrik alanlarda takip edebilme yeteneğini gösterse de, bugüne kadar bu genelleştirilmiş metrik hata izleme yeteneğinin nihai bilişsel yapısı üzerine kapsayıcı bir çalışma yapılmamıştır. Bu tez, metrik üstbilişin özel bir türü olarak zamansal hata izleme yeteneği üzerindeki farklı etkileri araştırmaktadır. Bu, kişinin kendi zamansal hata büyüklüğünü (yani, tahminlerin hedef süreden sapma miktarı) ve yönünü (yani, hedef sürenin altında mı yoksa üstünde mi kaldığını) (1) dışsal ve (2) içsel faktörler olmak üzere iki temel kategori altında incelemeyi içermektedir. Birinci çalışma (Bölüm 2), bu iki tür etkinin etkisini bütünsel bir şekilde araştırarak genel zamansal hata izleme yeteneği üzerinde endojen (içsel) ve eksojen (dışsal) geri bildirimin etkisini test etmektedir. Potansiyel bir dışsal faktör olarak, ikinci çalışma (Bölüm 3), algısal üstbiliş için daha önce gösterilmiş olan işler bellek (İB) yükünün azaltıcı etkisinin zamansal hata izleme yeteneğine uygulanıp uygulanamayacağını araştırmaktadır. Son olarak, üçüncü çalışma (Bölüm 4), zamansal hata izleme yeteneği üzerindeki farklı içsel faktörleri (örneğin, motor eylem, hata sahipliği ve hata sahipliğine yönelik inanç) araştırmaktadır. Genel olarak, birinci çalışmanın sonuçları, dışsal geri bildirimin deneme bazlı (çevrimiçi) zamansal hata izleme yeteneği üzerinde tutarlı bir etkisi olmadığını göstermiş ve aynı zamanda çevrimiçi ve deneyin sonunda genel zamanlama performansını rapor etmeyi ifade eden genel zamansal hata izleme yeteneği arasındaki ayrımı işaret etmiştir. Ek olarak, ikinci çalışma, farklı türdeki İB yükünün çevrimiçi zamansal hata izleme yeteneği üzerinde azaltıcı bir etkisi olmadığını göstermiştir. Son olarak, üçüncü çalışma, motor eylem ve zamansal hatanın sahiplenildiği inancının, zamansal hata izleme yeteneğinin iki bileşeni olarak güven ve hata yönlülüğü yargıları üzerindeki farklı etkilerini göstermiştir. Sonuç olarak, üç çalışmada bu tez (1) çevrimiçi ve genel zamansal farkındalık biçimleri arasındaki ayrımı ortaya koymuş (2) zamansal farkındalığın algısal üstbilişe göre farklı özelliklerini vurgulamış ve (3) zamansal hata izleme yeteneğinin iki bileşeni olarak hata büyüklüğü ve yön izleme özelliklerinin farklı olduğunu keşfetmiştir.

Author's Note: Chapter 2 has been published in *Quarterly Journal of Experimental Psychology*. Chapter 3 has been published in *Psychological Reports*. Chapter 4 is accepted for publication at *Attention, Perception & Psychophysics.*

Öztel, T., & Balcı, F. (2023). Humans can monitor trial-based but not global timing errors: Evidence for relative judgements in temporal error monitoring. *Quarterly Journal of Experimental Psychology*, *76*(9), 2155-2163.

Öztel, T., & Balcı, F. (2023). Temporal error monitoring does not depend on working memory. *Psychological Reports, 00332941231187121.*

Öztel, T., & Balcı, F. (Accepted). Temporal error monitoring: Does ownership of errors matter? *Attention, Perception & Psychophysics*

ACKNOWLEDGEMENT

Foremost, I would like to express my sincere gratitude to my thesis co-advisors Prof. Fuat Balcı for always supporting me on my PhD journey, even from far away. Without his support and guidance since 2015 in my mid-undergraduate, I could not achieve what I achieved today. His training style and point of view in scientific practice will always be a role model throughout my career. Having spent nearly a decade together in a supervisor-student relationship, he will always remain not only an invaluable collaborator and lifelong mentor; but also, an elder brother to whom I can consistently turn for guidance. Additionally, I would like to thank the rest of my committee members for their intellectual contributions with their valuable feedback on this dissertation.

I would also like to thank Prof. Sami Gülgöz from Koç University who always made me feel his support through all the difficult times throughout my PhD journey.

I want to thank my dearest friends from Koç University Psychology graduate program Hazal Kartalkanat, Gamze Turunç Bayrakdar, Demet Özer and Sura Ertaş for all the support, intellectual (and non-intellectual) contributions, and of course, all the play and laughter. They are one of my greatest gains from Koç University. I especially want to thank Hazal for making a great neighbor and for turning my mondays from a syndrome into something to look forward to with our week-opening morning walks and the deep and not-so-deep conversations we shared. Genuine thanks go to another dearest ones Zeynep Demirbilek Akçay and Elif Dinek Yılmaz for all the fun we have had together since 2013, even when we are far away. I feel blessed with having these great women in my life, I love you guys!

Next, I want to thank my parents and grandmother for their unconditional love and sharing their wisdom with me in every period of my life. No need to say much, I love you beyond words.

Finally, I want to thank you, my dear fiancé Doruk Engin, for showing your strong personality while getting along with me with your great calmness and patience, for always being there to cry and laugh, support and criticize, and for being my ultimate best friend forever. You are another one from whom I learned a lot about various perspectives of life ranging from sports to arts, history to celebrity gossip, politics to traveling. I am a big fan of yours.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

The error monitoring aspect of metacognition refers to the ability to report and reflect upon self generated errors without necessitating external feedback (Yeung & Summerfield, 2012).

The error monitoring ability is typically investigated within the composition of performances in two tasks: In the first order tasks, participants perform a basic discrimination task which includes distinguishing "signal" (or target) from "noise" (or non-target). In the second order task, participants report an evaluation of their own performance in the first order task. Gathering confidence ratings is one of the most popular methods for evaluating first order performance, where the degree of match between the confidence ratings and actual performance is used to quantify error monitoring performance. Based on such metrics, previous work has indeed demonstrated that human participants can match their trial-by-trial confidence ratings to their categorization accuracy in perceptual decision making (e.g., Fleming et al., 2012b; 2012c; also see: Fleming et al., 2012a) and memory domains (e.g., Vandenbroucke et al., 2014) in an online fashion. But virtually in all of these domains, error monitoring performance is investigated in two alternative forced choice (2AFC) tasks where the decision accuracy is binary (e.g., correct or incorrect). Thus, this paradigmatic approach falls short of capturing most daily life situations that involve errors with magnitude and direction. For example, a person can substantially overcook a dish or slightly undercook it. Both of these outcomes contain an error that has different directions and magnitudes.

Akdoğan and Balcı (2017), for the first time, investigated whether human participants can keep track of their timing errors by investigating this ability utilizing two different indices that captures the metric aspects of their timing errors: the magnitude and direction of the error. Crucially, this finding has been replicated many times by the subsequent studies (Kononowicz et al., 2019; see also: Öztel, Eskenazi & Balcı, 2020; Yallak & Balcı, 2020) and its scope has been expanded to include also numerosity (Duyan & Balcı, 2018;2019) and line length judgements (Duyan & Balcı, 2020 - for a detailed discussion, see: Yallak & Balcı, 2021), composing of a unitary form of metric error monitoring ability. However, this metric error monitoring ability has been found to carry distinctive features than the perceptual error monitoring ability, such that these two kinds of metacognitive processes are not correlated (Yallak & Balcı, 2021).

While the differentiation is documented (Yallak & Balcı, 2021), one still open question is how the cognitive architectures are differential for perceptual and metric error monitoring abilities. The perceptual error monitoring is found to be affected by many different factors such as working memory load (Maniscalco & Lau, 2015), endogenous and exogenous factors such as (1) motor actions and (2) psychophysical characteristics of the stimuli, and external feedback (3). Inspired by the previous literature on perceptual error monitoring, the current dissertation focuses on two potential guiding factors: (1) exogenous factors in the form of externally presented or (2) endogenous factors, in the form of internally represented. Overall, in three independent studies, this dissertation aims to elucidate the cognitive architecture of temporal error monitoring ability by investigating the effect of these two categories of informational sources on the temporal error monitoring; as this ability has recently demonstrated to differ from the traditionally investigated domains of metacognition such as meta-perception and metamemory (Yallak & Balcı, 2021).

Along with the online/trial-by-trial error monitoring, global error monitoring (i.e., being aware of average bias) can be regarded as another type of metacognitive process. Different from the trial-by-trial (online) error monitoring, individuals retrospectively evaluate their own performance throughout the experiment at the end of the given block or experimental session. Chapter 2 describes a study that investigated how various online exogenous influences such as trial by trial feedback as well as confidence and error direction judgements affect the global/overall monitoring of average error direction in the temporal settings.

Another potential extraneous factor that has been documented for perceptual metacognition (e.g., Maniscalco & Lau, 2015) is the working memory (WM) that is unique to information manipulation demand such that increased manipulation demand was detrimental for metacognitive accuracy. Chapter 3 presents a study which explores whether this detrimental effect can also apply to the temporal error monitoring ability with three different parameterization of the WM load.

Can metacognitive abilities be guided by the endogenous factors? Previous literature discusses metacognitive processes are fed by either the stimulus's psychophysical properties (i.e., exogenous cues; e.g., Fleming et al., 2015) or the agent's internal feedback mechanism that stems from motor actions (i.e., endogenous cues; Palser et al., 2018). Chapter 4 describes a study that investigated differential effects of endogenous factors (e.g. motor activation while making a response) on the two temporal error monitoring indices (confidence and error direction).

Chapter 2

HUMANS CAN MONITOR TRIAL-BASED BUT NOT GLOBAL TIMING ERRORS: EVIDENCE FOR RELATIVE JUDGEMENTS IN TEMPORAL ERROR MONITORING

2.1. Abstract

Humans can monitor the magnitude and direction of their temporal errors in individual trials. Based on the predictions of our model of temporal error monitoring that rely on a relative comparison of internal clock readings, we predict that participants would monitor their timing errors in individual trials, but not the direction of their global timing errors without external feedback. One study has indeed found that accurate self-monitoring of average timing biases required external feedback with directional information. The current study investigates how different sources of feedback (i.e., internal or external) affect performance in the self-monitoring of average timing bias. Four groups of participants were tested in a temporal reproduction task. Participants in the self-evaluation condition evaluated the direction and size of their time reproduction errors in individual trials. In the accurate feedback condition, participants received explicit trial-based feedback regarding the direction of their error while participants in the partially accurate feedback condition received trial-based feedback according to the accuracy of short-long judgements of another participant in the self evaluation condition. Participants in the control condition reproduced only the target duration without making any judgements regarding their reproduction performance or receiving any external feedback about it. Results showed that while participants accurately monitor timing errors in individual trials, in none of the experimental conditions were they more accurate than the chance level in terms of evaluating the direction of their average temporal bias. We discuss these results in terms of the temporal error monitoring model introduced by Akdoğan and Balcı. Thus, our findings suggest that external directional feedback does not have any informational value for global temporal bias judgements above and beyond internal self-monitoring.

2.2. Introduction

Metacognition refers to “cognition about cognition”/“higher order cognition” (Flavell, 1979), which is considered to be a part of consciousness (Fleming, Dolan, & Frith, 2012; Lau & Rosenthal, 2011; Lau, 2019). The error monitoring aspect of metacognition refers to the ability to report and reflect upon self-generated errors without necessitating external feedback (Yeung &

Summerfield, 2012). The current study investigated the effects of different types of feedback (i.e., no feedback, fully accurate feedback, partially accurate feedback, and internal feedback by self evaluation) on the self-monitoring of average error direction in the temporal settings.

The error monitoring ability is typically investigated by the trial-by-trial match between a decision performance and a corresponding confidence rating regarding the accuracy of that decision. Thus, a typical error monitoring hypothesis would predict a positive correlation between the decision performance and the confidence level. Based on such metrics, previous work has indeed demonstrated that human participants can match their trial-by-trial confidence ratings to their decision accuracy in perceptual decision making (e.g., Fleming & Dolan, 2012; Fleming, Huijgen, & Dolan, 2012; also see: Fleming, Dolan, & Frith, 2012) and memory domains (e.g., Vandenbroucke et al., 2014). But virtually in all of these domains, error monitoring performance is investigated in two alternative forced choice (2AFC) tasks where the decision accuracy is binary in nature (e.g., correct or incorrect). Thus, this paradigmatic approach falls short of capturing most daily life situations that involve errors with a magnitude and direction. For example, a person can substantially overcook a dish or slightly undercook it. Both of these outcomes contain an error that has different directions and magnitudes.

Akdoğan and Balcı (2017), for the first time, investigated whether human participants can keep track of their timing errors by investigating this ability utilising two different indices that capture the metric aspects of their timing errors. Participants were tested in a temporal reproduction paradigm where they were asked to reproduce a previously presented target duration as accurately as possible. After each trial, participants rated their level of confidence regarding the accuracy of their temporal reproductions in terms of their proximity to the target duration (i.e., confidence index). Finally, participants classified their reproductions as being shorter or longer than the target duration (i.e., error direction index). Results demonstrated a clear trial-by-trial match between the signed confidence judgements and the signed temporal errors, suggesting a temporal error monitoring ability. This finding has been replicated many times by subsequent studies (Doenyas et al., 2019; Kononowicz et al., 2019; Yallak & Balcı, 2021; Öztel et al., 2020b) and its scope has been expanded to include also numerosity (Duyan & Balcı, 2018, 2019) and line length judgements (Duyan & Balcı, 2020—for a detailed discussion see Yallak & Balcı, 2021).

Along with online/trial-by-trial error monitoring, global error monitoring can be regarded as another type of metacognitive process. In the latter, individuals retrospectively evaluate their overall performance throughout the experiment either at the end of the given block or the experimental session. A recent study by Rouault and colleagues (2019) found that when no external feedback was provided regarding their discrimination accuracy, participants reported their overall performance as being worse than it was in a perceptual discrimination task. In addition, the external feedback fine-tuned the error monitoring of average accuracy and participants with better online error monitoring were better at evaluating their overall accuracy. Together, these findings point to the enhancing effect of internally driven error monitoring and external feedback on the evaluation of overall performance in perceptual decision making (Rouault et al., 2019).

Thus far, research in the temporal error monitoring domain showed that humans can keep track of the magnitude and direction of errors in their trial-based timing behaviour. Whether human participants can also keep track of their average timing performance (i.e., bias) and whether this ability relies on external feedback are outstanding research questions that are investigated in this study. A recent study by Riemer and colleagues (2019) investigated the effect of external feedback on timing bias judgements in a temporal reproduction task. In absolute feedback condition, after each reproduction participants received non-directional feedback regarding the amount of deviation between their reproduction and the target duration. In the signed feedback condition, this feedback was also coupled with information regarding the direction of errors (e.g., early vs. late). At the end of the experiment, participants reported on average, how much and towards which direction their reproductions fell in relation to the target duration. Results showed that participants in the signed feedback condition reported their global temporal bias more accurately compared with the absolute feedback condition (Riemer et al., 2019). As the trial-by-trial temporal error monitoring is assumed to take place in all conditions implicitly, the only difference between the experimental conditions is the nature of the feedback. Authors thus concluded that monitoring of the overall timing biases (i.e., global error monitoring) does not benefit from the ability to self-monitor error direction in individual trials. Instead, the effect is due to the fact that global error monitoring requires external feedback that includes directional information regarding the error that is committed. Complementing the above-mentioned results, Overbeek and colleagues (2005) demonstrated that the trial-by-trial external feedback resulted in comparable performance monitoring signals (i.e., ERN) for obsessive compulsive disorder patients (who normally demonstrate hyperactivity in the ERN signals, for example, Fitzgerald et al., 2005) compared with the healthy controls, pointing to the improving effect of feedback on error monitoring. This study also demonstrates the improving effect of external feedback on error monitoring performance.

As suggested by Riemer and colleagues (2019), the reason that monitoring overall timing biases requires trial-by-trial directional feedback might be because internal error representations are subject to memory interference. But it is also possible that metric error monitoring is only achievable in individual trials due to mechanistic constraints imposed by relative comparisons as in the model of Akdoğan and Balcı (2017). This model depicts the error monitoring process as a comparison between the threshold first crossing times of the two independent temporal integrators for (a) temporal estimation and (b) behavioural action. Hence, if the speed of both clocks changes towards a direction resulting in an overall bias, although this model still accounts for the relative comparison, it is blind to this overall bias (for a more detailed discussion see Öztel & Balcı, 2020a). In line with the null predictions of Akdoğan and Balcı's (2017) model regarding temporal bias monitoring, we have recently shown that although participants can keep track of their timing errors on a trial-by-trial basis, they cannot keep track of biases induced by external stimulus characteristics in the bisection task (Öztel & Balcı, 2020a).

The present study investigated the effect of different types of feedback (internally driven self-evaluation, fully vs. partially accurate explicit feedback, control condition with neither element) on the timing performance and global monitoring of overall timing biases. At the end of the experiment, participants reported whether their reproductions were on average shorter or longer than the target duration. Specifically, based on the model described in Akdoğan and Balcı (2017), we predicted that participants would only be able to monitor their trial-by-trial temporal errors. However, this temporal error monitoring ability will not generalise itself in the global monitoring of the mean timing errors in the absence of feedback. The implicit trial-by-trial temporal error monitoring is assumed for all conditions.

As metric error monitoring would be assumed in all of these four conditions, any difference between the no-feedback and self-evaluation condition in terms of the global error monitoring performance would be attributed to making metric error monitoring explicit. More specifically, we predicted similar performance in the feedback and self evaluation conditions if explicit metric error monitoring (via explicit report of trial-by-trial confidence and error direction in the self-evaluation condition) can take a feedback-like form for the global error monitoring judgements (as in the directional feedback condition in Riemer et al., 2019). Finally, we predicted that the global error monitoring performance in the control condition should be diminished compared with the feedback and self-evaluation conditions as we expect no explicit feedback-like mechanism to be held in the control condition.

This study expands the recent findings (e.g., Riemer et al., 2019) by investigating the effect of different feedback types on global error monitoring in a more comprehensive fashion that addresses internal, in addition to external, influences. Furthermore, this study expands the scope of the recent findings by becoming grounds for the direct comparison between trial-based and global monitoring of temporal errors.

2.3. Method

2.3.1. Participants. A total of 168 adults participated in this study; 131 of them were Koç University undergraduate students who received extra course credit for their participation and 38 were volunteers who responded to our ads on social media. Out of 168, 28 left the experiment without completing it; 30 were discarded from the formal analyses as their reproduced durations were shorter than 1s or longer than 4s in more than half of the total experimental trials (the target was 2 s). These exclusions aimed at preventing any artificial support for the hypothesis that the participants could correctly report their mean error direction. Thus, 110 were included in the formal analyses (37 male, M_{age} = 22.12, SD_{age} = 3.79; 3 did not report age, 5 were left-handed, and 1 reported being ambidextrous). Each condition consisted of 27 (except for the partially accurate feedback condition which consisted of 29) participants. Note that this study was conducted online during the COVID-19 pandemic. Thus, the factors behind the relatively high attrition rate might include internet interruptions, difficulty in paying attention to the task, environmental disruptions, and feeling extra comfortable with leaving the experiment as they are not accompanied by an experimenter.

2.3.2. Stimuli and apparatus. The target duration of 2,000ms was presented with a 6.1×6.1cm red square centered on the screen. The reproduction stimulus was a blue square of the same size. The experiment was built on Psychopy Builder mode (Peirce et al., 2019). The stimulus presentation and response recording were controlled via the Pavlovia online experiment platform (Bridges et al., 2020).

2.3.3. Procedure. The task was temporal reproduction. Participants were randomly assigned to the four conditions. Before the testing session, participants completed five practice trials to learn the target duration and how a temporal reproduction would be performed. The testing session started with presentation of the constant target duration of 2,000ms, which was presented at every 10 trials. Then, participants were asked to start the reproduction with a space button press and terminate the reproduction by pressing the button again when they thought the elapsed time was equal to the target duration.

After each time reproduction, participants in the self evaluation condition were asked to report how confident they were regarding the accuracy of their reproduction compared with the target duration by pressing 1, 2, or 3 buttons (1=low confidence and 3=high confidence) on their keyboard. After the confidence rating, participants reported whether they thought their reproduction was shorter or longer than the target duration by pressing the “S” or “L” buttons (“S” for shorter and “L” for longer; short–long judgements). In the feedback conditions (fully and partially accurate), participants received feedback regarding the direction of deviation from the target duration in the middle of the screen (“Your reproduction was SHORT [LONG] compared with the target duration” in Turkish); there was no self-evaluation in these conditions.

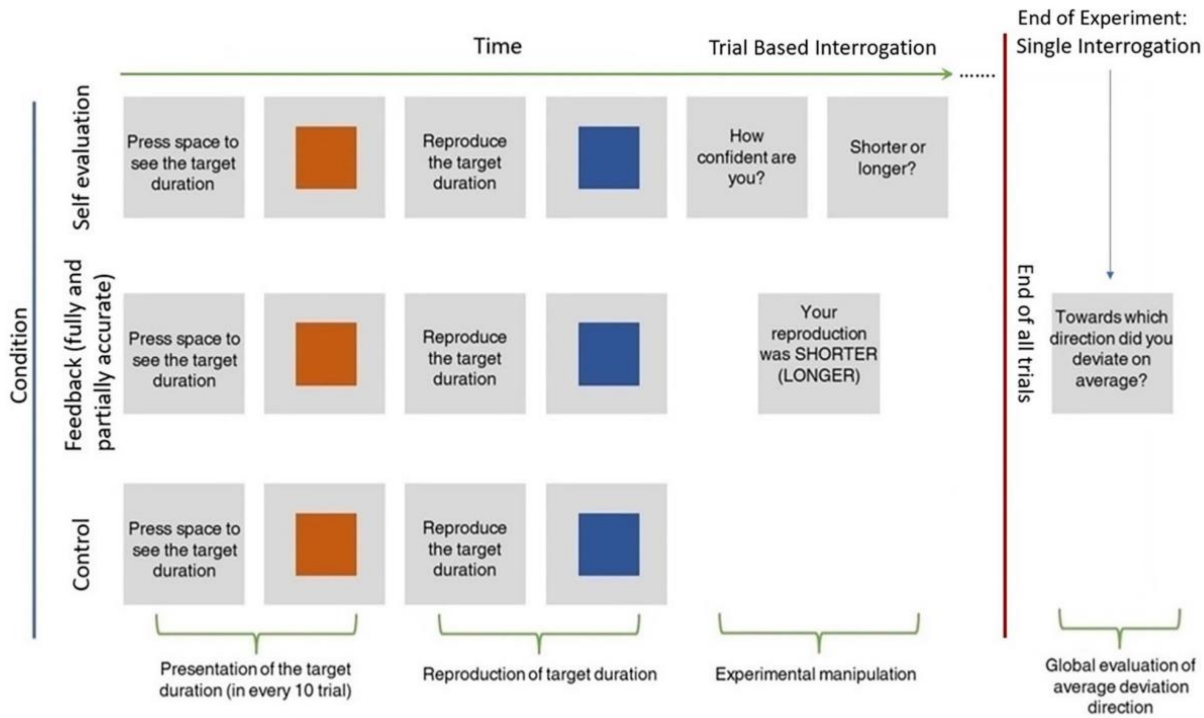


Figure 2.1. Schematic representation of a trial in all conditions.

The distinct nature of the feedback provided in fully and partially accurate feedback conditions was as follows. Participants in the fully accurate feedback condition were provided with always accurate feedback according to the empirical temporal error direction (i.e., whether the temporal reproduction in the given trial was shorter or longer than the target duration). Participants in the partially accurate feedback condition received directional feedback according to the proportions of the short–long responses given by the participants in the self-evaluation condition (i.e., short–long feedback probabilities were yoked from another participant's responses in the self-evaluation condition). In this way, each participant in the partial feedback condition was matched with a participant in the self-evaluation condition (i.e., participants that relied on their own internal and imperfect feedback mechanism) in terms of the accuracy of the feedback that was being utilised. Thus, this procedure helped us to control for the discrepancy between the fully accurate feedback condition and the self-evaluation condition in terms of the accuracy of the feedback (i.e., external information that provides feedback with perfect accuracy and internal information that provides feedback with imperfect accuracy due to noise in the information processing; e.g., Kepecs et al., 2008). Note that fully random feedback would not control for the discrepancy between the fully accurate feedback and self-evaluation conditions (where the feedback was internally driven from a noisy source) in terms of feedback accuracy.

At the end of the experiment, participants in all conditions were asked to report towards which direction on average they thought their reproductions deviated from the target duration throughout the experiment using the "S" and "L" buttons ("S" for "short" and "L" for "long"; global short–long judgement). We did not provide error magnitude information in any of the feedback conditions as we assessed the effect of feedback/self-evaluation on only the correct monitoring of the global error directions without global error magnitudes. In all conditions, the key mappings were visible to the participant until a response was recorded and participants were instructed not to count or utilise any kind of chronometric strategy (Rattat & Droit-Volet, 2012). Each condition consisted of 200 trials. Figure 2.1. illustrates the schematic representation of a trial in each condition.

2.4. Analytical approach. Before data analysis, we combined confidence ratings with the short–long judgements in the self-evaluation condition in a way that combines the directional and magnitudinal information of the temporal errors in a single variable. For example, if the reproduction was classified as "short" with "low" confidence, the response was coded as −3 (minus for "short" and 3 for "high amount of temporal error"). On the contrary, if the reproduction was classified as "long" with "high" confidence, the response was coded as +1 (positive for "long" and 1 for "little amount of temporal error"). Consequently, a signed confidence judgement variable with six levels ranging between −3 (too short, low confidence) and 3 (too long, low confidence) was obtained, without the inclusion of zero. Then, we z-score transformed the reproduced durations and finally, we applied a listwise deletion to reproduced durations that are below −3 and above 3 SDs. Note that the removal of the outlier trials was applied only for investigation of the temporal error monitoring performance in the self-evaluation condition (where participants evaluated their temporal reproduction performance on a trial-by-trial basis). On the contrary, we did not remove these trials for analysis of the global error monitoring accuracy in any of the experimental conditions to prevent any biases in this index.

2.5. Results

2.5.1. Temporal reproduction performance. To investigate whether there was an overall bias in the timing behaviour in each experimental condition, we investigated whether reproduced durations in each experimental condition deviated from the target duration. Separate one sample t-tests revealed that the mean reproduced durations were significantly different than the target duration only in fully and partially accurate feedback conditions— fully accurate feedback condition: $t(26) = 2.19$, $M = 2.06$, $SD = 0.14$, $p = .04$, Cohen's d = 0.42, Bayes $factor_{10} = 1.56$, which demonstrates the data are 1.56 times more likely under the alternative hypothesis than the null hypothesis; partially accurate feedback condition—$t(28)=2.5$, $M=2.37$, $SD=0.78$, $p=.018$, Cohen's d=0.47, Bayes $factor_{10}=2.74$, depicting that the data are 2.74 times more likely under the alternative hypothesis than the null hypothesis. Note that for all Bayes factors we used a default prior of 0.707 as the scale parameter (r) of a Cauchy distribution. However, when alpha (α) levels are corrected for the Type I error rate using Holm–Bonferroni correction, these effects disappeared (where Holm–Bonferroni corrected $\alpha=0.0125$ and 0.016 for partially and fully accurate feedback hypotheses, respectively). Moreover, the mean reproduced durations were not significantly different from the target duration in either self-evaluation or the control condition (all ps>0.05; all Bayes $factor_{10}<1$). Together, these results reveal that none of the experimental manipulations resulted in timing bias compared with the target duration.

Furthermore, to investigate whether the timing behaviours in each experimental condition are comparable, we compared the reproduced durations across four experimental conditions with a between-subject one-way analysis of variance (ANOVA) and found no difference, Welch's $F(3, 50.94)=1.67$, $p=.19$. This result highlights that the manipulation applied in each condition yielded similar first-order timing performance. Figure 2.2.a illustrates the mean reproduced durations across experimental conditions.

Finally, we compared the coefficient of variation (CV, as an index of precision of the temporal reproductions where lower values depict higher precision) of the reproduced durations across the experimental conditions. Here, if results point to a diminished timing precision (i.e., higher CV) in any of the experimental conditions (coupled with disrupted global error monitoring), its potential reflection in the global reports of error directions would hint at a mediating mechanism critical for the effect of our experimental manipulations on the global error monitoring ability. A between-subject one-way ANOVA revealed a statistical difference between the experimental conditions in terms of CV values, Fisher's $F(2,78)=4.58$, $p=.006$. Holm–Bonferroni corrected post hoc comparisons revealed only higher mean CV value for the partially accurate feedback condition compared with the control condition, $M_{\text{partially accurate feedback}}=0.499$ (0.296), $M_{\text{control}}=0.287$(0.124), Games–Howell corrected $t(38.06)=-3.54$, $p=.006$; where Holm–Bonferroni corrected $\alpha=0.008$. Figure 2.2.b shows the mean CV values across experimental conditions.

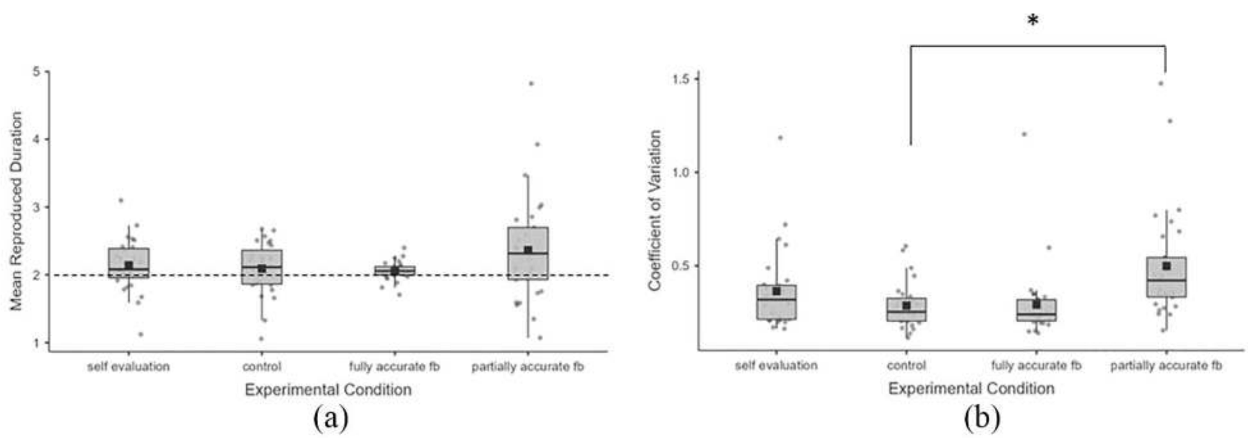


Figure 2.2. (a) The mean reproduced durations and (b) CV values across experimental conditions. The target duration is marked with a black dashed line. In both panels, error bars represent the standard error (SE) of the estimates, the grey dots depict the individual data points, and the black squares depict the distribution mean. The asterisk depicts statistical significance at $\alpha=0.05$ level.

2.5.2. Temporal error monitoring ability in the self-evaluation condition. To evaluate the temporal error monitoring ability in the self-evaluation condition in a single model, we conducted a linear mixed effects analysis. We included participants as random intercept and z-scored reproductions as fixed slope in the model, where the dependent variable was the recoded confidence ratings. The full model was

Recoded confidence ratings ~ timing error (z-score transformed reproduction) + (1 | participant)

where ~ stands for "predicted from" and 1|participants stands for "random intercept across participants." This model aims to investigate the linear relationship between the timing errors and the recoded confidence ratings, where a significant positive relationship between the two variables would point to temporal error monitoring.

Linear mixed effects model revealed a statistically significant effect of timing error on recoded confidence ratings (ß=0.63, *SE*=0.03, *p*<.001). These results point out that participants could monitor the direction and magnitude of their temporal errors on a trial-by-trial basis, which is in line with the previous work (e.g., Akdoğan & Balcı, 2017; Kononowicz et al., 2019; Öztel et al., 2020b). Figure 2.3. illustrates this linear relationship between timing errors and the recoded confidence judgements. Note that we ran the same model using the recoded confidence ratings as being linearly spaced between 1 (too short, low confidence) and 6 (too long, low confidence). The results and conclusions remain the same when data are organised in this fashion as well.

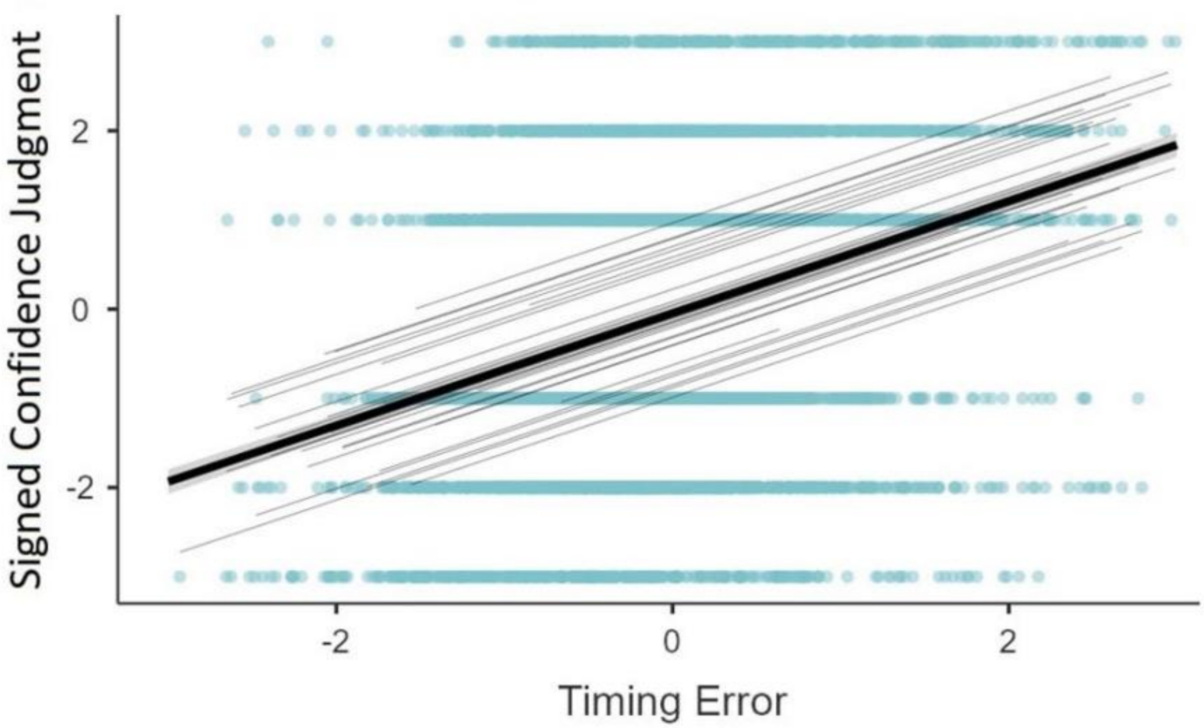


Figure 2.3. The relationship between timing errors (z-score transformed reproductions) and recoded confidence rating in self-evaluation condition.

Furthermore, we investigated if the participants' mere short-long judgement accuracies were above chance level to determine whether the trial-by-trial temporal error monitoring and global error monitoring performance are indeed distinct processes as suggested by Riemer et al. (2019). One sample t-test revealed that on average, participants performed higher than chance level, *t*(26)=5.6313, *p*<.001, *M*=0.61(0.10), *Median*=0.6, demonstrating the directional aspect of the temporal error monitoring ability in an isolated fashion. Judgement accuracy for overall timing biases. We first categorised the accuracy of global judgements in a way that captures the match between the overall bias judgement and the actual sign of the bias in mean reproduced duration compared with the target duration. For this, the global bias judgement was categorised as "accurate" if it matched the sign of the mean reproduced duration and "inaccurate" in case of a mismatch. We first investigated whether participants in all conditions could make correct global judgements regarding their mean reproduced durations (pooled across conditions). The proportion test revealed that participants' accuracy in global judgements was at a chance level ($proportion_{correct}$=0.46, *p*=.8). We further investigated the proportion correct for each condition separately. In none of the experimental conditions (except for the partially accurate feedback condition), the proportion of participants who correctly responded to the global error monitoring judgement question was significantly higher than the chance level (self evaluation: $proportion_{correct}$=0.30, one-tailed proportion test *p*=.99; control: $proportion_{correct}$=0.56, *p*=.35; fully accurate feedback: $proportion_{correct}$=0.63, *p*=.12; partially accurate feedback: $proportion_{correct}$=0.38, *p*=.9; where alternative hypothesis is proportion>0.5)

2.6. Discussion

Recent work has consistently supported the idea that human online error monitoring ability extends to metric domains that is reflected in a correct trial-by-trial match between the directional error magnitude judgements and actual signed deviations from a target. Critically, this online error monitoring ability does not require external feedback. One study, however, found that monitoring of the average timing bias required explicit trial-by-trial directional feedback (Riemer et al., 2019). Based on this finding, Riemer et al. (2019) claimed that feedback provides information above and beyond internal self-evaluation for monitoring average timing biases. The current study examined the effect of different types of feedback on global temporal error monitoring judgements with more conditions than those used in Riemer et al. (2019). Based on the predictions of the model introduced by Akdoğan and Balcı (2017), we hypothesized that participants would be able to monitor their temporal errors in an online fashion, but this ability would not translate to the global monitoring of timing bias directions in the absence of feedback.

None of the experimental conditions resulted in biased temporal processing as reflected in no difference between the time reproductions and target duration (after correction for multiple comparisons). This finding contradicts Droit-Volet and Izaute (2009)'s finding that demonstrates a rightward shift in bisection point (i.e., slowed down temporal processing) in children receiving feedback in the temporal bisection task. However, the timing precision was substantially lower in the partially accurate feedback condition compared with the control condition. The lack of the effect of consistent feedback on timing precision corroborates earlier work using temporal bisection with adult participants (Droit-Volet and Izaute, 2009), whereas accurate feedback reduces the variability in the timing performance (Bader & Wiener, 2021). Thus, the partially accurate feedback might have misled the internally driven feedback during the implicit temporal error monitoring, which is assumed to be common in all experimental conditions. The misleading effect of the partially accurate feedback might have resulted in the imprecise timing performance in this condition compared with the other conditions.

Consistent with earlier work, our results also demonstrated that participants in the self-evaluation condition could monitor the direction as well as the magnitude of their temporal errors on a trial-by-trial basis (e.g., Akdoğan & Balcı, 2017; Kononowicz et al., 2019; Yallak & Balcı, 2021; Öztel et al., 2020b). However, this trial-by-trial temporal error monitoring ability did not apply to the global monitoring of the temporal bias directions in any of the conditions. Together, these results suggest that the mechanisms that underlie accurate trial-based error monitoring do not inform the participants of their overall biases and neither does fully accurate feedback.

In the case of internal and external feedback, what seems to matter with respect to global error monitoring performance is the accuracy of trial-based error directionality judgements and interestingly, although not statistically significant, participants in the self-evaluation and partially accurate feedback conditions tended to perform worse than even the control condition. This finding shows that self-evaluation might serve as a feedback mechanism, which is detrimental to global error monitoring performance, presumably due to less than perfect accuracy of trial-based judgements. Future research is needed to address these interesting trends.

The current study does not rule out the possibility that the global error monitoring ability simply reflects a mere memory process without necessitating active error monitoring. This possibility applies also to previous global error monitoring studies (e.g., Riemer et al., 2019). On the contrary, one can argue that the trial-by-trial online error monitoring also constitutes such a memory process that is less prone to the interference/decay effects due to the short delay between performance and metacognitive judgement (Riemer et al., 2019), instead of relying on the cumulative record of trial-based performances as in the case of global error monitoring. For instance, one can reach above chance error monitoring performance when s/he relies on the comparison of the working memory representation of the already reproduced interval and a random sample from the reference memory (a possibility inspired by Gibbon et al., 1984). We, however, argue that any cognitive process that reflects on one's own decision/ behaviour can be regarded as a form of metacognitive processing (e.g., performance monitoring). Under this rationale, global and online error monitoring does not differ phenomenologically. Nevertheless, the current study, along with that of Riemer et al.'s (2019), suggests that these processes are distinct from one another. Future studies can also interrogate the perceived global biases in every trial and thus in a cumulative fashion. Such data would provide richer information about whether participants can keep track of global errors and the factors that affect the accuracy of such judgements (e.g., exponentially weighted history of directional feedback).

The primary explanation for robust trial-based temporal error monitoring but null overall temporal bias monitoring comes from the temporal error monitoring model described in Akdoğan and Balcı (2017), which predicts that the monitoring of the timing biases is not possible in the absence of external feedback. In line with this model's predictions, in a recent study, Öztel and Balcı (2020a) demonstrated that human adults were not able to keep track of the temporal biases that are induced via the physical properties of the to-be-timed stimuli: participants' confidence ratings of the temporal bisection performance matched the stimulus induced leftward shift in the bisection points.

The inability to keep track of average temporal biases might also stem from the fact that error directionality information fades away due to overloaded working memory (WM) capacity (i.e., 200 trials per condition). To this end, Coutinho et al. (2015) found that excessive load in WM negatively affects uncertainty monitoring performance (see also Maniscalco & Lau (2015)). But this account does not explain the findings of Riemer et al. (2019), who demonstrated better than chance error monitoring performance for average timing biases in the signed feedback condition (with 50 trials per condition).

The results obtained in the current study are in parallel with this finding for the inherent temporal biases that are internally induced. However, while contributing to the discussion regarding the possible distinctions in these two error monitoring processes, current findings contradict the results obtained by Riemer et al. (2019), which demonstrated a facilitating effect of signed feedback and by Brocas et al. (2018), which demonstrated inherent temporal bias monitoring ability. Although the obtained results in the current study seemingly contradict those obtained by Brocas et al. (2018), it should be noted that their experimental design captured only the temporal biases that are induced with a monetary payoff scheme without the involvement of any online influences. This kind of difference between the two studies may have very well led to the seemingly contradicting results.

One reason why our result seemingly contradicts that obtained by Riemer et al. (2019) might be because our feedback condition did not include any magnitude information. Thus, it may be that the improving effect of signed feedback observed by Riemer et al. (2019) occurred via its combination with the magnitude information. Moreover, the global error direction judgement in the current study was a 2AFC question where participants could not report the magnitude of their overall deviation. This may have also led to seemingly contradicting findings to those obtained by Riemer et al. (2019). Finally, it should be noted that receiving the global error monitoring as onetime only does not capture the cumulative change in the global error direction judgements throughout the experiment. Moreover, this kind of measure of global error monitoring might make the performance vulnerable to working memory overload when asked at the end of the experiment, which can also introduce noise in evaluation of the temporal errors globally. Although such memory interference could be reduced with explicit feedback in Riemer et al. (2019), this was not the case for the current study. The trial-by-trial error monitoring, on the contrary, is less subject to such interferences introduced by the working memory load, as the evaluation of the timing error is immediate (as discussed in Riemer et al., 2019).

Another limitation of the current study stems from the fact that the experiment was run online instead of a strictly controlled laboratory environment. Future studies are needed to elucidate the critical procedural differences that might have led to these differential findings.

2.7. Conclusion

Our study demonstrated that none of the trial-by-trial influences resulted in an improvement in monitoring of the overall timing biases. In addition, the fully consistent external feedback also did not have an improving effect on the global error monitoring accuracy, which seemingly contradicts the earlier results of Riemer et al. (2019). We therefore conclude that although it correctly matches the corresponding temporal error direction and magnitude, online self-evaluation does not serve as an internal feedback-like mechanism that would result in correct overall temporal bias judgements.

Chapter 3

TEMPORAL ERROR MONITORING DOES NOT DEPEND ON WORKING MEMORY

3.1. Abstract

Working memory (WM) and metacognition has been documented to be in a reciprocal relationship. This study aims to address if temporal error monitoring performance can be diminished with increased working memory load. We hypothesized that if temporal error monitoring has commonalities with perceptual error monitoring, temporal error monitoring performance should be diminished by increased working memory load. Participants completed a temporal error monitoring task in a dual task design in which the secondary task was a letter alphabetization task. Results revealed no disrupting effect of WM load on either confidence or short-long judgments as being different metrics of temporal error monitoring ability. These results demonstrate that unlike perceptual error monitoring, WM and temporal error monitoring have distinct processing mechanisms. With this result, the current study suggests that temporal and perceptual error monitoring may partially rely on different mechanisms. Results are discussed within A Theory of Magnitude (ATOM), pacemaker-accumulator model and temporal error monitoring frameworks.

3.2. Introduction

Metacognition refers to “cognition about cognition” (Flavell, 1979), and also is argued to be a part of consciousness (Fleming et al., 2012a; Lau, 2019; Lau & Rosental, 2011). Metacognition serves as a supporting mechanism for learning (e.g., Schraw et al., 2006) and behavioral adaptation (e.g., Brown, 1987). The error monitoring aspect of metacognition comes from the ability to report and reflect upon self-generated errors without necessitating external feedback (Yeung & Summerfield, 2014). Working memory (WM) is another high-order cognitive ability that helps us hold and manipulate a limited amount of information for a limited amount of time (Frith & Dolan, 1996). Both of these processes are limited in their capacity (De Loof et al., 2015; Gayet et al., 2016) and thus the interplay between metacognition and WM emerges as an important mechanistic question. The current study addressed this issue by investigating the effect of working memory load on temporal error monitoring performance.

The error monitoring ability is often operationalized with the match between objective performance and subjective confidence ratings. Previous work demonstrated that human participants can match their confidence ratings to their categorization accuracy in perceptual decision making (e.g., Fleming & Dolan, 2012b, 2012c; also see: Fleming et al., 2012a) and memory judgments (e.g., Vandenbroucke et al., 2014). In all of these domains, error monitoring is investigated in two alternative forced choice (2AFC) tasks where the decision accuracy is binary in nature (e.g., correct or incorrect). But many daily life situations involve errors that have metric characteristics: the error that is committed has a magnitude direction. For example, one can overcook or undercook a dish at different temperatures.

Akdoğan and Balcı (2017) investigated whether human participants can keep track of these aspects of their timing errors. Participants were tested in a temporal reproduction task where they were asked to reproduce a previously presented target duration as accurately as possible. After each trial, participants rated the accuracy of their temporal reproductions in terms of their proximity to the target duration. Lastly, participants classified their reproductions as shorter or longer than the target duration. Results demonstrated a clear trial-by-trial match between the signed confidence judgements and the signed temporal errors, suggesting a temporal error monitoring ability. This finding was replicated by subsequent studies (Kononowicz et al., 2019a; Yallak & Balcı, 2021; Oztel et al., 2021), and its scope has been expanded to include also numerosity (Duyan & Balcı, 2018, 2019) and line length judgments (Duyan & Balcı, 2020 - for a detailed discussion, see Yallak & Balcı, 2021).

Working memory capacity includes active manipulation of information available in the maintenance storage, which can help increase storage capacity (such as sorting, calculating, rotating, comparing - Rypma et al., 1999; 2002; Curtis & D'Esposito, 2003). These information manipulation strategies seem to share functional characteristics with the error monitoring component of metacognition, where error monitoring also involves the comparison of maintained information to the target information (Falkenstein et al., 1990, 2000; Gehring et al., 1993; Coles et al., 2001; see also: Ullsperger & Von Cramon, 2004) and acting upon it with various compensatory behaviors (e.g., post error correction, post-error slowing). For example, according to global workspace theory, metacognition/error monitoring occurs as a result of a comparison process: one maintains the representation of a target perception (i.e., that is objectively correct) as well as the executed behavior in WM (Shea & Frith, 2019; see also: Yoo et al., 2020) during a metacognitive activity. The executed behavior is then subject to a comparison with the target representation to detect the match (or a possible mismatch) between them. Finally, one is to act upon the result of this comparison and with a behavioral reflection (such as confidence) and/or compensation (like slowing down after an error - post-error slowing: e.g., Rabbitt, 1966, and hypercorrections: e.g., Butterfield & Metcalfe, 2001; Metcalfe & Miele, 2014; for a detailed discussion: Vidal et al., 2020). Thus, metacognitive processes also involve manipulation of maintained information in a form of comparison with target and actual behavior (global workspace theory: Shea & Frith, 2019; mismatch theory: Falkenstein et al., 1990, 2000; Gehring et al., 1993, Coles et al., 2001; see also: Ullsperger & Von Cramon, 2004) along with corresponding reflection. The global workspace theory, thus, brings about the idea that WM can act as a base resource for metacognition. The above-mentioned overlaps at the algorithmic level are reinforced by overlaps also at the implementation level (Marr, 1982).

To this end, many studies demonstrated commonalities between metacognition and working memory at the neural level, including dorsal, and rostral parts of the prefrontal cortex (PFC). Specifically, functional magnetic resonance imaging (fMRI) studies provide support for the enhanced activity in rostrolateral and left dorsolateral PFC (Fleming et al., 2012c). In support of these findings, Rounis and colleagues (2010) demonstrated the causal role of bilateral dlPFC to perceptual metacognition by utilizing transcranial magnetic stimulation (TMS). Together, the critical involvement of dlPFC has been documented.

Regarding the neural basis of WM, the active manipulation and monitoring of information is underlain by the dorsolateral PFC (e.g., Bor et al., 2003; Rypma et al., 2002; Curtis & D'Eposito, 2003; Rowe et al., 2000, 2001; Postle et al., 1999; dual-task paradigm: D'esposito et al., 1995; Cohen et al., 1997; but see: Davis et al., 2018). For instance, one direct evidence showed that disruption of the activity of dlPFC by TMS diminished performance in the alphabetization of the given letter (Postle et al., 2006; but see: Wang et al., 2018). Taken together these results demonstrate causal support for the role of dLPFC on information manipulation in the WM, as in the case of metacognitive processing, raising the commonality of both cognitive mechanisms at the neural level.

One behavioral manifestation of the neural commonalities between WM and metacognition was demonstrated by Maniscalco and Lau (2015) in perceptual error monitoring domain. In their study, Maniscalco and Lau (2015) tested participants in a 2AFC task where participants were required to classify which side of the screen the target stimulus was presented and rate their confidence regarding their classification accuracy. At the beginning of each trial, participants saw a letter array consisting of either one or four letters. Participants were asked to mentally alphabetize the letters, where they were asked to indicate whether a given letter was assigned a correct alphabetical order at the end of the trial. Results demonstrated that although a load of passive maintenance of items (i.e., number of letters to be maintained in the WM) did not affect the metacognitive performance, the increased difficulty of alphabetization (i.e., manipulation difficulty) reduced the metacognitive performance (Maniscalco & Lau, 2015; see also: De Loof et al., 2013, 2015). Taken together, these results demonstrate how different types of WM load can affect metacognitive accuracy differently and emphasize the possible shared cognitive mechanisms between WM and metacognition. These behavioral results also align with the neuropsychological data that point to the involvement of dLPFC in both WM manipulation (e.g., Bor et al., 2003) and error monitoring processes (e.g., Shekhar & Rahnev, 2018).

Given the recent finding demonstrating the differentiation between the error monitoring performance in perceptual decision-making and temporal error monitoring (Yallak & Balcı, 2021), the current study aims at testing whether working memory overload affects temporal error monitoring performance. While there is no causal role of a specific brain region documented for this ability as in the case of perceptual error monitoring yet, a recent electroencephalogram (EEG) study reported temporal error monitoring-related activity of precuneus (Kononowicz et al., 2019a; 2019b), which is also implicated in WM during a delayed matching to sample task (Luber et al., 2007). Taken together, these results point to a potential overlap between temporal error monitoring and WM. In light of these findings, we hypothesized that if WM shares common mechanisms (Maniscalco & Lau, 2015) such that it acts as a base resource for temporal error monitoring (Shea & Frith, 2019), the temporal error monitoring performance should be disrupted as a result of increased WM load. If, on the other hand, the temporal error monitoring and the WM have distinct mechanisms such that representations that are to be acted upon are not stored in WM (as discussed in global workspace theory, Shea & Frith, 2019), the temporal error monitoring performance should remain insensitive to the increases in the WM load. We tested our hypotheses in active manipulation and passive maintenance of information which are two different WM indices.

3.3. Method

3.3.1. Participants. 95 Koç University undergraduate students completed this online experiment in return for extra course credit. Data from 47 individuals were discarded due to not completing the experiment (n = 18) and unreasonable reproduction distributions (i.e., too short reproduction trends around 2100 milliseconds 700 milliseconds that hints the aim to complete the experiment earlier than it would normally take, or too long reproduction trends around 3300 milliseconds 600 milliseconds that hints the participant may have left the experimental device, or reproduction trends that hint explicit counting; n = 29). Forty-eight participants were included for the formal analysis (32 females, 44 right handed, M_{age} = 21.58±4.45). Since we collected data in an online environment, we decided on the number of participants as the double of Maniscalco and Lau's (2015) experiment. All the exclusions were decided on before the data analysis was commenced.

3.3.2. Stimuli and Apparatus. The target duration of 2500 milliseconds was presented with a 6.1 x 6.1 red square placed on the center of the screen. The reproduction stimulus was the same as the target duration stimulus, except that the color of the square was blue. We used all the digits from one to nine for the WM task in Experiment 1. The experiment was prepared in PsychoPy Builder (Peirce et al., 2019). All participants completed the experiment on their personal computers. All the stimulus presentations and response recordings were controlled via the Pavlovia online experiment platform (Bridges et al., 2020).

3.3.3. Procedure

We used letter alphabetization as a WM task (also see Maniscalco & Lau, 2015). The task started with a presentation of the target duration, which was followed by the presentation of a letter array consisting of either one (low load condition) or four (high load condition) letters randomly chosen from F, G, H, J, K, L, M, N, P, R, S, T (Maniscalco & Lau, 2015; Q was discarded from the list as it is not part of the Turkish alphabet). The letter length condition was randomly chosen and presented with equal probability. The letter array was presented in the center of the screen for 2 seconds and followed by a fixation cross for 500 milliseconds. Participants were instructed to remember the letter array and mentally alphabetize them. In order to prevent participants from rehearsing the letter array, the blue square (timing stimulus) appeared immediately after the fixation cross and participants were instructed to terminate the reproduction by pressing the spacebar. The participants' aim was to reproduce the target duration as accurately as possible. After each reproduction, participants reported their confidence and the direction of deviation from the target (short-long judgement). Finally, participants were presented with a probe letter along with a digit from one to four to depict the location of the probe letter in the letter array in that trial. Participants pressed 1 (no) if they thought the digit did not match the correct alphabetical order of the probe letter and 2 (yes) otherwise. The correct response was "no" in half of the trials. Participants received auditory feedback regarding the accuracy of their WM responses (high-pitched beep sound for correct and low-pitched beep sound for incorrect responses) for 200 milliseconds along with a fixation cross. The target duration was reminded to the participants every six trials. Participants had a 1-minute mandatory break in every 50 trials. The task consisted of 400 trials in total and took approximately 90 minutes to complete. Figure 3.1. illustrates the experimental procedure.

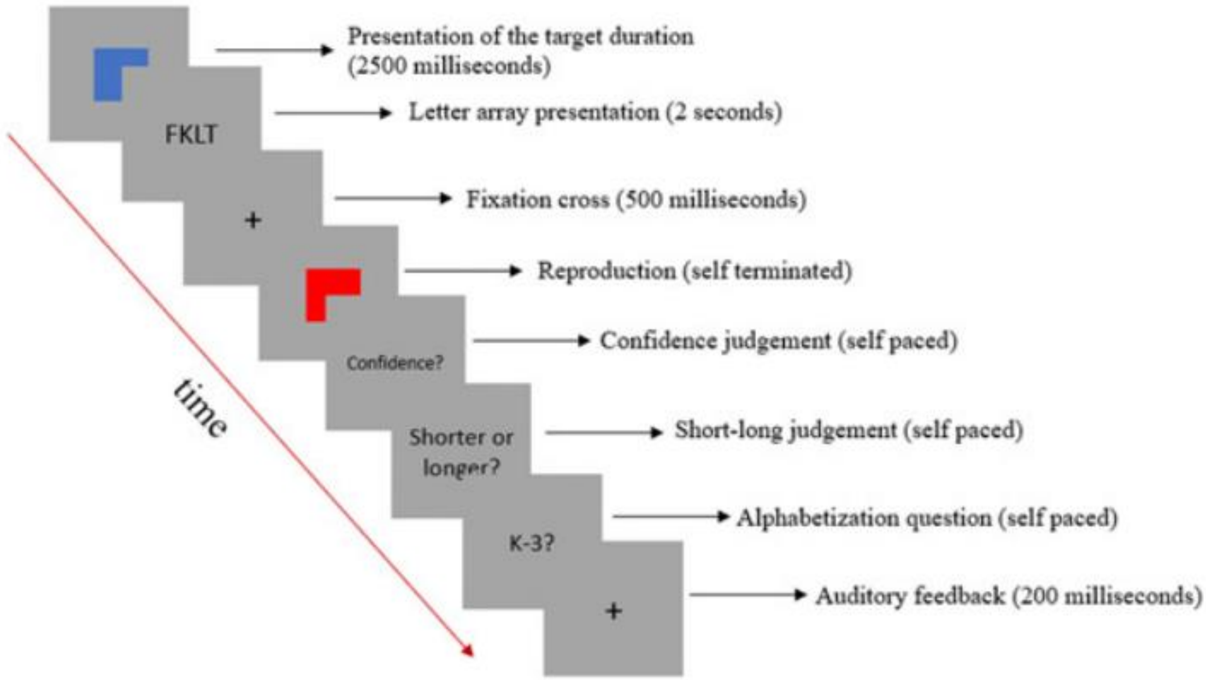


Figure 3.1. Schematic representation of one trial in the experiment

3.4. Analytical Approach

3.4.1. Data Filtering

We first discarded all reproductions that were shorter than half of the target duration (1750 milliseconds) and longer than twice the target duration (5000 milliseconds). We then z-score transformed the remaining reproduced durations (i.e., (temporal reproduction in the nth trial - mean of the temporal reproductions throughout the experiment)/ standard deviation of the temporal reproductions throughout the experiment, will be hereafter referred to as "signed timing error"). Before performing any statistical analyses, we applied listwise deletion to all trials that are above 3 standard deviations in reproduced duration as these trials could artificially confirm our hypothesis regarding the temporal error monitoring ability (e.g., participants could reproduce very long or very short durations on purpose to correctly match the subsequent confidence and short-long judgements about these extraordinarily short-long reproductions). Based on this filtering method, 13.7% of the whole data is discarded from the dataset for the formal analyses in total.

3.4.2. Definitions of Two Different Working Memory Indices: Sorting Difficulty Determination and Number of Letters

The working memory load was quantified with two different indices. The first one was the mere number of the letters that the letter array contained (either one or four, will be hereafter referred to as “letter number conditions”). This metric aimed to test the effect of mere working memory maintenance demand on temporal error monitoring performance. The second metric was the alphabetization/sorting difficulty of the given letter array. This metric aimed to test the effect of working memory manipulation, in addition to the maintenance demand on the temporal error monitoring performance.

The sorting difficulty level was determined only for the four letters condition where an alphabetization took place. Before the formal analyses, the letter array was divided into three consecutive substrings so that letter stimulus was paired with the subsequent letter stimulus. After this division, the letter pairs were checked to see if they were alphabetized. For each pair of substrings that are not alphabetized, the alphabetization difficulty was increased by one difficulty point. As a result of this, four difficulty points were established, from very easy (0, which depicted a letter string stimulus that was fully alphabetized) and very hard (3, which depicted a letter string stimulus that was fully non-alphabetized), which was then divided into “easy” and “hard” sorting difficulty conditions. For example, given a letter string stimulus of “TKFL”, the sorting difficulty would be determined as follows: the first pair of substrings “TK” is not alphabetized (“K” comes earlier than “T”), which would increase the sorting difficulty by one point. The second pair of substrings “KF” is again not alphabetized, which would again increase the sorting difficulty by one point. The final pair of substrings “FL” is alphabetized, which would not result in any increase in the sorting difficulty level. In total, the sorting difficulty point of the given letter array can be determined as 2 (hard sorting difficulty).

3.5. Results

3.5.1. Manipulation Check: **WM Accuracy and Reaction Times.** We first compared the mean WM accuracy levels across the letter number conditions to test if the four-letter condition indeed induced a higher WM load than the one-letter condition. A paired sample t-test revealed that participants were more accurately responded the WM question in the one letter condition compared to the four letters condition ($t(47) = 8.01$, $p < .001$, Cohen's d = 1.16; $M_{\text{one letter}} = 0.79$ (0.18); $M_{\text{four letters}} = 0.57$ (0.13), note that accuracy in both one letter and four letters condition was statistically higher than chance level (one letter: $t(47) = 11.28$, $p < .001$, Cohen's d= 1.63; four letters: t (47) = 3.91, $p = .0003$, Cohen's d = 0.56).

We also compared the mean WM accuracies across the levels of WM sorting difficulties. A paired sample t-test revealed a significant difference between high load-easy and high load-hard conditions in terms of WM accuracy (t (47) = 2.30, $p = .03$, Cohen's d = 0.33; $M_{\text{high load-easy}} = 0.59$ (0.14), $M_{\text{high load-hard}} = 0.56$ (0.13)). This result demonstrates that the WM load manipulation was effective.

We also investigated whether the effect of WM load manipulation on accuracy also translated to the letter sorting reaction times. A paired sample t-test revealed a significant difference between the letter number conditions (t (47) = 2.94, $p = .005$) in terms of mean letter sorting reaction times (will be referred to as WM RTs hereafter), demonstrating a higher median RT for the four-letter condition ($M_{\text{one letter}} = 1.54$ (0.71); $M_{\text{four letters}} = 2.27$ (1.95)). Finally, we compared the WM RTs across the sorting difficulty levels. However, this result did not apply to mean WM RTs ($p = .36$).

3.5.2. Time Reproductions, Coefficients of Variations and Confidence Ratings

We further investigated the biasing effect of WM load on time reproductions. A paired sample t-test revealed a significant difference across the letter number conditions (i.e., either one or 4 letter conditions; t (47) = 2.0549, $p = 0.045$, Cohen's d = 0.3; $M_{\text{four letters}} = 2.73$ (0.3), $M_{\text{one letter}} = 2.69$ (0.31)) but no significant difference between sorting difficulty conditions for the high load condition (i.e., four letters, t (47) = 0.29, $p = 0.77$, Cohen's d = 0.04). This result demonstrates the biasing effect of the increased number of letters on the reproduced durations.

We also compared the coefficient of variations (CV) of the reproductions across different manipulation conditions. We calculated CV by dividing the standard deviations of reproductions by their means as in the following equation: $CV_{\text{reproduction}} = SD_{\text{reproduction}}/M_{\text{reproduction}}$.

This calculation is widely used as an index of temporal precision where smaller values indicate higher temporal precision (e.g., Lewis & Miall, 2009). There was no significant difference in reproduction CVs between different conditions (letter number conditions: t (47) = 0.97, p = .34, Cohen's d = 0.14; sorting difficulty conditions: t (47) = 0.57, p = .57, Cohen's d = 0.08), demonstrating that the different indices of WM did not lead to a change in temporal precision. This result also applied to confidence ratings (letter number conditions: t (47) = 1.87, p = .07, Cohen's d = 0.27; sorting difficulty conditions: t (47) = 0.51, p = .61, Cohen's d = 0.07). Descriptive statistics for reproductions, CVs and confidence ratings are provided in the Supplementary Online Material (SOM).

3.5.3. Hypothesis testing for the effect of WM load on temporal error monitoring. We investigated the effect of different WM load on confidence and short-long judgements in an isolated fashion. To investigate the effect of WM load on temporal error monitoring ability with three different load indices (i.e., letter number, sorting difficulty conditions and WM accuracy) in a single model, we ran three different linear mixed model analyses as depicted below:

Model 1. Confidence judgement ~ absolute timing errors (absolute of z-score transformed reproductions) * number of letters +1|participant where in all models, "1|participant" reads as "random intercept across participants"

All models included all lower terms. Additionally, (absolute) timing errors were level 1 and the number of letters was level 2 variables in all models. We performed the hypothesis testing analyses for both the complete (N = 77) and individual based filtered data (N = 48). For the individual based filtered data, linear mixed effects revealed a significant main effect of absolute timing errors (ß = 0.103, SE = 0.008, p < .001, R^2 conditional = 0.36) and number of letter conditions ($M_{\text{one letter - four letters}}$ = 0.046, SE = 0.015, p = .002) on isolated confidence judgements. However, the interaction between the two independent variables was not statistically significant (p = .13). Similar results were obtained for the complete dataset (ß = 0.10, SE = 0.007, R^2 conditional= 0.37, p < .001; $M_{\text{correct-incorrect}}$ = 0.053, SE = 0.013, p < .001; the interaction term p = .15). These results demonstrate a biasing effect of WM load such that the one letter condition followed a higher confidence judgement than the four letter condition, but this effect did not reflect itself on the diminished temporal error monitoring performance as indexed with the match between timing errors and confidence judgements. The first row of the left panel in Figure 3.2. depicts the relationship between the absolute timing error and confidence judgement across the levels of number of letter condition. To investigate the effect of WM accuracy on the confidence judgements, we ran Model 2:

Model 2. Confidence judgement ~ absolute timing errors * WM accuracy +1| participant.

For the individual based filtered data, a similar pattern was observed for the effect of WM accuracy on the confidence judgements in isolation. Overall, participants could correctly match their subjective confidence judgements to the magnitude of their temporal errors (ß = 0.102, SE = 0.008, $p < .001$, R^2 conditional = 0.36), demonstrating a temporal error monitoring ability regardless of the WM accuracy. Moreover, participants were more confident in their temporal reproductions when accurate on the letter sorting task ($M_{\text{correct-incorrect}} = 0.03$, SE = 0.016, $p = .045$). However, this effect did not apply to the interaction between the absolute timing errors and WM accuracy (p = .75), revealing no modulating effect of WM accuracy on the temporal error monitoring performance as indexed with the confidence judgements in isolation. A similar pattern was observed for the complete dataset (ß = 0.103, SE = 0.007, $p < .001$; $M_{\text{correct-incorrect}} = 0.03$, SE = 0.014, R^2 conditional = 0.37, $p = .04$; interaction term $p = .98$). Second row of the left panel in Figure 3.2. depicts the relationship between the absolute timing error and confidence judgement across the levels of WM accuracy.

Finally, we investigated the effect of sorting difficulty on confidence judgements with the following model:

Model 3. Confidence judgement ~ absolute timing errors * sorting difficulty +1| participant.

Note that in this model, we discarded all 1-letter trials as no mental alphabetization was made in those trials. Such a discard results in a sorting difficulty with two levels (high load-easy and high load-hard). For the individual-based filtered data, although an overall temporal error monitoring ability was observed (ß = 0.093, SE = 0.011, $p < .001$, R^2 conditional = 0.37), Model 3 revealed no biasing effect or interaction effect for the WM load indexed with the sorting difficulty (all ps > 0.05). The same pattern was observed for the complete data (ß = 0.096, SE = 0.009, $p < .001$, R^2 conditional = 0.38; all other ps > 0.05). The third row of the left panel in Figure 3.2. depicts the relationship between the absolute timing error and confidence judgement across the levels of sorting difficulty. We ran the logistic versions of the models depicted above for the short-long judgements. Except instead of the absolute values, we used the signed timing errors to keep the directionality information for the short-long classifications.

For the individual-based filtered data, a similar scenario was observed in short-long judgements. For Model 1, logistic mixed effects model revealed only the main effect of signed temporal errors (ß = 0.45, exp(ß) = 1.57, *SE* = 0.019, $p < .001$, R^2 conditional = 0.34) on short-long judgements, revealing a higher probability of "long" response as the reproduced durations prolonged with respect to the mean reproduced durations. The main effect of letter conditions and the interaction between the signed temporal errors and letter conditions were not statistically significant ($p > .05$). A similar pattern was observed for the complete dataset (ß = 0.09, *SE* = 0.003, $p < .001$, R^2 conditional = 0.24; all other *p*s > 0.05). The first row of the right panel in Figure 3.2. depicts the relationship between the signed timing error and short-long judgement across the levels of the number of letter condition.

For both WM accuracy (Model 2) and letter sorting (Model 3) models with the individual-based filtered data, only the main effect of the signed temporal error was significant (Model 2: ß = 0.46, exp (ß) = 1.58, *SE* = 0.02, $p < .001$, R^2 conditional = 0.34; Model 3: ß = 0.43, exp (ß) = 1.53; *SE* = 0.027, $p < .001$, R^2 conditional = 0.36; all other *p*s > 0.05). A similar effect was observed in the complete data as well (Model 2: ß = 0.085, *SE* = 0.003, $p < .001$, R^2 conditional = 0.24; Model 3: ß = 0.08, *SE* = 0.004, $p < .001$, R^2 conditional = 0.24, all other *p*s > 0.05). The second and third rows of the right panel in Figure 3.2. illustrate the relationship between the signed timing error and short-long judgement across the levels of WM accuracy and sorting difficulty, respectively.

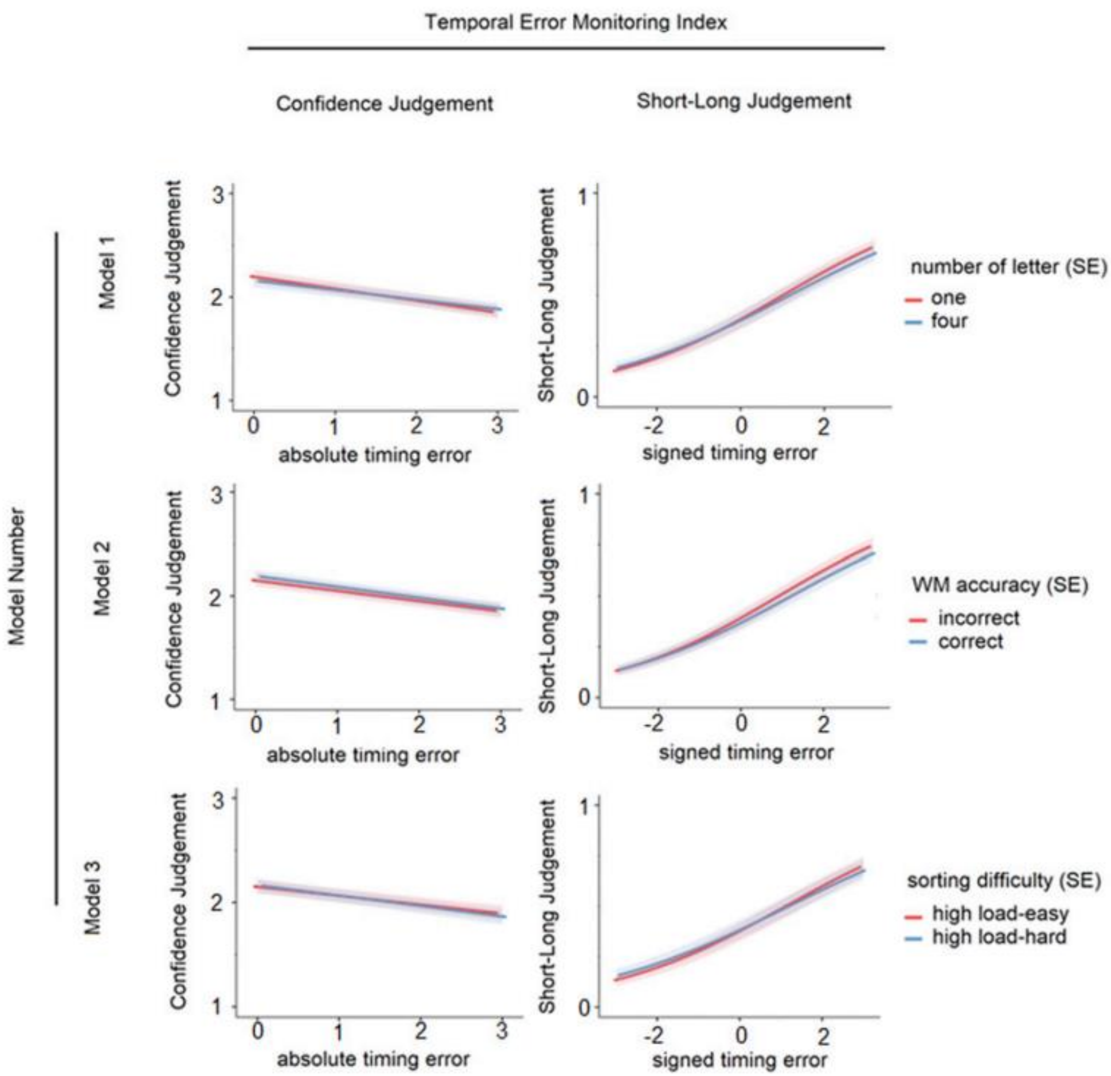


Figure 3.2. The relationship between timing errors (absolute and signed) and temporal error monitoring indices (confidence and short-long judgements) across the levels of different types of WM load (N = 48). Each row depicts the different WM load type and each column depicts different temporal error monitoring indices. Shaded areas depict standard error of the estimates (SE).

3.6. Discussion

This study investigated the effect of working memory load on temporal error monitoring performance. If temporal error monitoring ability has common domain characteristics with perceptual error monitoring, we hypothesized a diminished temporal error monitoring ability for a high WM load condition as in the case of Maniscalco and Lau's (2015) study. If, on the other hand, these two metacognitive abilities demonstrate domain-specific characteristics, we predict the metric error monitoring performance might be insensitive to the different WM load conditions.

Our manipulation check analyses revealed that the accuracy in the low load condition, as indexed with the number of letters (i.e., low WM maintenance load) was higher compared to the high load condition. While pointing to an effective manipulation of the WM load, the exact source of this effect cannot be disentangled as both the number of letters and the sorting difficulty vary at the same time in such a comparison. As a result, it is not known whether the effect of manipulation here stemmed from (a) the mere number of letters, (b) the sorting difficulty, or (c) the additive effect of the two. Nevertheless, the comparison across the sorting difficulty levels in four-letter conditions revealed that the WM manipulation was effective even when the number of letters was controlled for.

Although the difference in the accuracy levels across the number of letter conditions was observed in the corresponding RTs, such a difference was not observed in the sorting difficulty RTs. This result is in line with Maniscalco and Lau (2015). While the observed difference in the accuracy across the difficulty levels might represent the interference in the memory trace with the increased difficulty, the RTs might instead represent the search in the memory traces. Thus, as the given task does not require a differential search in the memory traces across difficulty levels, comparable RTs might be expected.

We found a biasing effect of an increased number of letters on the reproduced durations. This finding is in line with A Theory of Magnitude (Walsh, 2003) that predicts all sensory information with metric characteristics (time, space and number) are processed by a shared common mechanism which results in a more A-more B-like perceptual pattern. For example, in line with this prediction, Oliveri et al. (2008) observed that numerical magnitudes altered time as a linear function of the numerical value resulted in temporal overestimation in a way that demonstrates a “more A-more B” like pattern, as predicted by ATOM (Bueti & Walsh, 2009; Walsh, 2003). Similar to Oliveri et al. (2008), we found that reproduced durations were significantly longer in the four-letter condition than in the one-letter condition, pointing to a more A (i.e., number of letters in the current study)-more B (i.e., longer reproduced durations), which supports ATOM’s prediction.

The temporal disruption manifested as biased temporal reproductions as a result of WM load is also in line with pacemaker-accumulator model (Church, 1984; Gibbon, 1977; Treisman, 1963). According to the Scalar Timing Theory, temporal information is represented as pulses generated by a pacemaker. These pulses are transferred to an accumulator that sends the accumulated pulses to the WM. The number of pulses transferred to the WM is then compared with the reference memory. Overloading WM resources can, thus, result in temporal distortions.

The linear mixed-effects models that aim to investigate the effect of WM load on temporal error monitoring ability replicated the previous results highlight the temporal error monitoring ability of human participants (e.g., Akdoğan & Balcı, 2017; Kononowicz et al., 2019a). Furthermore, we could not find the diminishing effect of WM load on temporal error monitoring for either indices (i.e., confidence and short-long judgements), which contradicts previous findings (Maniscalco & Lau, 2015). This result points to the dissociation between the perceptual and metric error monitoring domains which is in line with the previous findings (Yallak & Balcı, 2021). Furthermore, these results contradict global workspace theory, which argues that the target and behavioral representations to be compared for metacognition to emerge are held in WM. This contradiction also emphasizes the distinction between metacognition in perceptual and metric domains.

The parallel results that are obtained for the two different temporal error monitoring indices (i.e., confidence and short-long judgements) are in line with Akdogan and ˘ Balcı's (2017) model for temporal error monitoring. According to this model, temporal error monitoring occurs as a result of a comparison of the threshold crossing time of two internal timers. These two timers simultaneously integrate temporal evidence for (a) time estimation and (b) motor action. The increased amount of discrepancy between the two internal timers manifests itself as a decreased confidence judgement regarding the temporal decision. The direction of the time that it takes for the motor action to occur with respect to the temporal estimation manifests itself as a short-long judgement such that if the motor action timer hits the threshold earlier than the estimation timer, the reproduction will be classified as "shorter" than the target duration. The two clocks that process temporal information simultaneously are thus seemingly insensitive to an exogenous WM load to the same extent, which is in parallel with the temporal error monitoring model predictions of Akdogan and Balcı (2017).

Although in parallel with the temporal error monitoring model predictions, this result contradicts a recent finding from our laboratory that points to distinct sensitivities for confidence and short-long judgements to the endogenous motor information that is discussed to guide error monitoring ability (e.g., Palser et al., 2018, Öztel and Balcı, submitted): In a series of experiments, we recently demonstrated that confidence judgements are mainly guided by the agency of the error (i.e., correctly reflects the error amount only that belongs to the agent that commits the error), whereas owning the error introduces noise to the short-long judgements by disrupting the correct classification of the error direction to some extent compared to the errors that are committed by others. This result indicates that confidence and short-long judgements are sensitive to the endogenous influences to a different extent. On the other hand, the current study demonstrates that neither index of temporal error monitoring is sensitive to external influence (as induced by the WM load manipulation).

Taking the possibility that temporal and perceptual error monitoring have different neural mechanisms into account, temporal error monitoring might also be sensitive to a specific domain of the working memory load. Hence, the type of task that we used to induce a working memory load (mental alphabetization) might have fallen short of affecting both temporal error monitoring indices (i.e., confidence and short-long judgements) simultaneously. An example of such a working memory load can be mental arithmetics, which is demonstrated to be correlated with IPS and dLPFC activity (Belkacem et al., 2020). This possibility can be investigated in future studies.

This study has several limitations. The first limitation is that the study was conducted during the pandemic and in an online environment, thus it may not be directly compared to the obtained results by Maniscalco and Lau's (2015) experiment. Further, although earlier work demonstrates the diminishing effect of increased WM load on perceptual metacognition (Maniscalco & Lau, 2015), a replication of this diminishing effect on perceptual metacognition needs to be established in an online environment for a more valid and direct comparison between the two studies. In addition to the possibility that perceptual and metric error monitoring have different processing mechanisms (Yallak & Balcı, 2021), another reason why the current study found no effect of WM load on temporal error monitoring can be the difficulty of the dual task paradigm: It might be that the mental alphabetizing task was too difficult so that participants gave up on it and focused completely on the temporal error monitoring task. This might have resulted in dual task turning into a single task and thereby hindering the intended cognitive load effect of mental alphabetizing on temporal error monitoring.

This interpretation of the current data also aligns with the attentional gate model of interval timing (Zakay & Block, 1995) reflected in the prolonged temporal reproductions in high WM load conditions. The attentional gate model introduces a dynamic attention component to the classical pacemaker-accumulator model. This component serves as a gate between the pacemaker and the accumulator components by transferring the number of pulses emitted by the pacemaker as a function of attention paid to the to-be-timed duration. The model predicts that if the attention is diverted from the to-be-timed duration, it would be perceived as shorter than it actually is because some of the emitted pulses get lost due to attention diversion. Oppositely, if the attention is fully paid to the to-be-timed duration, it would be perceived as longer because none of the pulses would be lost. In the current study, we observed prolonged reproductions as a result of WM load increase, which can be explained with attention paid less to the to-be-timed duration during testing with increased dual task difficulty.

We also acknowledge that the percentage of data that did not satisfy the predetermined inclusion criteria was relatively high (49% of the whole dataset). It should be noted that the current study was conducted during the COVID-19 pandemic in an online environment. The high rejection rate of participants is somewhat typical for online studies. For example, Burns et al. (2014) could use only 47.67% of the data collected from the MTurk platform. 40% of the unused data were discarded due to inattentiveness. Previous studies reported a high careless response rate (up to 45.9% Brühlmann et al., 2020) and dropout rates (up to 69%, Sauter et al., 2020; Gagne &´ Franzen, 2021) for online experiments that might be due to nonpresence of the experimenter (Al-Salom & Miller, 2019; Hardre et al., 2012), lack of motivation, distraction and technical limitations (e.g., problems with internet connections or hardware related issues that cannot be standardized; Gagne & Franzen, 2021 ´). These factors might have led to the relatively high rate of data discard due to dropout/bad quality data in the current study. In any case, the fact that similar results were gathered without data exclusion shows that our exclusion criteria did not bias the results. Moreover, we could not perform a power analysis as we used linear mixed effects for the formal analyses, which requires post hoc power analysis from an obtained effect. Thus, the two times larger sample size than Maniscalco and Lau (2015) could be problematic without formal power analysis. Note that we obtained the null effect of WM indices on temporal error monitoring even with a rather large sample size, which seemingly points to no effect of WM load on temporal error monitoring.

3.7. Conclusion

This experiment aimed to test if the observed detrimental effect of WM load on perceptual metacognition also applies to the metric error monitoring performance in the temporal domain. Results demonstrated that WM load did not affect the temporal error monitoring indices. Although these results seemingly contradict those obtained by Maniscalco and Lau (2015), they are predicted by the findings of Yallak and Balcı's (2021) study, which demonstrated a differentiation between perceptual metacognition and metric error monitoring abilities. While this study is in need of replicating Maniscalco and Lau's (2015) study in an online environment, its preliminary results suggest that the WM load has differential effects on temporal error monitoring ability and thus has a distinct processing mechanism with perceptual metacognition. This possibly distinct characteristic may lend itself to neural processing that is unshared with perceptual metacognition, which should be addressed by future studies.

Chapter 4

TEMPORAL ERROR MONITORING: DOES AGENCY MATTER?

4.1. Abstract

Error monitoring is the ability to report one's errors without relying on feedback. Although error monitoring is investigated mostly with choice tasks, recent studies discovered that participants parametrically keep track of also the magnitude and direction of their temporal, spatial, and numerical judgment errors. We investigated whether temporal error monitoring relies on internal generative processes that lead to the to-be-judged first-order timing performance. We hypothesized that if the endogenous processes underlie temporal error monitoring, one can monitor timing errors in emitted but not observed timing behaviors. We conducted six experiments to test this hypothesis. The first two experiments showed that confidence ratings were negatively related to error magnitude only in emitted behaviors, but error directionality judgments of observed behaviors were more precise. Experiment 3 replicated these effects even after controlling for the

motor aspects of responding. The last three experiments demonstrated that belief of agency (i.e., believing that the error belongs to the self or someone else) was critical in accounting for the confidence rating effects observed in the first two experiments. The precision of error directionality judgments was higher when participants did not have agency on the observed behavior. These results show that confidence is sensitive to belief, and short-long judgment is sensitive to the actual agency of timing behavior (i.e., whether the behavior was emitted by the self or someone else).

4.2. Introduction

"Cognition about cognition" is arguably the simplest definition of metacognition (Flavell, 1979), which plays a role in learning (e.g., Schraw, Crippen, & Hartley, 2006) and behavioral adaptation (e.g., Brown, 1987). The error monitoring component of metacognition involves self-evaluating one's errors (e.g., reporting low confidence after errors; Yeung & Summerfield, 2012). What guides the correct matching of confidence ratings to the actual performance is unknown. We investigated whether self-monitoring of error magnitude and direction (i.e., metric error monitoring) relies on the endogenous generative processes that lead to the first-order timing performance or on the exogenous factors that are merely observational cues. We further investigated how the belief in the agency of observed behavior mediated this relationship.

The error monitoring component of metacognition is often operationalized via the match between the objective performance (i.e., first-order performance accuracy; Fleming & Lau, 2014) and the subsequent confidence rating (second-order judgment) regarding performance accuracy. Within this framework, previous research exhibited correct (i.e., better than chance) monitoring of perceptual (e.g., Fleming, Huijgen, & Dolan, 2012) and memory judgment (e.g., Flavell & Wellman, 1975) errors in individual trials. Regarding the information-processing basis of error monitoring, one view asserts that the motor actions carry critical importance for accurate metacognitive processing (for review see: Anzulewicz et al., 2019). For example, Siedlecka et al. (2016) found lower metacognitive accuracy when confidence judgments were made before compared to after making a decision. The speed of motor actions also predicts confidence judgments about perceptual decisions (Palser et al., 2018); faster motor actions are followed by high confidence. This finding corroborates the negative relation between confidence ratings and response times (Rahnev et al., 2020). Finally, Gajdos and colleagues (2019) found that subthreshold motor activity lead to overconfidence bias while maintaining metacognitive accuracy intact (see also Charles et al., 2020; Filevich et al., 2020). Together, these findings support the contribution of the response-related signals to the metacognitive judgments about memory and perceptual decision-making performance.

Many of the studies mentioned above were conducted using two-alternative forced choice (2AFC) tasks where the to-be-evaluated performance is either correct or erroneous (binary nature of accuracy). However, errors in real-life actions often have metric attributes (e.g., magnitude and direction). Such quantitative characteristics of the errors are especially critical when achieving a "goal" necessitates the magnitude estimation (e.g., judging whether it is safe to cross the road depending on the distance and speed of an approaching vehicle). Relatedly, a recent line of research demonstrates that participants can correctly evaluate the accuracy of their magnitude judgments and reproductions. For example, Akdoğan and Balcı (2017) asked participants to reproduce a target duration, after which participants rated their confidence level regarding the proximity of their reproduction to the target (i.e., timing error). Finally, participants were asked to judge whether their time reproduction was shorter or longer than the target. They found that signed confidence ratings tracked the timing errors' amount and direction (i.e., under vs. over-reproduction). Subsequent work supported this conclusion in temporal (Kononowicz, Roger & Wassenhove, 2018; Doenyas et al., 2019; Öztel, Eskenazi & Balcı, 2020), numerical (Duyan & Balcı, 2018; 2019) and spatial judgments (Duyan & Balcı, 2020), which constitutes a general metric error monitoring ability (Yallak & Balcı, 2021).

The current study, for the first time, investigates the extent to which metric error monitoring relies on endogenous processes that lead to first-order timing performance and the role of belief of agency in this relationship. We also investigated whether this presumed reliance of metric error monitoring on endogenous generative processes differs for confidence vs. error directionality judgments. For this reason, instead of utilizing unitary response variables to assess the temporal error monitoring ability (e.g., continuous rating scheme that unifies error direction and magnitude in a single response variable as in Öztel & Balcı, 2023), we assessed the ability to monitor the temporal error magnitude and direction in separate variables (i.e., confidence and error direction judgments, respectively). We hypothesized that a) if the temporal error monitoring ability relies on endogenous instead of exogenous cues (e.g., one as the observer of own or someone else's behavior), error monitoring performance would be superior for monitoring committed compared to observed timing responses. On the other hand, if the temporal error monitoring relies on exogenous observational cues, one would expect similar error monitoring performance irrespective of the error agency. b) If what differentiates the error monitoring performance for committed and observed timing behaviors is differential belief in agency, then we expected better metric error monitoring performance when the participants believed that the observed performance belonged to them. Finally, c) if the two indices of temporal error monitoring ability (error directionality and confidence) are independent, we expected differential sensitivity for these indices to the endogenous and exogenous influences.

In support of our primary hypotheses, Bates et al. (2005) showed differential brain signals during self-committed and observed errors using 2AFC paradigms. This procedural detail is important because, given the dissociation between the perceptual and temporal error monitoring abilities (i.e., absence of correlation between the temporal and perceptual error monitoring abilities as demonstrated in Yallak & Balcı, 2021, along with the correlation between temporal, spatial and numerical error monitoring abilities), these results do not readily apply to temporal error monitoring. Thus, we investigated how a) committed vs. observed temporal errors, b) motor involvement, and c) belief about the agency of temporal errors affected the error monitoring performance. The first two experiments compared the temporal error monitoring performance for committed vs. observed timing performance. The third experiment tested the same question while controlling for the motor aspects of responding that are an integral part of committed responses. In the fourth and fifth experiments, we examined how the induced belief of agency about observed first-order performance (thus also controlling for motor involvement) affected the temporal error monitoring performance. In the sixth experiment, we examined whether the effect of agency belief depends on the design characteristics due to attentional and prediction-based differences (randomized across trials instead of blocks: Pfabigan et al., 2014). Each participant was tested in only one of the six experiments.

4.3. Experiment 1

The first experiment tested whether the temporal error monitoring ability relies on internal or external factors by evaluating error monitoring in self-performance vs. observed timing behaviors.

4.3.1. Method

4.3.1.1. Participants. Twenty-seven of the thirty participants were Koç University students who received one extra course credit for their participation. Data from twenty-one participants were included in the formal analyses (18 female, 19 right-handed, $M_{age} = 20.7 \mp 1.63$, 1 participant did not specify age). Two participants were discarded from the dataset due to unreasonable reproduction distributions (i.e., too short or too long reproductions or clear signs of explicit counting such as reproductions too closely distributed around the target duration with very little variance). Six participant did not complete the experiment. One participants were excluded due to clear signs of attentional distraction (e.g., emailing the experimenter during the experiment). Informed consent was obtained from each participant. This study was approved by Koç University Institutional Review Board (IRB, protocol number: 2020.113.IRB3.053).

4.3.1.2. Stimuli and Apparatus. The timing stimulus was a blue square placed at the center of the screen (height x width = 0.2 x 0.2 proportion of the screen size that adjusts according to the screen size). The short and long durations were 1500 and 3000 milliseconds. The experiment was programmed in the Builder mode of PsychoPy (Peirce et al., 2019). The stimuli were presented, and responses were recorded via the Pavlovia online experimentation platform, which has been tested to have under four milliseconds of reaction time precision in all web browsers under the optimized conditions (Bridges et al., 2020).

4.3.1.3. Procedure. The experiment was conducted in a 2 (experimental condition: self-reproduction vs. observation) or x 2 (target duration: short or long) full factorial within-subject design. The experimental conditions were randomized across the four blocks, and each participant completed the experimental conditions in a different order. Participants were instructed not to count or use any strategy to help them estimate the durations throughout the experiment (Rattat et al., 2012).

Each trial began with the presentation of the target duration, which was constant within the test block as in Akdoğan & Balcı (2017) to avoid any potential temporal context effects (e.g., Bausenhart et al., 2014; Gu & Meck, 2011; Jazayeri & Shadlen, 2010). In the self-reproduction condition, participants reproduced the target duration at any time they liked. The reproduction began with a space button press and terminated with a second space button press. Participants saw the identical blue square in the center of the screen throughout their reproduction. After each reproduction, participants were asked to rate how confident they were regarding their reproduction's accuracy (i.e., closeness to the target duration) with 1, 2, and 3 response keys for low, middle, and high confidence, respectively. Finally, participants were asked to classify their reproductions as shorter or longer than the target duration with V and N response buttons, respectively. No feedback was provided. The reason why participants were always asked to make a confidence judgements before the short-long judgements were two fold: (1) the confidence judgements aimed to evaluate the reproduction accuracy instead of short-long judgements and for this reason, (2) asking confidence judgements after short-long judgements could introduce a confusion for participants. Figure 4.1. illustrates the trial representation of the task. Crucially, the online data collection (2021-2022) approach is expected to improve the generalizability of the obtained results by providing a more naturalistic environment for participants to complete the experiment compared to the highly controlled laboratory experiments.

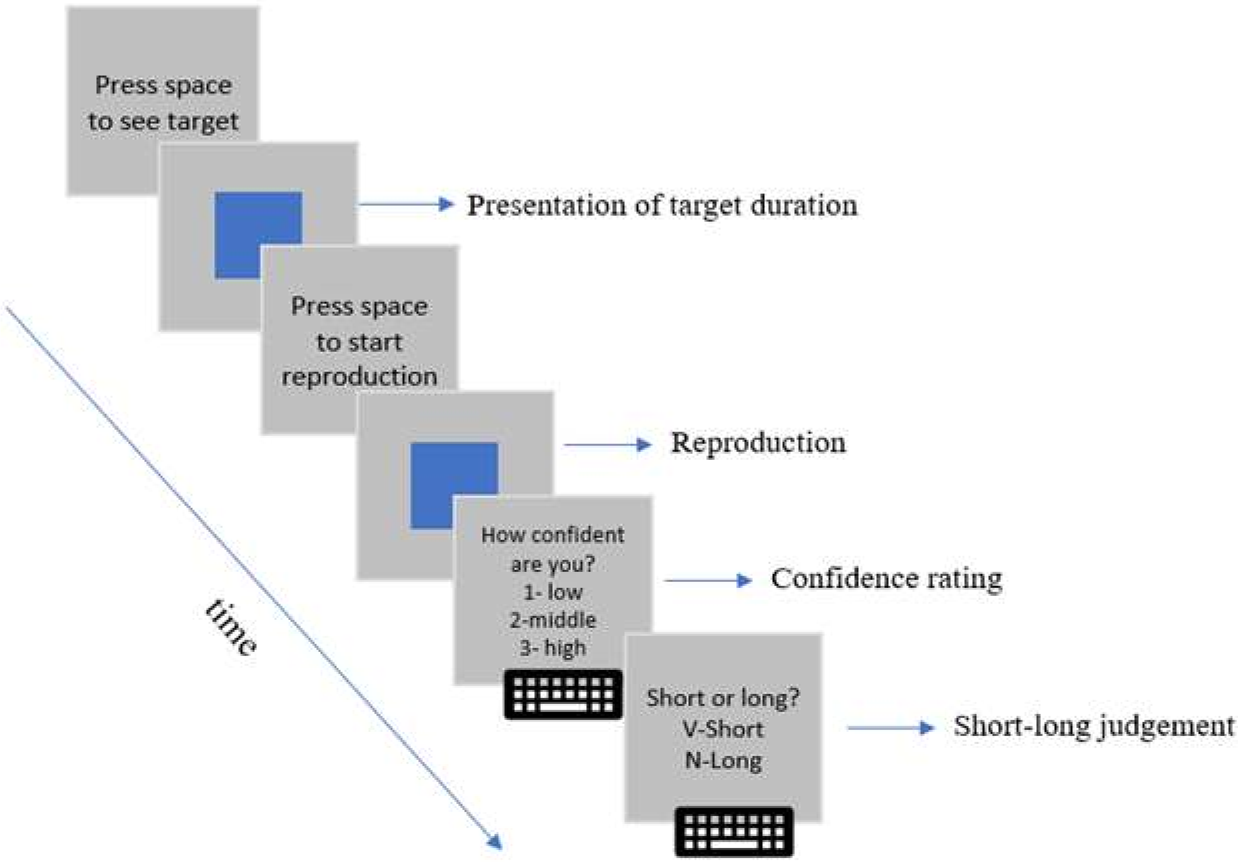


Figure 4.1. Trial representation of the task. The presentation of the target duration was 1.5 or 3 seconds, depending on the test block. Similarly, the confidence rating and the short-long judgments were done according to either observed or performed reproduction, depending on the test block.

The observation condition was the same as the self-reproduction condition, except that participants did not reproduce the target duration and saw the reproduced duration of another participant instead (tested in Akdoğan and Balcı (2017) - Experiment 4). After observing each reproduction, participants were asked to judge how confident they were about the accuracy of the observed reproduction. Finally, participants were asked to classify the observed reproduction as shorter or longer than the target duration. All the response-key mappings were visible until a response was recorded.

Participants took one minute break between each test block. At the beginning of each test block, participants were informed whether the upcoming test block would be an "observation" or "self-reproduction" condition. At the beginning of each test block, participants completed two training trials to practice the conditions of the corresponding block. In the training trials of the "observe" condition, participants saw a normally distributed random duration with the mean and standard deviation of the previous participant's reproduced durations. Each test block consisted of 60 trials. Thus, the experiment consisted of 4 x 60 = 240 trials, which took around 40 minutes to complete. The data from Akdoğan and Balcı (2017) is available as part of the Confidence Database (Rahnev et al., 2020). All experiments in this study were run in Turkish.

4.3.2. Analytical Approach

4.3.2.1. Data filtering. Before the data analysis, we z-score transformed the time reproductions (will be hereafter referred to as "signed timing error"). Trials that were three standard deviations above the mean of the reproduced durations in self, and observed reproductions separately for short and long target durations were excluded from the data analysis (i.e., smaller than -3 z score or larger than 3 z score). The reason for this filtering method was to avoid any artificial support for the temporal error monitoring hypothesis since these reproductions would be too easy to monitor in terms of their magnitude and direction. Note that although this criterion was not preregistered, it is a more conservative approach to data analysis and it was applied before the data analysis. As a result of this filtering method, 1.1% of the data were discarded from the whole dataset.

Our participant inclusion criteria were solely based on the following rules about the reproduction performance: a) unimodally distributed reproductions, b) no apparent drift in the reproductions over the course of the test session, and c) clustering of reproductions around the target interval (e.g., as opposed to very short intervals).

4.3.2.2. Linear mixed-effects analysis. In order to capture all data in a single model, we tested our hypothesis in a linear mixed-effects model. For the confidence judgments, we used absolute timing error (i.e., the absolute value of the signed timing errors), the target duration (i.e., short or long) to be reproduced, and the experimental condition (i.e., self-reproduction or observation) as fixed effects on the regression slope. We introduced random intercept across the participants. The full model formula is as depicted below:

Model 1: Confidence judgments ~ absolute timing error * target duration * experimental condition + (1|participant) where ~ refers to "predicted from" and 1|participant refers to "random intercept across participants." This model was performed to investigate how closely participants could match their and others' error magnitude irrespective of the error direction to their confidence judgments.

For the short-long judgments, we tested our hypothesis in a logistic mixed effects model, which was as follows: Model 2: Short-long judgments ~ signed timing error * target duration * experimental condition + (1|participant). The short-long judgments in the model involved the raw short or long classification of the reproduction with respect to the target duration, regardless of its accuracy. We ran this model to investigate how precisely participants could classify their own and others' error direction. These models were used for hypothesis testing in all experiments and included all lower terms. For all categorical comparisons, we used repeated contrast scheme in Jamovi; which calculates the difference between the two subsequent levels of the categorical variable (see also: https://gamlj.github.io/rosetta_contrasts.html).

We did not have any a priori expectation for the temporal error monitoring ability to be different for participants with different mean confidence and short-long judgments. Moreover, the high complexity of random effect structures often results in convergence and singular fit issues. Thus, to avoid unnecessarily complex random effect structures and keep the models as parsimonious as possible, we did not include random slopes for the experimental conditions in any of our mixed effects models. Nevertheless, to capture the variability induced by all cluster variables (for a more detailed discussion, see Barr et al., 2013), we attempted to fit all models with an additional random slope component. As expected, all the models that we attempted to fit with more complex random effect structures either raised singular fit issues (potentially due to overfitting) or failed to calculate the model R^2 value, which we presented in Supplementary Online Materials (Table S1.3.).

For simplicity, we only report the results that are critical for our hypotheses in the main text. The remaining results are reported in Supplementary Online Materials (SOM).

4.3.2.3. Statistical Reporting and Labeling. For the sake of simplicity, we deliberately omit the name of the continuous variables from the ß labeling, which depict slope difference for all relevant interaction terms.

4.3.3. Results

4.3.3.1. Confidence judgment. Linear mixed effects model revealed a statistically significant three-way interaction effect of absolute timing errors, target duration, and experimental condition on confidence judgments ($ß_{short-long * self reproduction-observation} = 0.16$, *SE* = 0.07, *p* = 0.015); the negative slope that depicts the relationship between the absolute timing errors and confidence judgments was steeper for the self-reproduction condition; and that this negative slope difference was significantly different for short and long target durations. Simple slopes revealed a significant negative relationship between absolute timing error and confidence judgment for all conditions except for the long duration- observation condition ($ß_{long observation} = -0.014$, *SE* = 0.03, *p* = 0.67; for all remaining simple slopes *p*s <0.001). This result revealed that participants demonstrated overall temporal error monitoring ability in all conditions, except for the observation condition with a long target duration. Figure 4.2.A and 4.2.B in the left column illustrates the relationship between absolute timing error and confidence judgments across the experimental conditions for long and short target durations, respectively.

4.3.3.2. Short-Long judgment. Logistic mixed effects model revealed a statistically significant three-way interaction between signed timing errors, target duration, and experimental condition ($ß_{short - long * self reproduction - observation} = -0.36$, exp(ß) = 0.70, *SE* = 0.16, *p* = 0.024). This result demonstrates that the difference in the steepness of the logistic curve between self-reproduction and observation conditions is significantly lower for the short target duration. Simple slopes demonstrated that participants could correctly classify their temporal error direction in both experimental conditions ($ß_{self reproduction} = 0.85$, exp(ß) = 2.38, *SE* = 0.057, *p* < 0.001; $ß_{observation} = 0.93$, exp(ß) = 2.54, *SE* = 0.055, *p* < 0.001). However, the two-way interaction between the experimental condition and signed timing error was not statistically significant (*p* = 0.29). Figures 4.2.C and 4.2.D in the left column illustrate the relationship between signed timing error and short-long judgments across the experimental conditions for long and short target durations, respectively.

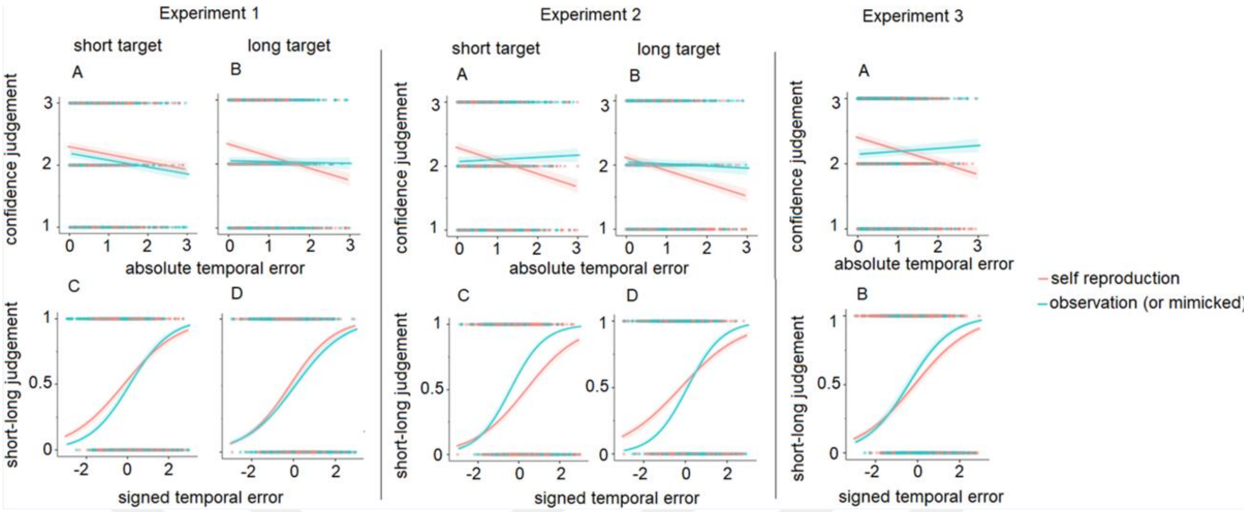


Figure 4.2. The relationship between timing errors and subjective timing error judgments across target durations and experimental conditions for Experiment 1 (most left), Experiment 2 (middle), and Experiment 3 (most right). In all panels, the shaded area depicts the standard error (SE) of the estimates.

4.3.4. Interim Discussion

Participants could not correctly match their confidence judgments to the observed timing errors only in the observed-long target duration condition. This result demonstrates that confidence judgments are guided more by endogenous than exogenous factors. Finally, a steeper logistic function that depicts the relationship between timing error and short-long judgments for the observation condition compared to the self-reproduction condition suggests that endogenous cues might have a disruptive effect or additional information provided in the observation condition would have an enhancing effect on the precision of directional error judgments. However, this result was driven mainly by the target duration.

4.4. Experiment 2: Replication

Experiment 2 aimed to test whether the effects we observed in Experiment 1 are replicable with a different set of stimuli. Furthermore, the observed patterns that are differential across target durations in Experiment 1 could be due to the fact that the production of short and long target durations was gathered from different participants in Akdoğan and Balcı's (2017) Experiment 4 data). To control for this difference, we used the same participants' reproductions of both short and long duration in the self condition of Experiment 1 in the "observed" condition of Experiment 2.

The third objective of Experiment 2 was to test whether the effects obtained in Experiment 1 could be generalized to different sets of observed temporal reproductions, which would also improve the ecological validity of the obtained results.

4.4.1. Method

4.4.1.1. Participants. 59 Koç University undergraduate students participated in this study in return for one extra course credit. 44 of them completed this study. 21 individuals were included in the formal analysis (17 female, all right-handed, M_{age} = 21.71 ∓ 4.47). Other participants were excluded due to unreasonable reproduction distributions (as in Experiment 1) or several other reasons such as participating in Experiment 1 before, completing the experiment more than once, and other technical reasons (e.g., internet connection problems, software crashes, non-Turkish speakers etc.).

4.4.1.2. Stimuli, Apparatus, and Procedure

All stimuli, apparatus, and procedure remained the same as in Experiment 1, except that in the observation condition, the reproduced durations were matched to the participants' reproduced durations in Experiment 1.

4.4.1.3. Analytical Approach

We applied the same data filtering method as in Experiment 1 before the analyses. This resulted in no data loss for Experiment 2.

4.4.2. Results

4.4.2.1. Confidence judgments. The interaction between the absolute timing error and experimental condition on isolated confidence judgments revealed a statistically steeper negative slope for the self-reproduction than the observation condition ($ß_{self\ reproduction\text{-}observation}$ = -0.20, SE = 0.04, $p < 0.001$). The simple slopes of the self condition were statistically significant for both target durations in the self condition ($ß_{self\ reproduction,\ short}$ = -0.204, SE = 0.035, $p < 0.001$; $ß_{self\ reproduction,\ long}$

$= -0.20$, $SE = 0.033$, $p < 0.001$), whereas for the observation condition, the simple slopes were not significant ($ß_{observation, short} = 0.034$, $SE = 0.036$, $p = 0.345$; $ß_{observation, long} = -0.033$, $SE = 0.036$, $p = 0.360$). This result indicates that participants could not monitor the magnitude of the observed timing errors. Figure 4.2.A and 4.2.B in the middle column illustrates the relationship between absolute timing error and confidence judgments across the experimental conditions for long and short target durations, respectively.

4.4.2.2. Short-Long Judgments. The interaction between the experimental condition and signed timing error was statistically significant ($ß_{self\ reproduction\text{-}observation} = -0.48$, $\exp(ß) = 0.62$, $SE = 0.09$; $p < 0.001$), revealing an overall steeper logistic function for the observation condition. Additionally, as in Experiment 1, simple slopes demonstrated that participants could correctly monitor the direction of their temporal errors in both experimental conditions ($ß_{self\ reproduction} = 0.73$, $\exp(ß) = 2.07$, $SE = 0.055$, $p < 0.001$; $ß_{observation} = 1.21$, $\exp(ß) = 3.34$, $SE = 0.065$, $p < 0.001$). Figures 4.2.C and 4.2. D in the right column of the panels illustrate the relationship between signed timing error and short-long judgments across the experimental conditions for long and short target durations, respectively.

4.4.3. Interim Discussion

Experiment 2 demonstrates a higher precision in error direction classification for the observed timing errors than self-performed ones. In Experiment 2, this difference in observed and self-performed errors did not depend on the target durations (cf. Experiment 1).

The significant negative relationship between error magnitude and the confidence ratings was only observed for errors committed by the participants themselves and not for observed error magnitudes. This result partially corroborates the results of Experiment 1, which demonstrated a null effect only for the “long” target duration in observed timing errors. Thus, consistent with the conclusion of Experiment 1, the accurate matching of confidence judgments to the temporal error magnitudes seems to be guided more by the endogenous than the exogenous factors.

The reason why the results of Experiment 2 did not fully match those of Experiment 1 might be because the observed durations of Experiment 1 were gathered from a between-subject design in Akdoğan and Balcı (2017) experiment while the observed durations of Experiment 2 were gathered from the same participants (within-subject design during Experiment 1).

4.5. Experiment 3: Controlling for the motor action

In two experiments, the results that pointed to differential temporal error monitoring performance might be because of a) motor action that is emitted during the error commitment and/or b) the agency (i.e., owning the error or not) across experimental conditions. Experiment 3 aimed to control for the motor involvement during the error commitment in two experimental conditions. To address this, instead of passively observing others' temporal reproductions, participants in Experiment 3 mimicked the previous participants' reproductions (i.e., "mimicked observation" condition instead of mere "observation" condition). On the other hand, if it is the agency that caused the differential error monitoring ability across conditions in Experiments 1 and 2, their results should be replicated in Experiment 3.

For the sake of experimental and analytical simplicity, we included only the long target duration in the experimental procedure from Experiment 3 forward. Note that testing a single test duration (3000 ms) in the rest of the experiments yielded a one less predictor for the linear mixed effects models.

4.5.1. Method

4.5.1.1. Participants. 79 Koç University undergraduate students participated, and 56 completed this study. Due to various reasons listed in Experiment 2, 27 of them were included in the formal analysis (21 females, 3 left-handed, $M_{age} = 20.66 \mp 1.62$).

4.5.1.2. Stimuli, Apparatus and Procedure. The procedure was the same as in the first two experiments. In the self-reproduction condition, participants reproduced the target duration themselves (both short and long durations, in a blockwise randomized order) and made confidence and short-long judgments regarding their own reproductions (motor-self blocks). Participants were asked to mimic a previous participant's reproduction in the mimicked reproduction condition. Upon the beginning of the reproduction, participants pressed the spacebar (as if they initiated the reproduction). Upon the termination of the reproduction (i.e., the disappearance of the stimulus that represents the reproduction), participants pressed the spacebar the second time (as if they terminated the reproduction themselves). If the participants pressed the space button before the blue square disappeared from the screen or later than 500 milliseconds upon its disappearance, they received a warning on the screen ("Please do not press the space button before you see that the blue square disappeared from the screen" and "You should be faster", respectively). The reproductions mimicked in this condition were gathered from the "self-reproduction" conditions of Experiment 1. Finally, as in the case of self-reproduction condition, participants reported their confidence and short-long judgments regarding the reproductions they mimic with respect to the target duration. Note that discarding the short durations from the experimental design in the rest of the experiments yielded one less predictor for the linear mixed effects models compared to the first two experiments.

4.5.1.3. Analytical Approach

Prior to the analyses, we applied the same data filtering method as in the previous experiments. Additionally, the trials where participants failed to mimic the observed reproductions were also discarded. This filtering resulted in a 1.1% discard of the whole dataset.

4.5.2. Results

4.5.2.1. Confidence Judgments. There was a significant interaction between the absolute timing error and the experimental condition ($ß_{\text{self reproduction-mimicked reproduction}} = -0.23$, $SE = 0.046$, $p < 0.001$), pointing to a steeper slope for the self-reproduction compared to the mimicked reproduction condition. Simple slopes demonstrated that participants could correctly match their confidence judgments to the amount of their timing error only in the self-reproduction condition ($ß_{\text{self reproduction}} = -0.189$, $SE = 0.03$, $p < 0.001$) but not in the mimicked condition ($ß_{\text{mimicked reproduction}} = 0.044$, $SE = 0.035$, $p = 0.202$). These results replicate the results obtained in the first two experiments, even after equating the motor demands of the two conditions. Figure 4.2.A in the right-most column of panels illustrates the relationship between absolute timing error and confidence judgments.

4.5.2.2. Short-Long Judgments. The two way interaction between the signed timing errors and the experimental conditions was statistically significant ($ß_{\text{self reproduction-mimicked reproduction}} = -0.236$, $\exp(ß) = 0.790$; $SE = 0.105$, $p = 0.024$), pointing to a more precise classification of the mimicked reproductions compared to the self-performed reproductions (simple slopes: $ß_{\text{self reproduction}} = 0.748$, exp(ß) = 2.112; $SE = 0.066$, $p < 0.001$; $ß_{mimicked\ reproduction} = 0.983$, exp(ß) = 2.673; $SE = 0.083$, $p <$ 0.001). Figure 4.2.B in the right-most column of panels demonstrates the relationship between the signed timing error and short-long judgments.

4.5.3. Interim Discussion

In three experiments, we demonstrated that participants could only correctly match their own temporal error magnitudes to their confidence ratings. However, this ability did not generalize to monitoring someone else's timing temporal error magnitude as an observer. This effect reversed for the temporal error direction monitoring such that participants could better categorize someone else's error directions than the direction of self-made errors. Experiment 3 demonstrates that agency, instead of motor involvement, is the critical factor that led to the difference across experimental conditions in terms of temporal error monitoring ability.

Although the motor action was controlled in the "mimicked observation" condition in Experiment 3, the obtained results might have occurred due to a) actual agency or b) mere belief of agency. Experiment 4 complemented Experiment 3 by eliminating motor response from both experimental conditions to directly test the "belief of agency" on temporal error monitoring.

4.6. Experiment 4

Experiment 4 used participants' own reproductions in both "belief of agency" and "belief of non agency" conditions, and no response was emitted for the first-order task during testing (observation only). Resultantly, participants made metacognitive judgments for their own temporal errors throughout the experiment but on the random half of the blocks, they were told that the reproduction belonged to a previous participant (i.e., "belief of non agency" condition). If the experimentally induced sense of agency guides the temporal error monitoring ability, one should observe similar results as in the first three experiments.

4.6.1. Method

4.6.1.1. Participants. 54 Koç University undergraduate students participated in Experiment 4 for one extra course credit (40 completed the experiment). Out of 40, 14 were discarded before the formal analyses due to various reasons, as in the previous experiments. Resultantly, data from 26 participants were used in the formal analysis (22 female, one did not specify gender, none reported left-handedness, $M_{age} = 21.24 \mp 1.56$).

4.6.1.2. Procedure. The experiment included three main phases. In the training phase, participants saw the target duration of 3000 milliseconds five times after which they were asked to reproduce the target duration five times. Each reproduction was demarcated by two successive space button presses.

In the second phase, participants reproduced the target duration a hundred times (reproduction part). They were shown the target duration upon each reproduction. In the third phase, participants were shown with their own time reproductions (random order) recorded during the second phase. Thus, participants observed their own durations without any motor action. At the end of each observation, participants first rated how confident they were (by pressing 1, 2 and 3 buttons on the keyboard where 1- low confidence, 2- medium confidence, 3-high confidence) regarding the proximity of the reproduction they observed to the target duration in two belief conditions (i.e., "belief of agency" condition, where instruction was "YOUR reproduction" and "belief of non agency" condition, where instruction was "OTHER'S reproduction"). The experimental conditions were presented in blocks and their order was randomized across participants. After the confidence judgments, participants reported the direction of the reproduction they observed with respect to the target duration (the reproduction was S- shorter, L- longer than the target duration). Note that participants did not emit motor responses for the first-order task in either experimental condition (controlling for the motor involvement). Participants took a one-minute break in the middle of both self-reproduction and the observation blocks. The experiment took approximately 35 minutes to complete.

4.6.2. Analytical Approach and Results

Prior to the data analysis, we followed the same filtering method and mixed effects models as in the previous experiments; 1.2% of the whole dataset was discarded before the formal analysis.

4.6.2.1. Confidence judgments. The simple slopes for both experimental conditions were statistically significant ($\beta_{\text{belief of non agency}} = -0.12$, $SE = 0.036$, $p < 0.001$; $\beta_{\text{belief of agency}} = -0.23$, $SE = 0.037$, $p < 0.001$). On the other hand, the significant interaction between the absolute timing error and the experimental condition ($ß_{belief\ of\ agency-belief\ of\ non\ agency}$ = -0.11, SE = 0.05, p = 0.036) pointed to a steeper negative slope for the "belief of agency" compared to the "belief of non agency" condition, which revealed a stronger temporal error monitoring ability in the self condition. The leftmost column of Figure 4.3.A illustrates the relationship between absolute timing error and confidence judgments across experimental conditions.

4.6.2.2. Short-long judgments. The main effect of the experimental condition and the interaction between the experimental condition and signed timing error on short-long judgments were not significant (main effect: $M_{belief\ of\ agency\ -\ belief\ of\ non\ agency}$ = 0.050, SE = 0.10, exp(ß) = 1.050, p = 0.61; interaction: $ß_{belief\ of\ agency\ -\ belief\ of\ non\ agency}$ = -0.03, SE = 0.15, p = 0.86). The leftmost column of Figure 4.3.B illustrates the relationship between signed timing error and short-long judgments across experimental conditions.

4.6.3. Interim Discussion

The slope differences showed that participants were better at monitoring the error amount when they believed the observed reproduction belonged to themselves. This result suggests that the confidence judgments are sensitive to the belief of agency.

This effect was not present in the short-long judgments. Participants could equally report the error direction regardless of the belief of agency. This could be simply due to the elimination of the response component; in other words, responding might be a source of variability for the short-long judgments (e.g., due to sensory-motor system variability).

4.7. Experiment 5

Experiment 5 aimed to replicate Experiment 4 again by using two different instructions. In the observation condition of Experiment 5, we used simulated data as a normally distributed random variable drawn from a distribution with parameter values estimated (i.e., the mean and the standard deviation (SD)) from the corresponding participant's empirical performance. This approach helped us to control for the differential reproduction distributions in terms of their precision, which we quantify with the coefficient of variation (CV) of the reproduced durations. The CV of the reproductions were calculated as follows: CV = $SD_{reproduction}$ / $M_{reproduction}$. This resulted in a 2 (instruction) x 2 (source of data) factorial design: belief of agency-empirical reproduction, belief of agency-simulated reproduction, belief of non-agency-empirical reproduction, and belief of non agency-simulated reproduction.

4.7.1. Method

4.7.1.1. Participants. 52 Koç University students participated in Experiment 5 for one extra credit. Eight did not complete the experiment. Out of 44, 19 of them were excluded before the formal analysis due to not completing the experiment, reproducing subsecond durations throughout the experiment, or too high CV that rendered matching the CV for the "simulated reproduction" condition difficult and other technical reasons. Resultantly, data from twenty-five participants were included in the formal analyses (21 females, $M_{age} = 20.7 \mp 1.31$).

4.7.1.2. Stimuli, Apparatus and Procedure. All stimuli, apparatus and procedures were kept the same as in Experiment 4. Except instead of showing participants their own, we simulated the reproductions in the "other" condition such that they matched the participants' mean reproduction and CV. This resulted in a 2 x 2 within-subject-blocked design where the reproductions either belonged to the participants (i.e., empirical reproduction) or not (i.e., simulated reproduction), and the instruction was either "YOUR" or "OTHER'S". Empirical reproductions were presented in a randomized order.

4.7.1.3. Analytical Approach

Prior to the data analysis, we applied the same trial-by-trial filtering method as in the previous experiments. This filtering method resulted in 0.9% discard of the whole dataset. Additionally, we discarded all participants whose CV was above 0.3 as higher than these values that are driven by the high standard deviations that resulted in unreasonable values in the distributions of simulated data for the "other" condition that we aimed to match to participants' CV. We performed the same linear mixed-effects models separately for confidence and short-long judgments as in the previous experiments.

4.7.2. Results

4.7.2.1. Manipulation Check. To test whether the simulated reproductions indeed matched the empirical reproductions in terms of the statistical parameters (i.e., the mean and the standard deviation), we compared the distributions of empirical and simulated reproductions. Linear mixed effects model revealed no statistically significant difference between the empirical and simulated reproductions ($M_{\text{simulated reproduction-empirical reproduction}} = 0.014$, $SE = 0.018$, $p = 0.43$; Bayesian paired sample t-test $BF_{01} = 2.951$, error % = 0.027, prior defined as Cauchy distribution), which reveals an overall match between the distribution of two types of reproduction. The non-difference of two distributions was supported by Kolmogorov-Smirnov test results (K-S test output and descriptive statistics associated with the two distributions can be found in Supplementary Online Materials; Table S2.12., S2.13. and Figure S2.1.).

4.7.2.2. Confidence Judgments. Simple slopes that depict the relationship between the absolute timing error and confidence judgments in both instruction types were statistically significant ($ß_{\text{belief of agency}} = -0.20$, $SE = 0.02$, $p < 0.001$; $ß_{\text{belief of non agency}} = -0.13$, $SE = 0.02$, $p < 0.001$). Importantly, the significant interaction between the instruction type and the absolute timing error on confidence judgments ($ß_{\text{belief of non agency-belief of agency}} = 0.07$, $SE = 0.03$, $p = 0.03$), pointed to a better temporal error monitoring when participants believed that the reproduction they observed belonged to themselves (belief of agency); regardless of the actual source of the data (empirical vs. simulated). The middle column of Figure 4.3.A illustrates the relationship between absolute z-score reproduction and confidence judgments across experimental conditions.

4.7.2.3. Short-Long Judgments. Simple slopes revealed that participants could correctly classify the observed reproductions as "long" in both reproduction type conditions ($ß_{\text{empirical reproduction}} = 1.11$, $SE = 0.06$, exp(ß) = 3.03, $p < 0.001$; $ß_{\text{simulated reproduction}} = 1.39$, $SE = 0.06$, exp(ß) = 4.01, $p < 0.001$). The significant interaction between the source of data and the signed timing error revealed a more sensitive classification of the observed reproductions as "long" when they were simulated, regardless of the belief condition ($ß_{\text{simulated reproduction-empirical reproduction}} = 0.28$, $SE = 0.09$, exp(ß) = 0.086, $p = 0.001$). All remaining effects were not statistically significant (all ps > 0.05). The middle column of Figure 4.3.B illustrates the relationship between signed timing errors and short-long judgments across experimental conditions.

4.7.3. Interim Discussion

Experiment 5 demonstrated that the confidence judgments were susceptible to the belief of agency. The worsening rather than elimination of error monitoring performance in the belief of non-agency condition of Experiment 5 (compared to the results of the earlier experiments) can be attributed to potential deviations in the CVs of first-order performance used for different reproduction conditions and the elimination of the motor component in Experiment 5. Interestingly, short-long judgments were more sensitive when the observed reproduction was simulated instead of owned irrespective of the belief conditions. Briefly, consistent with Experiment 4, the current results deem belief of agency as a critical factor in metric error monitoring performance in terms of confidence ratings. Experiment 5 also supported the idea that confidence and short-long judgments are dissociative indices of metric error monitoring ability with sensitivity to different error-related factors (i.e., agency on the errors or belief of agency).

4.8. Experiment 6

In all five experiments, we randomized the experimental conditions across the blocks instead of the trials. Thus, the effect of the blockwise randomization approach on the obtained results is not clear. Experiment 6 addressed this possibility. Experiment 6 aimed to test to what extent the blocking trials by the experimental condition previously shown to affect electrophysiological markers of error monitoring altered the obtained results. For instance, Pfabigan et al. (2014) found more distinct feedback-related negativity (FRN) peaks, larger effect sizes, and significant relation between FRN amplitudes and behavioral task performance measures in blocked vs. trial-based experimental design (critical manipulation being feedback).

Motivated by the differential effect by different predictability regarding the stimulus and experimental instructions in the blocked versus trial-based experimental design, we aimed to test how manipulating beliefs on a trial basis would change the obtained findings. We predicted that the effect of experimental manipulations (i.e., belief of agency on errors) would exert a weaker effect in trial-based manipulations. In Experiment 6, we tested our hypotheses in both lab and online environments (analyzed separately).

4.8.1. Method

4.8.1.1.1. Participants. *Lab experiment.* 32 Koç University undergraduate students participated in the laboratory counterpart of this study. Data from five participants were discarded from the formal data analysis due to unrealistic reproduction distributions. Thus, data from 27 participants were included in the formal analysis. (14 females, none reported left-handedness $M_{\text{age}} = 19.96 \mp 1.09$).

4.8.1.1.2. *Online study.* 44 Koç University undergraduate students participated in the online version of the study for one extra course credit. Thirty-three of them completed the study. Out of 33, 5 were discarded due to unrealistic reproduction distributions, and five were discarded due to technical problems. Data from 23 participants were included in the formal data analysis ($M_{\text{age}} = 20.5 \mp 1.57$, 17 females, one did not specify gender).

4.8.1.2. Stimuli, Apparatus and Procedure. Experiment 6's procedure was similar to Experiment 4 except that the experimental conditions were randomized across trials instead of across blocks. Thus, participants randomly received "YOUR" or "OTHER'S" instructions throughout the experiment according to the experimental condition. The experiment took approximately 35 minutes to complete. The experiment was run simultaneously in the lab and online setting. All stimuli presentations and response recordings were done in Pavlovia online experimentation platform during the online experiment. The lab experiment was done in PsychoPy on Windows 10 PC.

4.8.1.3. Analytical Approach

We applied the same data filtering method as in previous experiments for online and lab counterparts of Experiment 4 separately prior to the formal analyses; 1.2% and 1.6% of the whole datasets were discarded for online and lab data, respectively.

4.8.2. Results

4.8.2.1.1. Confidence Judgments. *Online experiment.* Neither the main effect of the belief condition on confidence judgments nor the interaction between the belief condition and absolute timing error were statistically significant (main effect of belief condition: $M_{\text{belief of agency - belief of non agency}} = 0.040$, $SE = 0.044$, $p = 0.36$; interaction: $ß_{\text{belief of agency - belief of non agency}} = 0.026$, $SE = 0.052$, $p = 0.61$; simple slopes: $ß_{\text{belief of agency}} = -0.32$, $SE = 0.04$, $p < 0.001$; $ß_{\text{belief of agency}} = -0.34$, $SE = 0.04$, $p < 0.001$). Together, these results indicate that participants could keep track of the amount of their timing errors even when they believed that they were rating confidence about someone else's temporal errors.

4.8.2.1.2. *Lab experiment.* The interaction between the experimental condition and absolute timing error was not statistically significant ($ß_{\text{belief of agency - belief of non agency}}$= -0.050, SE = 0.053, p = 0.35; simple slopes: $ß_{\text{belief of agency}}$ = -0.24, SE = 0.04, $p < 0.001$; $ß_{\text{belief of agency}}$ = -0.19, SE = 0.04, $p < 0.001$), revealing that the temporal error monitoring ability did not change across the belief conditions as in the online counterpart of the experiment. The right-most column of Figure 4.3. illustrates the relationship between absolute timing errors and confidence judgments across belief conditions (lab setting: Panel A and online setting: B).

4.8.2.2.1. Short-long Judgments. *Online experiment.* There was no interaction between the belief condition and signed timing error and the main effect of belief conditions on long judgment probability (main effect of belief condition: $M_{\text{belief of agency - belief of non agency}}$ = 0.18, SE = 0.10, exp(ß) = 1.20, p = 0.085; interaction: $ß_{\text{belief of agency - belief of non agency}}$ = 0.064, SE = 0.15, exp(ß) = 1.07, p = 0.68; simple slopes: $ß_{\text{belief of agency}}$ = 1.45, SE = 0.11, exp(ß) = 4.28, $p < 0.001$; $ß_{\text{belief of non agency}}$ = 1.39, SE = 0.11, exp(ß) = 4.01, $p < 0.001$).

4.8.2.2.2. *Lab experiment*. Neither the main effect of the belief condition nor the interaction effect was statistically significant (main effect of belief condition = $M_{\text{belief of non agency - belief of agency}}$ = -0.050, exp(ß) = 0.10, SE = 0.10, p = 0.62; interaction: $ß_{\text{belief of agency - belief of non agency}}$ = -0.13, exp(ß) = 1.14, SE = 0.16, p = 0.41; simple slopes: $ß_{\text{belief of agency}}$ = 1.80, SE = 0.12, exp(ß) = 6.08, $p < 0.001$; $ß_{\text{belief of non agency}}$ = 1.94, SE = 0.12, exp(ß) = 6.94, $p < 0.001$). The right-most column of Figure 4.3. illustrates the relationship between signed timing error and short-long judgments across belief conditions (lab setting: Panel C and online setting: D). Table 4.1. summarizes the procedural differences across all six experiments.

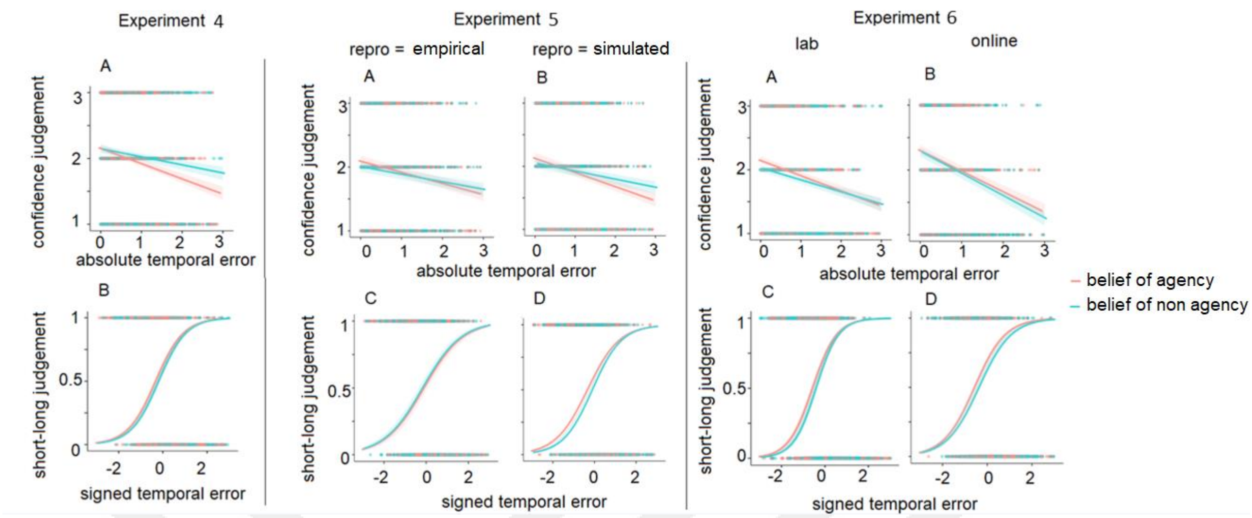


Figure 4.3. The relationship between timing errors and subjective timing error judgments across target durations and experimental conditions for Experiment 4 (most left), Experiment 5 (middle) and Experiment 6 (most right). The experimental conditions depicted in red express "YOUR (in "belief of agency" condition)" and the blue line expresses "OTHER'S (in "belief of non agency" condition)" instructions. In all panels, the shaded area depicts the standard error (SE) of the estimates.

Exp number	Exp Setting	Source of error in “non-self reproduction” condition	Distribution of exp conditions	Critical Manipulation	Motor Component in “self reproduction” Condition	Aim of the manipulation	Sample size	Cues for agency
1	Online	Another participant	Blockwise	Original Experiment	Present	To test the influence of motor actions	N = 21	Motor component
2	Online	Another participant	Blockwise	Replication (Experiment 1)	Present	To test the replicability of Exp 1	N = 21	Motor component
3	Online	Another participant	Blockwise	Controlling for motor response	Present	To disentangle the effect of motor action and agency	N = 27	Experimental instruction

4	Online	Participant's own data (in the training block)	Blockwise	Replace other with participant's own data	Absent	To disentangle the effect of belief of agency	N = 26	Experimental instruction
5	Online	Normally distributed random variable based on participant's performance	Blockwise	Replace own data with simulated data	Absent	To replicate Exp 4 with two sources of data	N = 25	Experimental instruction
6	Online and Laboratory	Participant's own data (in the training block)	Trialwise	Change blockwise design to trial-based design	Absent	To test whether the effect depends on attention	N = 27	Experimental instruction

Table 4.1. *Summary of methods in all experiments.*

4.8.3. Interim Discussion

Experiment 6 revealed that the temporal error monitoring ability did not depend on the belief of agency when induced trial-by-trial. Thus, whatever effect we observed regarding the effect of agency belief is particular to block-based designs. This is in line with our predictions derived from the differential predictability of instructions in block-based and trial-based manipulations (Pfabigan et al., 2014).

Overall, in experiments where the participants are not acting, the nature of agency turns into "error ownership". Thus, in experiments where participants merely observe and assess the performance (Experiment 3 to Experiment 6), any difference between self-made and other-made errors could be attributed to error ownership effects; which is introduced by the agency in the baseline blocks in Experiment 3 to Experiment 6.

4.9. General Discussion

This study aimed to examine the effects of different endogenous factors, such as producing the timed response and belief of agency on temporal error monitoring. In all six experiments, we observed a generally robust temporal error monitoring ability (at least for the self-performance conditions) replicating previous findings (Öztel, Eskenazi and Balcı, 2020; Akdoğan & Balcı, 2017; Yallak & Balcı, 2021).

We found a dissociation in terms of the sensitivity of confidence ratings and short-long judgments to experimental manipulations: confidence judgments were sensitive to the belief of agency whereas the short-long judgments were sensitive to the actual agency (more precise categorization in non agency condition).

Critically, these observed patterns were beyond the mere involvement of motor action. Thus, by showing their different sensitivities to different influences, this study demonstrates that these two indices of temporal error monitoring are disassociated.

Below, we discuss our findings at various levels of explanation: 1) information processing level, 2) phenomenological and neural levels and 3) attentional level.

Information processing level. One potential difference between the two temporal error monitoring indices might stem from their potentially differential underlying sources of information. In Akdoğan and Balcı's (2017) model of temporal error monitoring, the directional (i.e., short-long judgments) and magnitude (i.e., confidence judgments) monitoring results from the comparison of threshold crossing times of two clocks that keep perceptual time vs. motor action (by a second stage decision process). The order of the threshold crossings determines the error direction judgment such that earlier threshold crossing of the motor clock results in a "shorter than target" error judgment. Further, the amount of discrepancy between the threshold crossing times between the two clocks informs the confidence ratings - larger discrepancy leading to larger deviation generated by a second stage decision process. If participants do not engage their motor clock in the "other" conditions (even through emulation), this would disrupt the generation of the deviation metric, which in turn leaves confidence ratings unguided. That being said, participants can still compensate for the lack of a second clock engagement, simply by using the time interval between the threshold crossing of their perceptual time and the "observed reproduction" as a proxy for confidence ratings, which might be less robust and less automatic than an endogenous comparison process (e.g., due to lack of an informative scale). On the other hand, the extraction of error direction judgments based on such a proxy would be easier (not requiring a scale either) since it would only rely on temporal order. In fact, using a proxy for error directionality judgments instead of an endogenous comparison process might even improve the precision of the error directionality judgments by eliminating non-timing related variability (e.g., variability of t0).

It is also possible that more precise error direction judgments in the "observed" condition are due to the use of an additional source of information (i.e., observed time reproduction) to help decide the error direction judgment. Under this rationale, the participants would be assumed to use both the sensory and motor clock as in the "self" conditions but would also integrate the observed reproduction to inform the error direction judgment further. For instance, if the motor clock hits the threshold earlier than both the sensory clock and the observed reproduction, this would provide stronger evidence for "shorter than target" error judgments while the evidence would be equivocal if the threshold crossing of the motor clock is demarcated by the observed reproduction and the threshold crossing of the sensory clock.

A similar explanation for the obtained results at the information processing level could be that the motor command led to time compression compared to the conditions where no motor action was taken for error commitment, which yielded the observed difference in the temporal error monitoring ability. However, the differential temporal error monitoring ability was replicated even when the motor action taking was controlled, eliminating this possibility. Future studies are needed to investigate these possibilities.

Phenomenological and neural levels. What can be the phenomenological distinction between the two indices of temporal error monitoring? The leading distinction might be that short-long judgments have an absolute accuracy (i.e., temporal order), but confidence judgments capture continuous subjective states regardless of the actor that made the decision/error. When the errors belong to another individual, the observed durations are merely classified as short or long without any introspective component that brings about metacognitive processing (see earlier discussions). Under this rationale, the error direction classification turns into a first-order processing that relies on objective perceptual cues without the involvement of subjective feelings (e.g., comparison of the subjective timing and other's timing behavior). This might exempt the short-long judgments from any additional noise introduced by the introspective processing during second-order decision-making (i.e., metacognition - Maniscalco & Lau, 2014) that targets another individual's performance. Accordingly, the errors that do not have agency attribution are classified more precisely according to their directions with respect to the target. This may also underlie the differential sensitivity of two temporal error monitoring components to experimental manipulations. On the other hand, the fully subjective characteristics of confidence judgments render them tacitly or phenomenologically introspective. Thus, the belief of agency might be sufficient per se to guide subjective confidence. Note that the phenomenological account outlined above does not differentiate between those cases that involve a motor response vs. not and only refers to the nature of the second order judgment.

Critically, the above mentioned phenomenological differences can also reflect itself at the implementational (neural) level. For example, error direction monitoring could be mainly encoded in the medial frontal and motor cortices (see Schie et al., 2004), dorsal and rostral parts of the anterior cingulate cortex (ACC, Shane et al., 2008); as these areas were found to suggest shared common neural markers for observed and performed errors in the MFC (including ACC) and motor cortex ("your mistakes are mine", Kang et al., 2010; see also: Bates et al.,2005; Spinelli et al., 2018; Cochin et al., 1999; Miltner et al., 2004). Similarly, the confidence judgments might be encoded in the dACC, OFC, TPJ and dmPFC areas due to their documented involvement in self/other distinction (e.g., Walton et al., 2004; Jiang et al., 2022). Neuroimaging and/or neuromodulation studies would help determine the possible implementational level differences that might explain the differential sensitivity of different indices of temporal error monitoring to agency manipulations.

Finally, the differentiation between the confidence and short-long judgments in the current study could have been a result of "intentional binding", which refers to temporal constriction for the durations between motor action and its physical consequences (Haggard et al., 2002). Accordingly, the intentional binding would predict short response bias for the self-made reproductions in short-long judgments. While this is indeed the case for several experiments (Experiments 2, 3, and 5), it is not persistent in all experiments. Although this inconsistency could be a result of methodological differences between the experiments, the persisting steeper slope for "other" conditions in short-long judgments renders the intentional binding an insufficient explanation for the current results. Furthermore, the potential intentional binding effect is mainly attributed to the overall response bias; instead of the metacognitive ability. Taken together, regardless of potential intentional binding effects; the main finding of the current study reveals a distinction between the error magnitude (i.e., as operationalized with the confidence judgments) and directionality monitoring (i.e., short-long judgments).

Attentional level. An additional possibility is that participants may have paid less attention to someone else's performance (believed to), which would diminish their ability to monitor them correctly. If this were the case, however, one would expect to see a similar diminishing effect for the error direction monitoring ability as well (i.e., as indexed with the short-long judgments), where the precision of error direction classification ability was improved for someone else's performance. Thus, while this alternative explanation could hold for the error magnitude monitoring ability (i.e., indexed with the confidence ratings), it does not apply to the error direction monitoring ability where the reverse effect was found. Thus, the obtained difference between the two error monitoring indices could not be directly attributed to the differential attentional allocation to self-made versus other-made errors.

However note that the disassociation between error direction and confidence was unique to the blocked designs. We attribute this difference to the differential attentional factors induced by the experimental designs (instead of self/other distinction): Manipulating agency belief in a trial-by-trial manner is much more sensitive to attentional effects since it requires participants to pay attention to the constantly changing trial conditions throughout the experiment. Participants could overcome this cognitive load by not paying attention to the instructed conditions of the trials and treat them all as attention requiring.

Block-based manipulation would be exempt from this kind of diminished effect of manipulation since the manipulation is constant across all trials within a block. Thus, the induction of the belief would be sufficient for the effect of the manipulation to persist until the end of that block.

Given methodological approaches in the six experiments, the current study does not allow us to come up with any mechanistic explanations on how the discrimination of the agency takes place in the error direction judgments in Experiment 5. We think that this is most likely due to distributional differences between empirical and simulated data or a statistical fluke. However, our K-S test comparisons of the empirical and simulated RTs showed that in the majority of the cases (80% of these cases), these two datasets were not discriminable. It is likely that the differences in the source distribution types (e.g., inverse Gaussian vs. Gaussian) might be an informative source of information (S2.12). Future research is required to directly test the dependence of our results on such possible differences.

We had a relatively high rate of data discard due to response distributions of first-order timing performance that fall below the acceptable data quality in the current set of experiments. All of these exclusions were realized prior to data analysis. There can be many different reasons for this, which can be listed as differential motivation for completing the study across different dates of periods in experimentation (e.g., beginning and end of the semester), and uncontrollable attentional distractions during online experimentation. Nevertheless, the data discarding method that we followed was based on the reproduction distributions; rather than the temporal error monitoring ability. Thus, our exclusion does not favor (in fact may disfavor) the temporal error monitoring hypothesis. Further, this ability has also been validated in samples with less discard rate (e.g., Yallak & Balcı, 2021). Briefly, the effects obtained in the current study cannot be attributed to the relatively high discard rate in the current study. Nevertheless, as for any other scientific work, the further replication of our studies would benefit the robustness of our findings.

Overall, the current study demonstrated distinctive features (in terms of sensitivity) of the error magnitude and direction monitoring components of metric error monitoring, which can be addressed at different levels. Future studies utilizing neuromodulation methods (e.g., transcranial magnetic stimulation (TMS), transcranial direct current stimulation (tDCS)) would help elucidate the causal roles of different brain areas in different components of metric error monitoring and their modulation by the belief of agency. For instance, these studies can target supplementary motor area, left cerebellum, right posterior parietal cortex, and/or right extrastriate body area that have been implicated in the sense of agency (e.g., Yomogida et al., 2010). Regarding the independence of confidence and error directionality judgments as two core components of MEM, our recent research using tDCS shows that precuneus and anterior prefrontal cortex might be involved in confidence judgments but not error directionality judgments (Yallak, Çiçek, Eskenazi & Balcı, in preparation). Coupling this work with the functional magnetic resonance imaging (fMRI) could help reveal the underlying cognitive and neurological underpinnings of the observed distinction between confidence and short-long judgments.

Our research utilized explicit interrogation of timing errors (but see Yallak & Balcı, 2022 for evidence for metric error monitoring using implicit tasks) and it is possible that there are also latent signatures of temporal error monitoring that can be gathered from response times and key press pressure. Monitoring the state of skeletal muscles with electromyography (EMG) while performing the temporal errors might also help parse out the potential contribution of the motor preparation in a more granular and parametric fashion.

4.10. Conclusion

This study demonstrates that the critical factor determining the temporal error magnitude monitoring (in the form of subjective confidence) is the subjective belief of agency (i.e., whether the performance belongs to the participant). This factor exerts an effect on temporal error monitoring performance irrespective of motor involvement. Specifically, participants were better at matching their subjective confidence ratings to error magnitude when they had a belief of agency on the errors (i.e., absolute distance from the mean reproductions in standard deviation units) irrespective of the actual source of performance). For error direction monitoring, the actual source of the error appears to be more critical such that the direction of others' errors is precisely classified compared to one's own performance, particularly when there is a motor component to self-performance. Together, this study highlights the two aspects of metric error monitoring that possibly stem from separate read-out mechanisms with differential sensitivities to tacit beliefs and sensorimotor noise. However, these effects are also potentially sensitive to different attentional allocation strategies as observed absence in trial-based manipulation suggests, which should be investigated in future studies.

Chapter 5

5. General Conclusion

Given the documented distinction between perceptual and metric error monitoring (Yallak & Balcı, 2021), investigating whether the factors that affect the perceptual error monitoring ability also affect the metric error monitoring ability is an important question to investigate the cognitive architecture for metacognitive processes in the metric domains. This dissertation addressed this question in the temporal domain in three distinct studies.

Overall, the current dissertation demonstrated that temporal error monitoring ability is seemingly robust to exogenous influences, while it is differentially sensitive to different forms of endogenous influences. The differential sensitivity to endogenous and exogenous factors further highlights the distinctive cognitive architecture of temporal error monitoring abilities with respect to perceptual error monitoring. The main reason for this differential sensitivity to the exogenous factors for the perceptual and temporal error monitoring abilities might be lying in the differential error characteristics such that timing performance (in the currently investigated form) consist of motor components (i.e., reproduction) whereas perceptual performance does not. This being said, the current dissertation cannot posit an ultimate conclusion that temporal error monitoring is fully insensitive to exogenous factors since only a limited number of such factors has been tested. There might be other exogenous factors that are not considered by the current dissertation which could change the temporal error monitoring ability.

The current dissertation, for the first time, demonstrated the magnitude and error directionality components of the temporal error monitoring ability. While this reveals (partially) distinct components of metacognition, the current dissertation cannot directly explain in what aspect these two components exactly differ. For this reason, this distinction is subjected to further investigations at multiple levels including information processing, neural, phenomenological and psychophysical levels to further elucidate their unique characteristics. The analytical and methodological importance of this distinction lies within that these two components need to be accounted for by separated measures and statistical approaches to fully capture their idiosyncratic characteristics with respect to one another. In practice, this distinction carries crucial importance in a way that it could prevent potential increase in error magnitude that might be caused by inability to monitor the magnitude even in the case of correct error direction monitoring. For example, being aware of the error direction but not the magnitude may potentially end up with hypercorrections which may ultimately result in doubled error magnitudes. These insights carry crucial importance to develop potential training programs for improving performance at various sports branches in which keeping track of the time is of critical importance (e.g., swimming, athletics, horseback riding, synchronized swimming etc.).

Taken together, the current dissertation exhibits the differential effects of endogenous and exogenous factors on temporal error monitoring ability such that different types and components are affected from different factors differentially. Finally, the obtained results, so far, validated the idea that temporal error monitoring has a distinctive mechanism that may not be shared with the perceptual error monitoring (Yallak & Balcı, 2020). Together, these results of the current dissertation provide important insights regarding the cognitive architecture of the metric error monitoring ability.

The results presented in the current dissertation may not be directly applicable to the other metric domains that are not covered here. While previous studies demonstrated metacognitive correlation between three metric domains (Yallak & Balcı, 2021), the idiosyncratic features of each metric domain should be taken into account given that the metric commonality is assumed to be partial (e.g., A Theory of Magnitude (ATOM), Walsh, 2006). For this reason, future studies should still explicitly validate the commonality across three main metric domains (i.e., time, space and numerosity) in the metacognitive level as well by replicating the studies that are presented in the current dissertation in spatial and numerosity domains.

While this dissertation has investigated the cognitive architecture of temporal error monitoring ability, its exact neural underpinnings are vastly unknown. There are a number of electroencephalography (EEG) studies which point to precuneus for the potential region of interests (RoI, e.g., Kononowicz et al., 2019; Kononowicz & van Wassenhove, 2019; for a detailed discussion: Öztel & Balcı, 2024). Still, given the limited spatial resolution of EEG and the inability to establish the causal role of an area of interest and temporal error monitoring performance, future studies could benefit from utilizing fMRI and TMS methods, respectively. These methodological improvements could also be utilized to exhibit the possible neural distinction between confidence and short-long judgements.

6. References

Akdoğan, B., & Balcı, F. (2017). Are you early or late?: Temporal error monitoring. *Journal of Experimental Psychology: General, 146*(3), 347.

Al-Salom, P., & Miller, C. J. (2019). The problem with online data collection: Predicting invalid responding in undergraduate samples. *Current Psychology, 38*(5), 1258-1264.

Anzulewicz, A., Hobot , J., Siedlecka, M., & Wierzchoń, M. (2019). Bringing action into the picture. How action influences visual awareness. *Attention, Perception, & Psychophysics*, *81*(7), 2171-2176.

Aydoğan, T., Karşılar, H., Duyan, Y. A., Akdoğan, B., Baccarani, A., Brochard, R., ... & Balcı, F. (2023). The timing database: An open-access, live repository for interval timing studies. *Behavior Research Methods*, 1-11.

Bader, F., & Wiener, M. (2021). Awareness of errors and feedback in human time estimation. *Learning & Memory, 28*(5), 171-177.

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.

Barbey, A. K., Koenigs, M., & Grafman, J. (2013). Dorsolateral prefrontal contributions to human working memory. *Cortex, 49*(5), 1195-1205.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. J*ournal of Memory and Language, 68*(3), 255-278.

Bates, A. T., Patel, T. P., & Liddle, P. F. (2005). External behavior monitoring mirrors internal behavior monitoring: Error-related negativity for observed errors. *Journal of Psychophysiology*, *19*(4), 281-288.

Bausenhart, K. M., Dyjas, O., & Ulrich, R. (2014). Temporal reproductions are influenced by an internal reference: Explaining the Vierordt effect. *Acta Psychologica, 147,* 60–67. http://dx.doi.org/10.1016/j.actpsy.2013 .06.011

Belkacem, A. N., Kiso, K., Uokawa, E., Goto, T., Yorifuji, S., & Hirata, M. (2020). Neural processing mechanism of mental calculation based on cerebral oscillatory changes: a comparison between abacus experts and novices. *Frontiers in Human Neuroscience, 14*, 137.

Bor, D., Duncan, J., Wiseman, R. J., & Owen, A. M. (2003). Encoding strategies dissociate prefrontal activity from working memory demand. *Neuron, 37*(2), 361-367.

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. (2020). The timing mega-study: comparing a range of experiment generators, both lab-based and online. https://doi.org/10.7717/peerj.9414

Brocas, I., Carrillo, J. D., & Tarrasó, J. (2018). Self-awareness of biases in time perception. *Journal of Economic Behavior & Organization, 148*, 1-19.

Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms in F. E. Weinert & R. H. Kluwe, (Eds.) *Metacognition, motivation, and understanding. 65-116*. Hillsdale, NJ: Lawrence Erlbaum

Brühlmann, F., Petralito, S., Aeschbach, L. F., & Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology, 2*, 100022.

Bryce, D., & Bratzke, D. (2014). Introspective reports of reaction times in dual-tasks reflect experienced difficulty rather than timing of cognitive processes. *Consciousness and Cognition*, *27*, 254-267.

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.

Burns, G. N., Christiansen, N. D., Morris, M. B., Periard, D. A., & Coaster, J. A. (2014). Effects of applicant personality on resume evaluations. *Journal of Business and Psychology*, *29*, 573-591.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science, 280*(5364), 747-749.

Charles, L., Chardin, C., & Haggard, P. (2020). Evidence for metacognitive bias in perception of voluntary action. *Cognition*, *194*, 104041.

Church, R. M. (1984). Properties of the internal clock. *Annals of the New York Academy of sciences*.

Cochin, S., Barthelemy, C., Roux, S., & Martineau, J. (1999). Observation and execution of movement: similarities demonstrated by quantified electroencephalography. *European Journal of Neuroscience*, *11*(5), 1839-1842.

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., & Smith, E. E. (1997). Temporal dynamics of brain activation during a working memory task. *Nature, 386*(6625), 604-608.

Coles, M. G., Scheffers, M. K., & Holroyd, C. B. (2001). Why is there an ERN/Ne on correct trials? Response representations, stimulus-related components, and the theory of error-processing. *Biological Psychology, 56*(3), 173-189.

Coutinho, M. V., Redford, J. S., Church, B. A., Zakrzewski, A. C., Couchman, J. J., & Smith, J. D. (2015). The interplay between uncertainty monitoring and working memory: Can metacognition become automatic?. *Memory & Cognition, 43*(7), 990-1006.

Coutinho, M. V., Redford, J. S., Church, B. A., Zakrzewski, A. C., Couchman, J. J., & Smith, J. D. (2015). The interplay between uncertainty monitoring and working memory: Can metacognition become automatic?. *Memory & Cognition, 43*(7), 990-1006.

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences, 7*(9), 415-423.

D'esposito, M., Detre, J. A., Alsop, D. C., Shin, R. K., Atlas, S., & Grossman, M. (1995). The neural basis of the central executive system of working memory. *Nature, 378*(6554), 279-281.

D'Esposito, M., Postle, B. R., Ballard, D., & Lease, J. (1999). Maintenance versus manipulation of information held in working memory: an event-related fMRI study. *Brain and Cognition,* 41(1), 66-86.

Davis, S. W., Crowell, C. A., Beynel, L., Deng, L., Lakhlani, D., Hilbig, S. A., ... & Cabeza, R. (2018). Complementary topology of maintenance and manipulation brain networks in working memory. *Scientific Reports, 8*(1), 1-14.

De Loof, E., Poppe, L., Cleeremans, A., Gevers, W., & Van Opstal, F. (2015). Different effects of executive and visuospatial working memory on visual consciousness. *Attention, Perception, & Psychophysics, 77*(8), 2523-2528.

De Loof, E., Verguts, T., Fias, W., & Van Opstal, F. (2013). Opposite effects of working memory on subjective visibility and priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(6), 1959.

Doenyas, C. , Mutluer, T., Genç, E. & Balcı, F. (2019). Error Monitoring in Decision-Making and Timing is Disrupted in Autism Spectrum Disorder. *Autism Research. 12*(2): 239-248. doi: 10.1002/aur.2041.

Droit-Volet, S., & Izaute, M. (2009). Improving time discrimination in children and adults in a temporal bisection task: The effects of feedback and no forced choice on decision and memory processes. *Quarterly Journal of Experimental Psychology, 62*(6), 1173-1188.

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.

D'Esposito, M., Postle, B. R., & Rypma, B. (2000). Prefrontal cortical contributions to working memory: evidence from event-related fMRI studies. *Executive control and the frontal lobe: Current issues*, 3-11.

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke L. (1990). Effects of errors in choice reaction tasks on the ERP under focused and divided attention. In CHM Brunia, AWK Gaillard and A Kok (Eds), *Psychophysiological Brain Research*. Tilburg, The Netherlands: Tilburg University Press, pp. 192-195.

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.

Filevich, E., Koß, C., & Faivre, N. (2020). Response-related signals increase confidence but not metacognitive performance. *Eneuro*, *7*(3).

Fitzgerald, K. D., Welsh, R. C., Gehring, W. J., Abelson, J. L., Himle, J. A., Liberzon, I., & Taylor, S. F. (2005). Error-related hyperactivity of the anterior cingulate cortex in obsessive-compulsive disorder. *Biological Psychiatry*, *57*(3), 287-294.

Flavell, J. H., & Wellman, H. M. (1975). Metamemory.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34,* 906–911

Fleming, S. M., & Dolan, R. J. (2012b). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1338-1349.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443.

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012a). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society B*, 367, 1280–1286 doi:10.1098/rstb.2012.0021

Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *Journal of Neuroscience, 32*(18), 6117-6125.

Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. (2015). Action-specific disruption of perceptual confidence. *Psychological Science, 26*(1), 89-98.

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain, 137*(10), 2811-2822.

Frith, C., & Dolan, R. (1996). The role of the prefrontal cortex in higher cognitive functions. *Cognitive Brain Research, 5*(1-2), 175-181.

Gagné, N., & Franzen, L. (2021). How to run behavioural experiments online: best practice suggestions for cognitive psychology and neuroscience.

Gajdos, T., Fleming, S. M., Saez Garcia, M., Weindel, G., & Davranche, K. (2019). Revealing subthreshold motor contributions to perceptual confidence. *Neuroscience of Consciousness, 2019(*1), niz001.

Gayet, S., van Maanen, L., Heilbron, M., Paffen, C. L., & Van der Stigchel, S. (2016). Visual input that matches the content of visual working memory requires less (not faster) evidence sampling to reach conscious access. *Journal of Vision, 16*(11), 26-26.

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.

Gherman, S., & Philiastides, M. G. (2015). Neural representations of confidence emerge from the process of decision formation during perceptual choices. *Neuroimage*, *106*, 134-143.

Gibbon, J., Church, R. M., & Meck, W. H. (1984). Scalar timing in memory. Annals of the New York Academy of sciences, 423(1), 52-77.

Gibbon, J., Fairhurst, S., & Goldberg, B. (1997). Cooperation, conflict and compromise between circadian and interval clocks in pigeons. In Advances in psychology (Vol. 120, pp. 329-384). North-Holland.

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psychological Review*, *84*(3), 279.

Gorn, G. J., Chattopadhyay, A., Sengupta, J., & Tripathi, S. (2004). Waiting for the web: how screen color affects time perception. *Journal of Marketing Research, 41*(2), 215-225.

Gu, B. M., & Meck, W. H. (2011). New perspectives on Vierordt's Law: Memory-mixing in ordinal temporal comparison tasks. In A. Vatakis, A. Esposito, M. Giagkou, F. Cummins, & G. Papadelis (Eds.), Multidisciplinary aspects of time and time perception (pp. 67–78). Berlin, Germany: Springer-Verlag. http://dx.doi.org/10.1007/978-3-642-21478-3_6

Hardre, P. L., Crowson, H. M., & Xie, K. (2012). Examining Contexts-of-Use for Web-Based and PaperBased Questionnaires. *Educational and Psychological Measurement, 72*(6), 1015–1038. doi: 10.1177/0013164412451977

Harty, S., Robertson, I. H., Miniussi, C., Sheehy, O. C., Devine, C. A., McCreery, S., & O'Connell, R. G. (2014). Transcranial direct current stimulation over right dorsolateral prefrontal cortex enhances error awareness in older age. *Journal of Neuroscience, 34*(10), 3646-3652.

Iskander, M. (2019). Burnout, cognitive overload, and metacognition in medicine. *Medical Science Educator, 29*(1), 325-328.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience, 13,* 1020–1026. http://dx.doi.org/10 .1038/nn.2590

Jiang, S., Wang, S., & Wan, X. (2022). Metacognition and mentalizing are associated with distinct neural representations of decision uncertainty. *PLoS Biology*, *20*(5), e3001301.

Kang, S. K., Hirsh, J. B., & Chasteen, A. L. (2010). Your mistakes are mine: Self-other overlap predicts neural response to observed errors. *Journal of Experimental Social Psychology*, *46*(1), 229-232.

Kepecs, A., Uchida, N., Zariwala, H., Mainen, Z., 2008. Neural correlates, computationand behavioural impact of decision confidence. *Nature* 455, 227–231.

Kononowicz, T. W., & van Wassenhove, V. (2019b). Evaluation of Self-generated behavior: untangling metacognitive readout and error detection. *Journal of cognitive neuroscience, 31*(11), 1641-1657.

Kononowicz, T. W., Roger, C., & van Wassenhove, V. (2019). Temporal metacognition as the decoding of self-generated brain dynamics. *Cerebral Cortex, 29*(10), 4366-4380.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences 15*(8): 365–73.

Lau, H. (2019). Consciousness, metacognition, & perceptual reality monitoring. psyArxiv

Lewis, P. A., & Miall, R. C. (2009). The precision of temporal judgement: milliseconds, many minutes, and beyond. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1525), 1897-1905.

Luber, B., Kinnunen, L. H., Rakitin, B. C., Ellsasser, R., Stern, Y., & Lisanby, S. H. (2007). Facilitation of performance in a working memory task with rTMS stimulation of the precuneus: frequency-and time-dependent effects. *Brain Research, 1128*, 120-129.

Magno, E., Foxe, J. J., Molholm, S., Robertson, I. H., & Garavan, H. (2006). The anterior cingulate and error avoidance. *Journal of Neuroscience, 26*(18), 4769-4773.

Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: meta-d′, response-specific meta-d′, and the unequal variance SDT model. *The cognitive neuroscience of metacognition, 25*-66.

Maniscalco, B., & Lau, H. (2015). Manipulation of working memory contents selectively impairs metacognitive sensitivity in a concurrent visual discrimination task. *Neuroscience of Consciousness, (1).*

Marr, D. C. (1982). Vision: A computation investigation into the human representational system and processing of visual information. *San Francisco: Freeman*

McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., De Lange, F. P., & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *Journal of Neuroscience, 33*(5), 1897-1906.

Meyniel, F., & Dehaene, S. (2017). Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proceedings of the National Academy of Sciences*, *114*(19), E3859-E3868.

Miltner, W. H., Brauer, J., Hecht, H., Trippe, R., & Coles, M. G. (2004). Parallel brain activity for self-generated and observed errors.

Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *Journal of Neuroscience, 38*(14), 3534-3546.

Oliveri, M., Vicario, C. M., Salerno, S., Koch, G., Turriziani, P., Mangano, R., ... & Caltagirone, C. (2008). Perceiving numbers alters time perception. *Neuroscience letters, 438*(3), 308-311.

Overbeek, T. J., Nieuwenhuis, S., & Ridderinkhof, K. R. (2005). Dissociable components of error processing: On the functional significance of the Pe vis-à-vis the ERN/Ne. *Journal of Psychophysiology, 19*(4), 319-329.

Öztel, T., & Balcı, F. (2020). Stimulus-Induced Temporal Illusions: When Awareness Is Mesmerized by Time. *Timing & Time Perception, 8*(3-4), 363-380.

Öztel, T., & Balcı, F. (2023). Metric error monitoring as a component of metacognitive processing. *European Journal of Neuroscience.*

Öztel, T., Eskenazi, T., & Balcı, F. (2020). Temporal error monitoring with directional error magnitude judgements: A robust phenomenon with no effect of being watched. *Psychological Research,* 1-10. DOI: 10.1007/s00426-020-01379-0

Palser, E. R., Fotopoulou, A., & Kilner, J. M. (2018). Altering movement parameters disrupts metacognitive accuracy. *Consciousness and Cognition*, *57*, 33-40.

Patel, D., Fleming, S. M., & Kilner, J. M. (2012). Inferring subjective states through the observation of actions. *Proceedings of the Royal Society B: Biological Sciences, 279*(1748), 4853-4860.

Pavlovia. https://pavlovia.org/ (accessed on 24 March 2020).

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*. 10.3758/s13428-018-01193-y

Petrides, M. (2000). The role of the mid-dorsolateral prefrontal cortex in working memory. *Experimental Brain Research, 133*(1), 44-54.

Pfabigan, D. M., Zeiler, M., Lamm, C., & Sailer, U. (2014). Blocked versus randomized presentation modes differentially modulate feedback-related negativity and P3b amplitudes. *Clinical Neurophysiology*, *125*(4), 715-726.

Postle, B. R., Berger, J. S., & D'Esposito, M. (1999). Functional neuroanatomical double dissociation of mnemonic and executive control processes contributing to working memory performance. *Proceedings of the National Academy of Sciences, 96*(22), 12959-12964.

Postle, B. R., Ferrarelli, F., Hamidi, M., Feredoes, E., Massimini, M., Peterson, M., ... & Tononi, G. (2006). Repetitive transcranial magnetic stimulation dissociates working memory manipulation from retention functions in the prefrontal, but not posterior parietal, cortex. *Journal of Cognitive Neuroscience, 18*(10), 1712-1722.

Rahnev, D., Desender, K., Lee, A. L., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., ... & Zylberberg, A. (2020). The confidence database. *Nature Human Behaviour*, *4*(3), 317-325.

Rahnev, D., Nee, D. E., Riddle, J., Larson, A. S., & D'Esposito, M. (2016). Causal evidence for frontal cortex organization for perceptual decision making. *Proceedings of the National Academy of Sciences, 113*(21), 6059-6064.

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.

Riemer, M., Kubik, V., & Wolbers, T. (2019). The effect of feedback on temporal error monitoring and timing behavior. *Behavioural Brain Research, 369*, 111929.

Rouault, M., Dayan, P., & Fleming, S. M. (2019). Forming global estimates of self-performance from local confidence. *Nature Communications, 10*(1), 1-11. https://doi.org/10.1038/s41467-019-09075-3

Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., & Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience, 1*(3), 165-175.

Rowe, J. B., & Passingham, R. E. (2001). Working memory for location and time: activity in prefrontal area 46 relates to selection rather than maintenance in memory. *Neuroimage, 14*(1), 77-86.

Rowe, J. B., Toni, I., Josephs, O., Frackowiak, R. S., & Passingham, R. E. (2000). The prefrontal cortex: response selection or maintenance within working memory?. *Science, 288*(5471), 1656-1660.

Rypma, B., Berger, J. S., & D'esposito, M. (2002). The influence of working-memory demand and subject performance on prefrontal cortical activity. *Journal of Cognitive Neuroscience, 14*(5), 721-731.

Rypma, B., Prabhakaran, V., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. (1999). Load-dependent roles of frontal brain regions in the maintenance of working memory. *Neuroimage, 9*(2), 216-226.

Sauter, M., Draschkow, D., & Mack, W. (2020). Building, hosting and recruiting: A brief introduction to running behavioral experiments online. *Brain Sciences, 10*(4), 251.

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education, 36*(1), 111-139.

Shane, M. S., Stevens, M., Harenski, C. L., & Kiehl, K. A. (2008). Neural correlates of the processing of another's mistakes: a possible underpinning for social and observational learning. *Neuroimage*, *42*(1), 450-459.

Shekhar, M., & Rahnev, D. (2018). Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition. *Journal of Neuroscience*, *38*(22), 5078-5087.

Siedlecka, M., Paulewicz, B., & Koculak, M. (2020). Task-related motor response inflates confidence. *BiorXiv*.

Siedlecka, M., Paulewicz, B., & Wierzchoń, M. (2016). But I was so sure! Metacognitive judgments are less accurate given prospectively than retrospectively. *Frontiers in Psychology*, *7*, 218.

Smets, G. (1969). Time expression of red and blue. *Perceptual and Motor Skills, 29*(2), 511-514.

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.

Thönes, S., Von Castell, C., Iflinger, J., & Oberfeld, D. (2018). Color and time perception: Evidence for temporal overestimation of blue stimuli. *Scientific Reports, 8*(1), 1-8.

Tullis, J. G., & Fraundorf, S. H. (2017). Predicting others' memory performance: The accuracy and bases of social metacognition. *Journal of Memory and Language, 95*, 124-137.

Ullsperger, M., & Von Cramon, D. Y. (2004). Neuroimaging of performance monitoring: error detection and beyond. *Cortex, 40*(4-5), 593-604.

Vandenbroucke, A. R., Sligte, I. G., Barrett, A. B., Seth, A. K., Fahrenfort, J. J., & Lamme, V. A. (2014). Accurate metacognition for visual sensory memory representations. *Psychological Science*, *25*(4), 861-873.

van Schie, H. T., Mars, R. B., Coles, M. G., & Bekkering, H. (2004). Modulation of activity in medial frontal and motor cortices during error observation. *Nature Neuroscience, 7*(5), 549-554.

Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*(11), 483-488.

Walton, M. E., Devlin, J. T., & Rushworth, M. F. (2004). Interactions between decision making and performance monitoring within prefrontal cortex. *Nature Neuroscience*, *7*(11), 1259-1265.

Wang, J., Tian, J., Hao, R., Tian, L., & Liu, Q. (2018). Transcranial direct current stimulation over the right DLPFC selectively modulates subprocesses in working memory. *PeerJ, 6*, e4906.

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1990). What influences learning? A content analysis of review literature. *Journal of Educational Research, 84*, 30–43.

Wokke, M. E., Achoui, D., & Cleeremans, A. (2020). Action information contributes to metacognitive decision-making. *Scientific Reports, 10*(1), 1-15.

Yallak, E., & Balcı, F. (2021). Metric error monitoring: Another generalized mechanism for magnitude representations?. *Cognition*, *210*, 104532.

Yallak, E., & Balcı, F. (2022). Metric error monitoring for a cleaner record of timing. *Journal of Experimental Psychology: Human Perception and Performance, 48*(10), 1130.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310-1321.

Yeung, N., & Summerfield, C. (2014). Shared mechanisms for confidence judgements and error detection in human decision making. In *The cognitive neuroscience of metacognition* (pp. 147-167). Springer, Berlin, Heidelberg.

Yomogida, Y., Sugiura, M., Sassa, Y., Wakusawa, K., Sekiguchi, A., Fukushima, A., ... & Kawashima, R. (2010). The neural basis of agency: an fMRI study. *Neuroimage, 50*(1), 198-207.

Yoo, A. H., Acerbi, L., & Ma, W. J. (2020). Uncertainty is Maintained and Used in Working Memory. *bioRxiv*.

Yoshida, K., Saito, N., Iriki, A., & Isoda, M. (2012). Social error monitoring in macaque frontal cortex. *Nature Neuroscience, 15*(9), 1307.

Zakay, D., & Block, R. A. (1995). An attentional-gate model of prospective time estimation. *Time and the Dynamic Control of Behavior, 5*, 167-178.

8. Appendix

8.1. Supplementary Online Materials (Chapter 2)

Descriptives

Variable	N	Mean	Median	SD	SE
Confidence in Low Load	48	2.1102	2.0913	0.4267	0.0616
Confidence in High Load	48	2.0856	2.0806	0.4424	0.0639
Confidence in High Load-Easy	48	2.0882	2.0763	0.4480	0.0647
Confidence in High Load-Hard	48	2.0826	2.0697	0.4397	0.0635
Reproductions in Low Load	48	2.69407	2.66223	0.30598	0.04416
Reproductions in High Load	48	2.72542	2.73407	0.29841	0.04307
Reproductions in High Load-Easy	48	2.72432	2.73910	0.30474	0.04399
Reproductions in High Load-Hard	48	2.72860	2.76795	0.30165	0.04354
CV in Low Load	48	0.24972	0.24965	0.07073	0.01021
CV in High Load	48	0.24656	0.25084	0.06810	0.00983
CV in High Load-Easy	48	0.24728	0.25786	0.06875	0.00992
CV in High Load-Hard	48	0.24459	0.23183	0.07101	0.01025

Table 1. Descriptive statistics of confidence, reproductions and coefficient of variation (CV) values of reproductions across different working memory load conditions.

8.2. Supplementary Online Materials 1 (Chapter 4)

8.2.1. Comparison between Experiment 1 and Experiment 2. We investigated whether the reproductions in the observed condition differ across Experiment 1 and Experiment 2. To this end, we used the target normalized reproductions (i.e., reproduction / (divided by) target duration) as the dependent variable and experiment number (i.e., Experiment 1 or 2) and target duration as independent variables. We included the participants as random intercept. Linear mixed effects model revealed a significant interaction between the experiment number and target duration ($M_{\text{Experiment 1- Experiment 2 * long-short}} = 0.23$, $SE = 0.014$, $p < 0.001$). There was also a significant main effect of experiment number ($M_{\text{Experiment 1- Experiment 2}} = -0.12$, $SE = 0.0296$, $p < 0.001$) and target duration ($M_{\text{long-short}} = -0.089$, $SE = 0.007$, $p < 0.001$), which reveals that the reproductions in the observation condition in two experiments were significantly different.

8.3. Experiment 1

8.3.1. Confidence judgment. There was a correct overall match between the absolute timing errors to the confidence judgments (ß = -0.12, $SE = 0.02$, $p < 0.001$; intercept = 2.21, $SE = 0.06$, $p < 0.001$), pointing to a temporal error monitoring ability. The main effect of the experimental condition on confidence judgments was also significant ($M_{\text{self reproduction-observation}} = 0.19$, $SE = 0.03$, $p < 0.001$). This result demonstrates an overall higher confidence in the participants' timing reproductions. The interaction between the absolute timing errors and the experimental condition was also significant, revealing a steeper negative slope for the self-reproduction condition than the observation condition ($ß_{\text{self reproduction-observation}} = -0.09$, $SE = 0.03$, $p = 0.006$).

8.3.2. Short-long judgment. Logistic mixed-effects showed that the probability of reporting "long" reproduction increased with increasing signed timing errors (ß = 0.89, exp(ß) = 2.44, $SE = 0.04$, $p < 0.001$; intercept = 0.062, exp(ß) = 1.064, $SE = 0.11$, $p = 0.58$). There was also a general bias towards reporting "long" regardless of the target duration when the reproduction was performed by the participants (self-reproduction) compared to observing others' reproduction ($M_{\text{self reproduction-observation}} = 0.15$, exp(ß) = 1.67, $SE = 0.06$, $p = 0.014$).

8.4. Experiment 2

8.4.1. Comparison of Coefficient of Variation (CV) of the Reproduced Durations Across Conditions. We first investigated whether the participants' timing performances matched the durations in the observation condition. To this end, we used the coefficient of variation (CV) to index the participants' timing precision (i.e., lower values indicate higher precision) and compared it with the observation condition. Mixed ANOVA results revealed no significant interaction between target duration (i.e., short or long) and experimental condition (i.e., self reproduction or observation) on the CV of the reproductions ($F(1,40) = 0.04$, $p = 0.84$). This null effect also applied to the main effect of target duration ($F(1,40) = 0.01$, $p = 0.91$) and self-reproduction condition ($F(1,40) = 1.71$, $p = 0.2$). Together, these findings demonstrate no difference across the experimental and target duration conditions regarding their mean CV values (i.e., scalar property; Gibbon et al., 1997), thus pointing to an overall match across conditions regarding timing performance. Note that this analysis could not occur in Experiment 1 as the reproduced durations used in the observation condition were obtained in a between-subject design (i.e., as gathered from Akdoğan & Balcı, 2017).

8.4.2. Confidence Judgments. Linear mixed-effects model demonstrated a significant overall temporal error monitoring ability (i.e., negative relationship between absolute timing error and confidence judgments; ß = -0.10, $SE = 0.02$; $p < 0.001$; intercept: 2.14, $SE = 0.07$, $p < 0.001$), target duration ($M_{\text{short-long}} = 0.10$, $SE = 0.03$; $p = 0.001$) which reveals an overall higher confidence judgment for short compared to the long target duration and experimental condition ($M_{\text{self reproduction-observation}} = 0.15$, $SE = 0.03$; $p < 0.001$) pointing to overall higher confidence when the self performs the reproduction. Moreover, the interaction between the target duration and experimental condition on confidence judgment ($M_{\text{short-long*self reproduction - observation}} = 0.15$, $SE = 0.06$; $p = 0.013$) showed that the mean confidence judgment difference across target durations was larger when the participant themselves made the reproduction. Furthermore, no interaction was observed between the target duration and absolute timing errors ($p = 0.4$). Finally, the three-way interaction was not statistically significant ($p = 0.31$).

8.4.3. Short-Long Judgments. There was a significant positive relationship between the signed timing error and short-long judgments (ß = 0.97, exp(ß) = 2.63, $SE = 0.04$; $p < 0.001$; intercept =

0.07, exp(ß) = 1.07, $SE = 0.10$, $p = 0.49$), revealing an increased probability of reporting "long" as the signed timing error increased. The main effect of the experimental condition was also statistically significant ($M_{\text{self reproduction-observation}} = -0.35$, exp(ß) = 0.71, $SE = 0.06$; $p < 0.001$), which points to a "short" response bias for the reproductions performed by the participants themselves. This finding contradicts Experiment 1, which pointed to a reverse effect. The interaction between target duration and experimental condition was statistically significant ($M_{\text{short-long*self reproduction-observation}} = -1.05$, exp(ß) = 0.35, $SE = 0.13$; $p < 0.001$). However, the interaction between target duration and signed timing error was not statistically significant ($p = 0.49$). Similarly, the three-way interaction between target duration, experimental condition, and signed timing error on the short-long judgment was not statistically significant ($p = 0.55$).

8.5. Experiment 3

8.5.1. Confidence Judgments. We investigated the temporal error monitoring ability for confidence judgments as in the first two experiments. There was a significant negative relationship between the absolute timing error and the confidence judgments (ß = -0.07, $SE = 0.023$, $p = 0.002$; intercept = 2.28, $SE = 0.07$, $p < 0.001$). Moreover, there was a significant main effect of the experimental condition on the confidence judgments ($M_{\text{self reproduction-mimicked reproduction}} = 0.26$, $SE = 0.04$, $p < 0.001$), pointing to higher confidence for the temporal errors that belong to the participants themselves compared to the mimicked ones.

8.5.2. Short-long judgements. We investigated the relationship between the signed temporal errors and the short-long judgments in isolation from the confidence judgments. Logistic linear mixed effect analysis demonstrated a significant positive relationship between the signed temporal errors and the short-long judgments (ß = 0.866, exp(ß) = 2.376; $SE = 0.053$, $p < 0.001$; intercept = 0.27, exp(ß) = 1.31, $SE = 0.14$, $p = 0.06$). This result demonstrates an overall accurate temporal error classification regardless of the experimental conditions. Moreover, mimicked reproductions were classified to be longer than the reproductions that are performed by the participants ($M_{\text{self reproduction-mimicked reproduction}} = -0.411$, exp(ß) = 0.663; $SE = 0.083$, $p < 0.001$).

8.6. Experiment 4

8.6.1. Confidence judgments. As in all previous experiments, the linear mixed effects model revealed a negative relationship between the absolute timing error and confidence ratings (ß = -0.18, *SE* = 0.03, $p < 0.001$; intercept = 2.15, *SE* = 0.05, $p < 0.001$), demonstrating an overall temporal error monitoring ability regardless of the experimental condition. However, there was no main effect of the experimental condition on confidence judgments ($p = 0.61$), which demonstrated no biasing effect of the experimental manipulation on the mean confidence level.

8.6.2. Short-long judgements. For the short-long judgments, the generalized mixed effects model revealed only a significant positive relationship between signed timing error and short-long judgments (ß = 1.60, *SE* = 0.08, exp(ß) = 4.94, $p < 0.001$; intercept = 0.42, *SE* = 0.13, exp(ß) = 1.52, $p < 0.001$), which demonstrates an overall correct short-long classification of the observed durations.

8.7. Experiment 5

8.7.1. Confidence judgments. Linear mixed effects model showed a significant negative relationship between absolute timing error and confidence judgments (ß = -0.17, *SE* = 0.02, $p < 0.001$; intercept = 2.08, *SE* = 0.07, $p < 0.001$), revealing an overall temporal error monitoring ability as reflected in the correct match between temporal error magnitude and corresponding confidence. Furthermore, there was a main effect of instruction type on confidence judgments ($M_{\text{belief of non agency-belief of agency}} = -0.10$, *SE* = 0.03, $p = 0.0017$), which points to an overall higher confidence rating when participants believed that the reproduction belonged to themselves (i.e., belief of agency) compared to another participant (i.e., belief of non agency). The remaining effects were not statistically significant (all *p*s > 0.05).

8.7.2. Short-long judgments. The generalized mixed model revealed a significant positive relationship between signed timing error and the short-long judgments (ß = 1.25, *SE* = 0.05, exp(ß) = 3.48, $p < 0.001$; intercept = 0.23, exp(ß) = 1.26, *SE* = 0.13, $p = 0.07$), revealing an overall temporal error monitoring ability. The main effect of the source of data was also statistically significant ($M_{\text{simulated reproduction-empirical reproduction}} = 0.14$, *SE* = 0.07, exp(ß) = 1.15, $p = 0.03$), which demonstrates an overall tendency to classify the observed reproductions that are simulated as "long" with higher probability. Moreover, there was a significant interaction between the source of data and instruction type on short-long judgments ($M_{\text{simulated reproduction-empirical reproduction* belief of non}}$

$_{\text{agency- belief of agency}}$ = -0.37, *SE* = 0.13, exp(ß) = 0.69, $p < 0.01$). Bonferroni corrected post hoc comparisons revealed a significant difference between the agency-congruent (i.e., both the reproduction belonged to the participant and the instruction was "YOUR") and non agency-incongruent (i.e., the reproduction did not belong to the participant but the instruction was "YOUR") in terms of "long" classifications ($\exp(ß)_{\text{empirical reproduction\&belief of agency - simulated reproduction\&belief of agency}} = 0.72$; *SE* = 0.067, $p_{\text{bonferroni}} = 0.003$), suggesting that probability of responding "long" was higher for the empirical reproduction compared to simulated reproduction when participants were simply instructed to have agency.

8.8. Experiment 6

8.8.1.1. Confidence Judgments. *Online experiment.* As in the first three experiments, we investigated the relationship between absolute timing error and confidence ratings across the experimental conditions. Linear mixed effects model revealed a significant negative relationship between absolute timing error and confidence ratings (ß = -0.33, *SE* = 0.026, $p < 0.001$; intercept = 2.29, *SE* = 0.065, $p < 0.001$), which points to an overall temporal error monitoring ability regardless of the experimental condition.

8.8.1.2. Lab experiment. The linear mixed-effects model revealed a significant negative relationship between the absolute timing errors and confidence ratings (ß = -0.21, *SE* = 0.03, $p < 0.001$; intercept = 2.09, *SE* = 0.05, $p < 0.001$), pointing to an overall temporal error monitoring ability as reflected in correct error magnitude reports. Unlike the online experiment, there was a significant main effect of the experimental condition ($M_{\text{belief of agency - belief of non agency}} = 0.14$, *SE* = 0.04, $p = 0.002$), revealing higher confidence ratings in the belief of agency condition. This result is in line with the first three experiments that were conducted as online experiments.

8.8.2.1. Short-long Judgments. *Online experiment.* We investigated the relationship between the signed timing errors and short-long judgments across the belief conditions. As in the case of confidence ratings, the linear mixed model revealed only the significant relationship between signed timing error and short-long judgments (ß = 1.42, *SE* = 0.08, exp(ß) = 1.20, $p < 0.001$; intercept = 0.67, exp(ß) = 1.95, *SE* = 0.18, $p < 0.001$). Moreover, belief conditions had no significant main effect on long judgment probability (all *p*s > 0.05).

9.7.2.2.Lab experiment. The generalized mixed model revealed only the significant relationship between signed timing error and short-long judgments (ß = 1.87, *SE* = 0.09, exp(ß) = 6.50, $p < 0.001$; intercept = 0.93, exp(ß) = 2.52, *SE* = 0.21, $p < 0.001$).

8.9.1. Combined Analyses

To test the robustness of the obtained interaction terms (Brysbaert, 2019[1]), we combined the data associated with experiments with similar methodological approaches and re-ran the linear mixed effects models.

8.9.1.1. Experiment 1 and 2 combined: Confidence Judgments. As a result of combining the data associated with the first two experiments, linear mixed effects model revealed an overall main effect of the absolute timing error (ß = -0.107, *SE* = 0.012, $p < 0.001$; intercept = 2.18, *SE* = 0.047, $p < 0.001$), main effect of experimental condition ($M_{\text{self reproduction - observed}} = 0.17$, *SE* = 0.021, $p < 0.001$), main effect of target duration ($M_{\text{short - long}} = 0.081$, *SE* = 0.021, $p < 0.001$) on confidence judgments. Moreover, the interaction between the absolute timing errors and the experimental condition was statistically significant ($ß_{\text{self reproduction - observed}} = -0.143$, *SE* = 0.024, $p < 0.001$), demonstrating a steeper negative slope for self condition compared to the other condition. Critically, this interaction did not change across target durations ($ß_{\text{self reproduction - observed * short-long}} = 0.047$, *SE* = 0.048, $p = 0.033$). While all the simple slopes for the self condition was statistically significant in the negative direction ($ß_{\text{self reproduction, short}} = -0.16$, *SE* = 0.025, $p < 0.001$; $ß_{\text{self reproduction, long}} = -0.192$, *SE* = 0.024, $p < 0.001$), this was not the case for the simple slopes for the other condition ($ß_{\text{observed, short}} = -0.045$, *SE* = 0.024, $p = 0.064$; $ß_{\text{observed, long}} = -0.026$, *SE* = 0.024, $p = 0.28$).

8.9.1.2. Experiment 1 and 2 combined: Short-long Judgments. The combination of the data from the first two experiments yielded an overall significant main effect of signed timing errors on short-long judgments (ß = 0.91, *SE* = 0.029, exp(ß) = 2.48, $p < 0.001$; intercept = 0.061, *SE* = 0.074, exp(ß) = 1.06, $p = 0.41$). The main effect of the experimental condition on short-long judgments was also statistically significant ($M_{\text{self reproduction - observed}} = -0.091$, *SE* = 0.044, exp(ß) = 0.91, $p = 0.037$). However, the main effect of target duration on short-long judgments was not

[1] Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition, 2*(1).

significant ($M_{short\text{-}long}$ = 0.023, *SE* = 0.044, exp(ß) = 1.023, *p* = 0.598). There was a significant interaction between the signed timing error and experimental conditions ($ß_{self\ reproduction\ -\ observed}$ = -0.254, *SE* = 0.057, exp(ß) = 0.78, $p < 0.001$), pointing to a steeper slope for other condition. Similarly, the interaction between the experimental condition and target duration was statistically significant ($M_{self\ reproduction\ -\ observed*short\text{-}long}$ = -0.51, *SE* = 0.088, exp(ß) = 0.6, $p < 0.001$). However, the three-way interaction between target duration, experimental condition and the signed timing error was not statistically significant ($ß_{self\ reproduction\ -\ observed*short\text{-}long}$ = -0.12, *SE* = 0.11, exp(ß) = 0.89, *p* = 0.31). All simple slopes were statistically significant ($ß_{self\ reproduction,\ short}$ = 0.76, *SE* = 0.055, exp(ß) = 2.14, $p < 0.\ 001$; $ß_{self\ reproduction,\ long}$ = 0.80, *SE* = 0.055, exp(ß) = 2.23, $p < 0.001$; $ß_{observed,\ short}$ = 1.07, *SE* = 0.061, exp(ß) = 2.92, $p < 0.001$; $ß_{observed,\ long}$ = 0.999, *SE* = 0.057, exp(ß) = 2.72, $p < 0.001$).

8.9.1.3. Experiment 4 and 6 combined: Confidence Judgments. The combined model revealed a significant main effect of absolute timing error (ß = -0.24, *SE* = 0.02, p < 0.001; intercept = 2.17, *SE* = 0.33, $p < 0.001$) and experimental condition on confidence judgments ($M_{self\ reproduction\ -\ observed}$ = 0.067, *SE* = 0.026, *p* = 0.009), revealing higher confidence for the self made errors. However, the interaction between the absolute timing errors and the experimental condition was not statistically significant ($ß_{self\ reproduction\ -\ observed}$ = -0.045, *SE* = 0.0301, *p* = 0.14). The simple slopes were statistically significant in the negative direction for both experimental conditions ($ß_{self\ reproduction}$ = -0.26, *SE* = 0.022, $p < 0.001$; $ß_{observed}$ = -0.21, *SE* = 0.021, $p < 0.001$).

8.9.1.4. Experiment 4 and 6 combined: Short-long Judgments. The combined generalized linear mixed effects model revealed a significant main effect of signed timing errors on short-long judgments (ß = 1.62, *SE* = 0.047, exp(ß) = 5.056, $p < 0.001$; intercept = 0.66, *SE* = 0.10, exp(ß) = 1.94, $p < 0.001$). However, neither the main effect of experimental condition ($M_{self\ reproduction\ -\ observed}$ = 0.09, *SE* = 0.058, exp(ß) = 1.094, *p* = 0.12) nor the interaction between the experimental condition and signed timing error was statistically significant ($ß_{self\ reproduction\ -\ observed}$ = -0.025, *SE* = 0.090, exp(ß) = 0.98, *p* = 0.778). Simple slopes revealed significant positive slope for both experimental conditions ($ß_{self\ reproduction}$ = 1.61, *SE* = 0.065, exp(ß) = 4.99, $p < 0.001$; $ß_{observed}$ = 1.63, *SE* = 0.064, exp(ß) = 5.12, $p < 0.001$).

8.10. Descriptive Statistics

Table S1.1.

Descriptive statistics for reproductions for all experiments across conditions

Experiment	**Target duration**	**Experimental condition**	**Mean**	**SD**
Experiment 1	Short (1.5 secs)	Self reproduction	1.76579	0.50055
		Observation	1.41754	0.30136
	Long (3.0 secs)	Self reproduction	2.91717	0.77155
		Observation	2.91506	0.62318
Experiment 2	Short (1.5 secs)	Self reproduction	1.69227	0.41382
	Long (3.0 secs)	Self reproduction	3.07072	0.68997
	Short (1.5 secs)	Observation	1.76579	0.50055
	Long (3.0 secs)	Observation	2.91717	0.77155
Experiment 3	Long (3.0 secs)	Self reproduction	2.93184	0.74731

	Long (3.0 secs)	Mimicked reproduction	2.94213	0.80205
Experiment 4	Long (3.0 secs)	Belief of agency (Instruction: "YOURS")	3.13100	0.72508
	Long (3.0 secs)	Belief of non agency (Instruction: "OTHER'S")	3.12104	0.75059
Experiment 5	Long (3.0 secs)	Belief of agency (Empirical reproduction)	3.11343	0.63541
	Long (3.0 secs)	Belief of non agency (Empirical reproduction)	3.09059	0.63555
	Long (3.0 secs)	Belief of agency (Simulated reproduction)	3.05290	0.70929
	Long (3.0 secs)	Belief of non agency (Simulated reproduction)	3.11565	0.71439
Experiment 6 (laboratory)	Long (3.0 secs)	Belief of agency (Instruction: "YOURS")	3.25998	0.89804

	Long (3.0 secs)	Belief of non agency (Instruction: "OTHER'S")	3.27537	0.97408
Experiment 6 (online)	Long (3.0 secs)	Belief of agency (Instruction: "YOURS")	3.15583	0.63327
	Long (3.0 secs)	Belief of non agency (Instruction: "OTHER'S")	3.16572	0.68805

Table S1.2.

Descriptive statistics for reproductions for all experiments across short-long judgements

Experiment	Short-long judgement	Target duration	Mean	SD
Experiment 1	Short	Short (1.5 secs)	1.46937	0.40514
		Long (3.0 secs)	2.67436	0.61215
	Long	Short (1.5 secs)	1.71163	0.45582
		Long (3.0 secs)	3.14893	0.70211

Experiment 2	Short	Short (1.5 secs)	1.54549	0.38225
		Long (3.0 secs)	2.74524	0.72111
	Long	Short (1.5 secs)	1.90686	0.46042
		Long (3.0 secs)	3.24438	0.66191
Experiment 3	Short	Long (3.0 secs)	2.68320	0.69997
	Long	Long (3.0 secs)	3.16221	0.73403
Experiment 4	Short	Long (3.0 secs)	2.80202	0.73056
	Long	Long (3.0 secs)	3.37017	0.64328
Experiment 5	Short	Long (3.0 secs)	2.81192	0.61976
	Long	Long (3.0 secs)	3.33992	0.62360
Experiment 6 (laboratory)	Short	Long (3.0 secs)	2.73579	0.69224
	Long	Long (3.0 secs)	3.56811	0.92347
Experiment 6 (online)	Short	Long (3.0 secs)	2.84748	0.59611

	Long	Long (3.0 secs)	3.36196	0.62167

Table S1.3.

Linear mixed effects models with complex random effect structures. All models included the highest interaction term as random slope across participants as random effect structure

Experiment Number	Dependent Variable	Issue Encountered
1	Confidence judgments	cannot be computed due to too complex random effect structure
1	Short-long judgments	ar fit (possibly due to overfitting caused by too complex random effect structure)
2	Confidence judgments	cannot be computed due to too complex random effect structure
2	Short-long judgments	ar fit (possibly due to overfitting caused by too complex random effect structure)
3	Confidence judgments	ar fit (possibly due to overfitting caused by too complex random effect structure)
3	Short-long judgments	cannot be computed due to too complex random effect structure
4	Confidence judgments	cannot be computed due to too complex random effect structure
4	Short-long judgments	cannot be computed due to too complex random effect structure
5	Confidence judgments	cannot be computed due to too complex random effect structure
5	Short-long judgments	ar fit (possibly due to overfitting caused by too complex random effect structure)
6	onfidence judgments (lab)	cannot be computed due to too complex random effect structure
6	hort-long judgments (lab)	ar fit (possibly due to overfitting caused by too complex random effect structure)
6	idence judgments (online)	cannot be computed due to too complex random effect structure

6	rt-long judgments (online)	ar fit (possibly due to overfitting caused by too complex random effect structure)

Figure S1.1.

An example of discarded individual data due to unrealistic reproductions

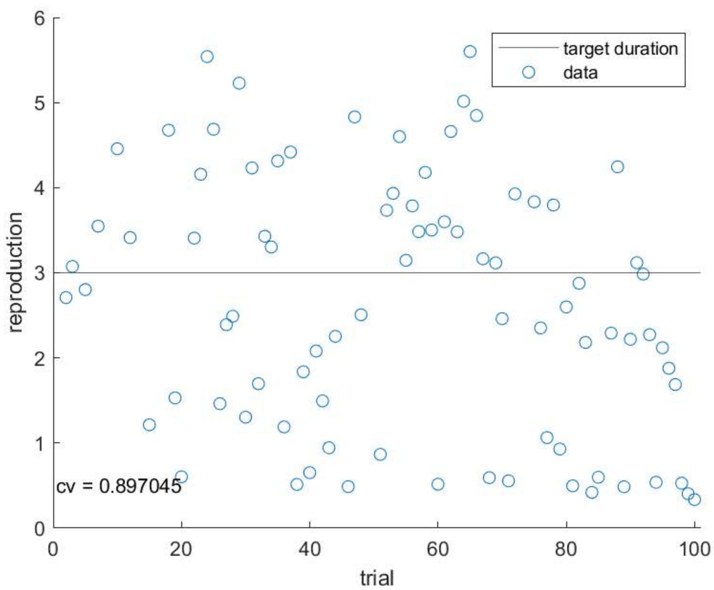


Figure S1.2.

An example of discarded individual data due to unrealistic reproductions

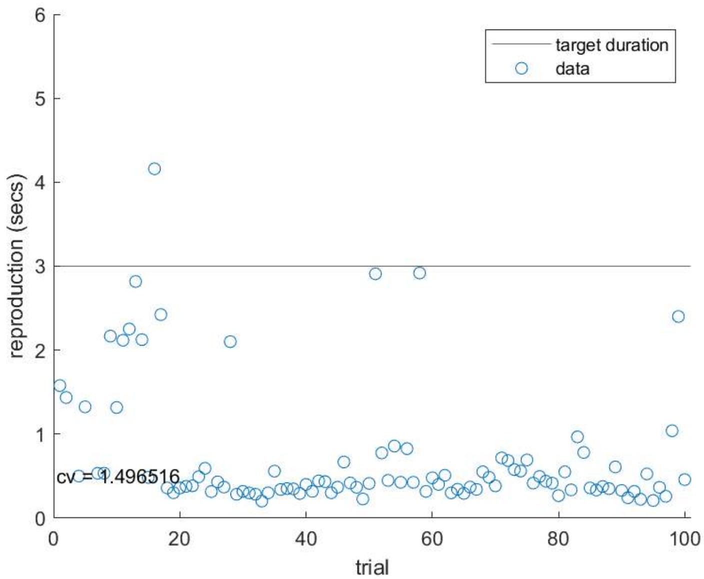


Figure S1.3.

A representation of data distributions for confidence (most left) and short-long judgements (middle) and reproduced durations (most right). Data presented in the figure are associated with an individual participant tested in Experiment 2.

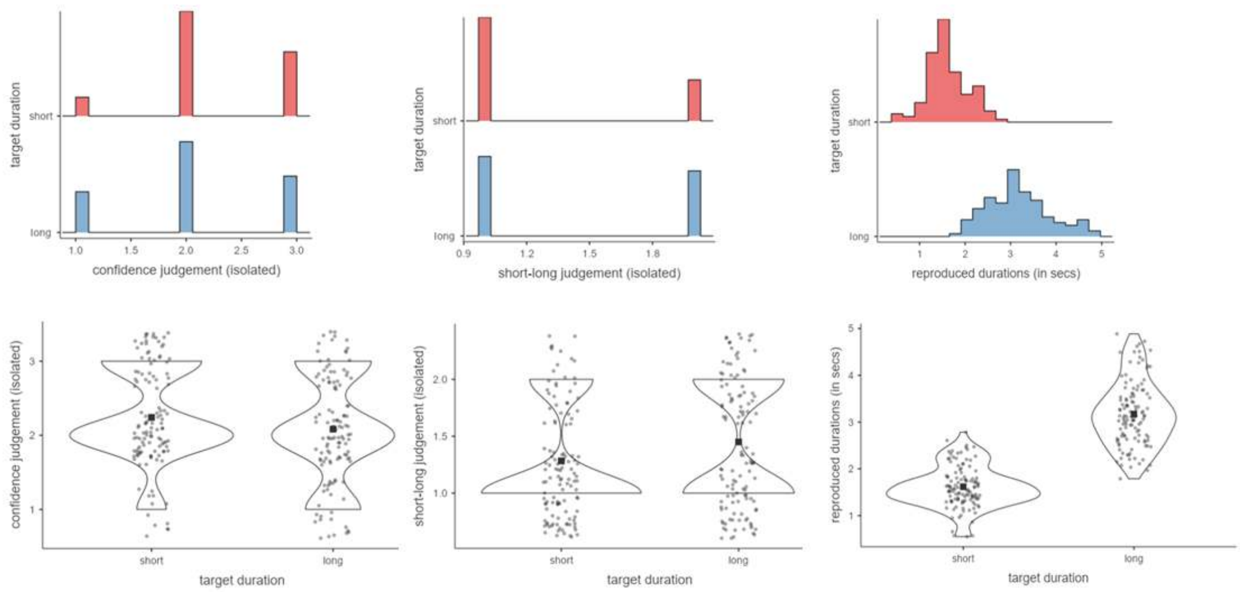


Figure S1.4.

The relationship between absolute timing error and confidence judgements across experimental conditions for per participant in Experiment 1.

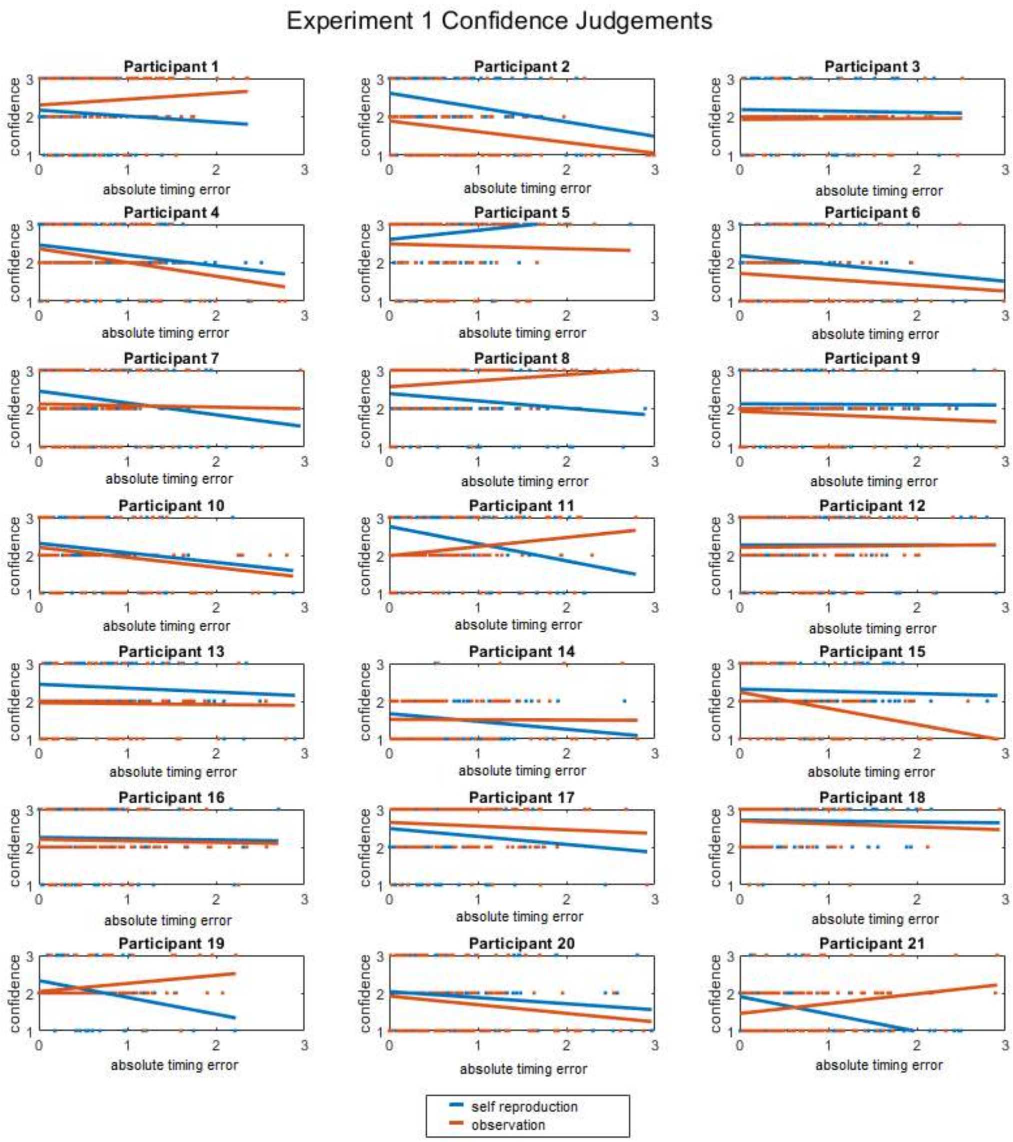


Figure S1.5.

The relationship between signed timing error and short-long judgements across experimental conditions for per participant in Experiment 1.

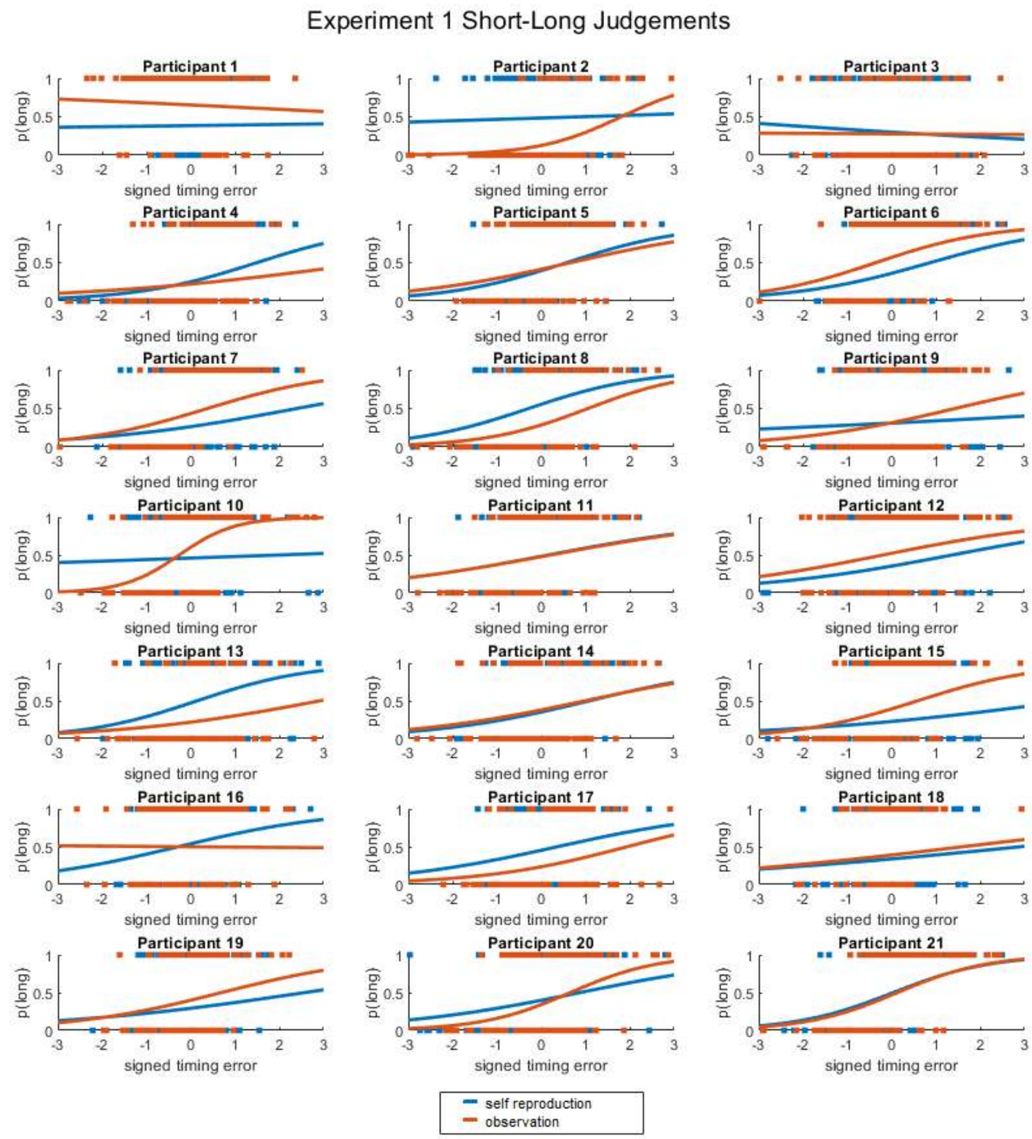


Figure S1.6.

The relationship between absolute timing error and confidence judgements across experimental conditions for per participant in Experiment 2.

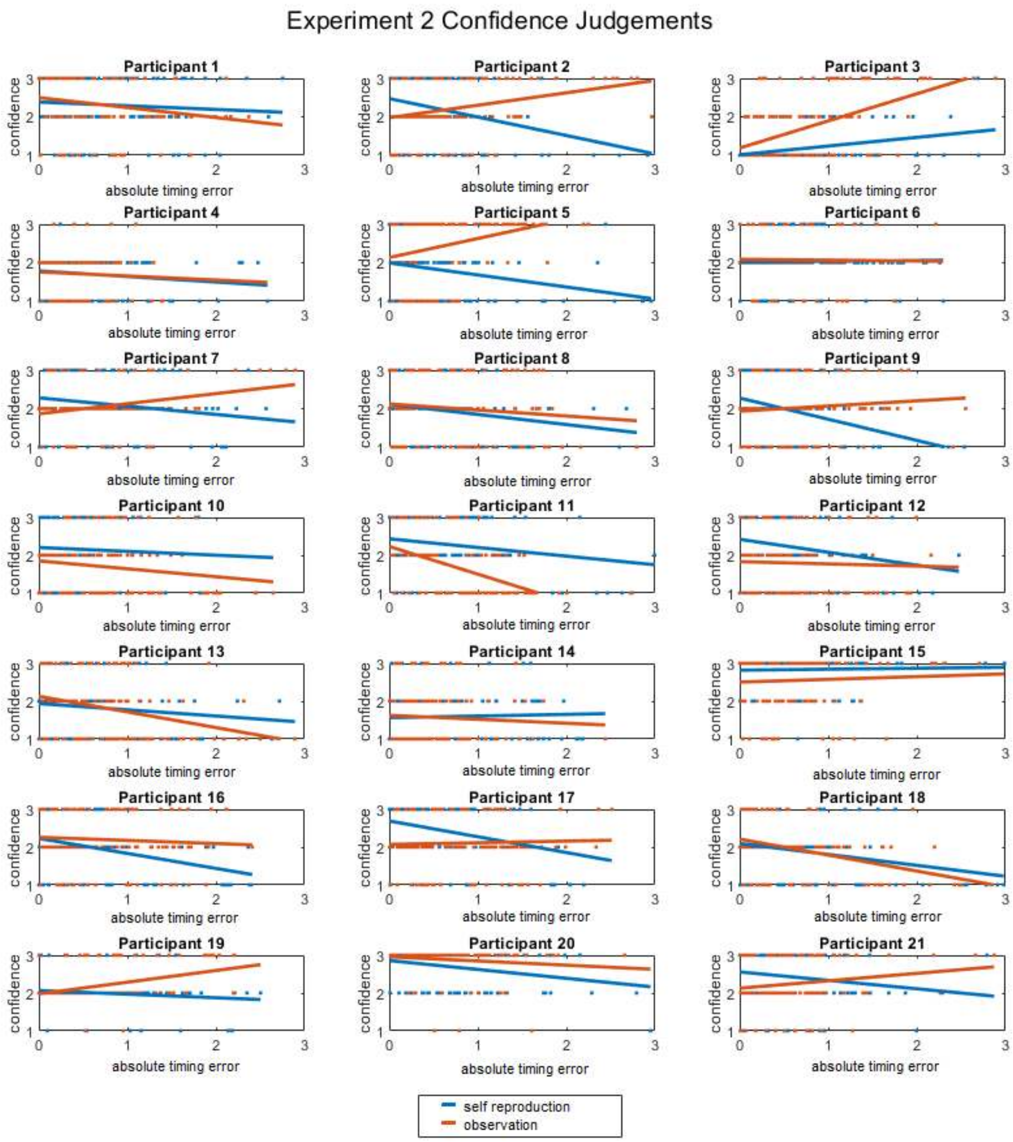


Figure S1.7.

The relationship between signed timing error and short-long judgements across experimental conditions for per participant in Experiment 2.

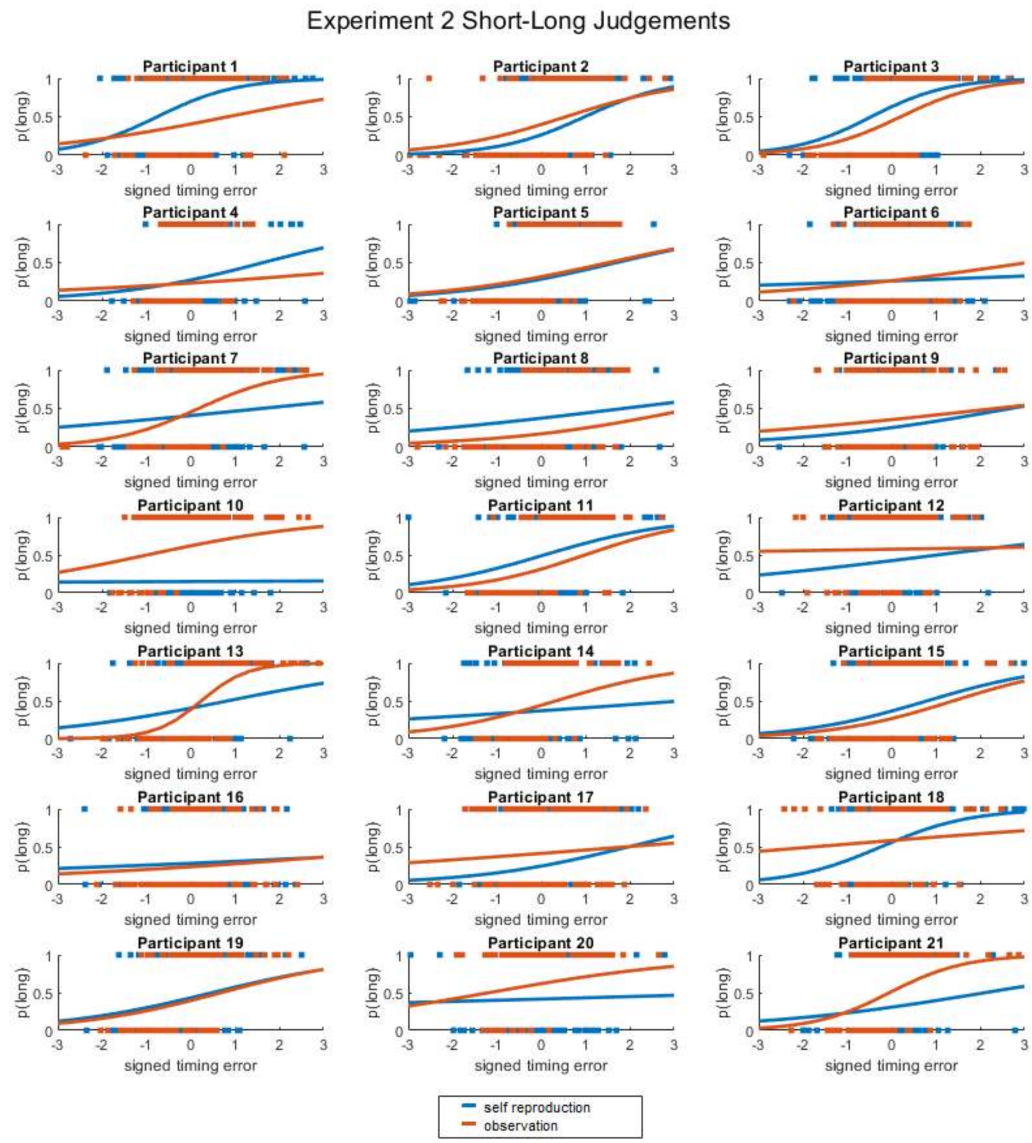


Figure S1.8.

The relationship between absolute timing error and confidence judgements across experimental conditions for per participant in Experiment 3.

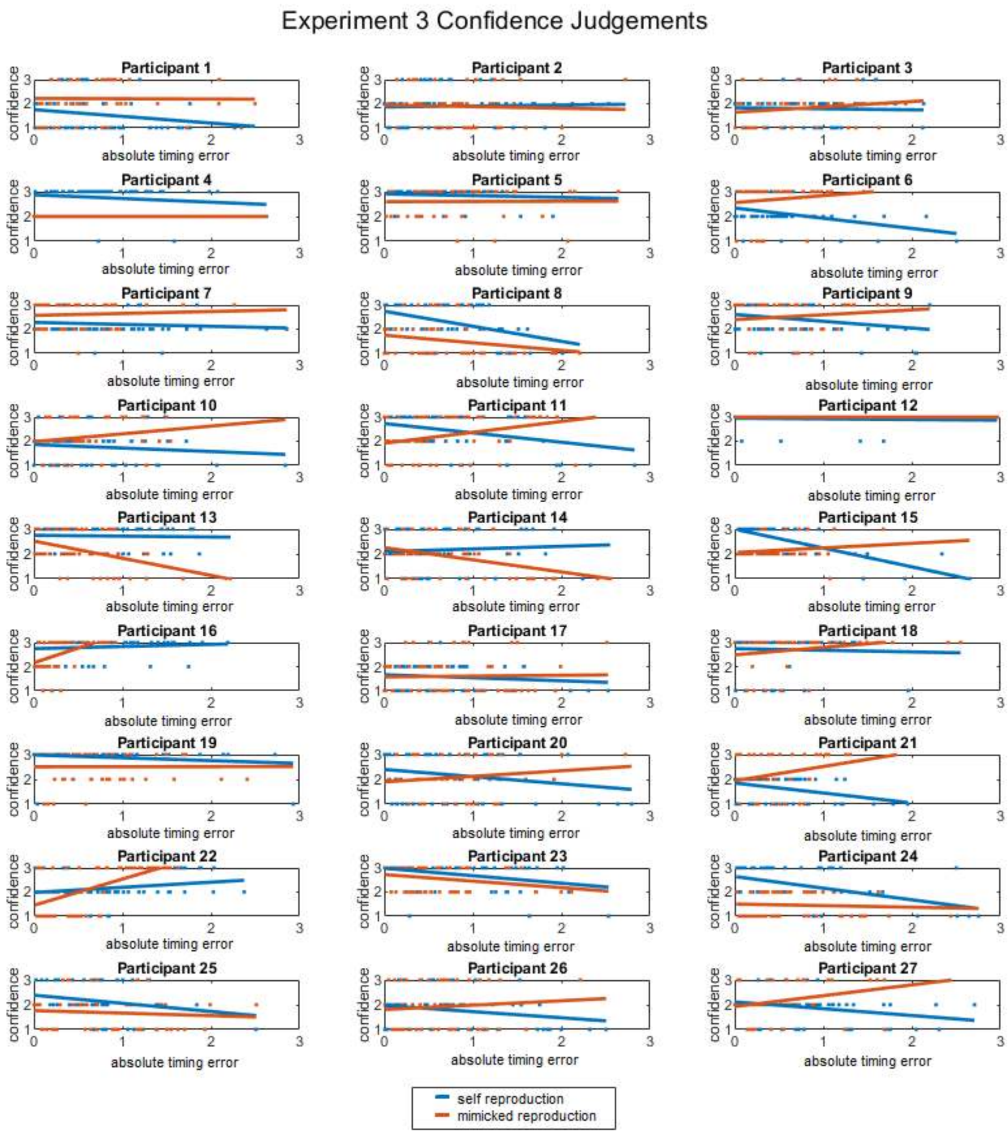


Figure S1.9.

The relationship between signed timing error and short-long judgements across experimental conditions for per participant in Experiment 3

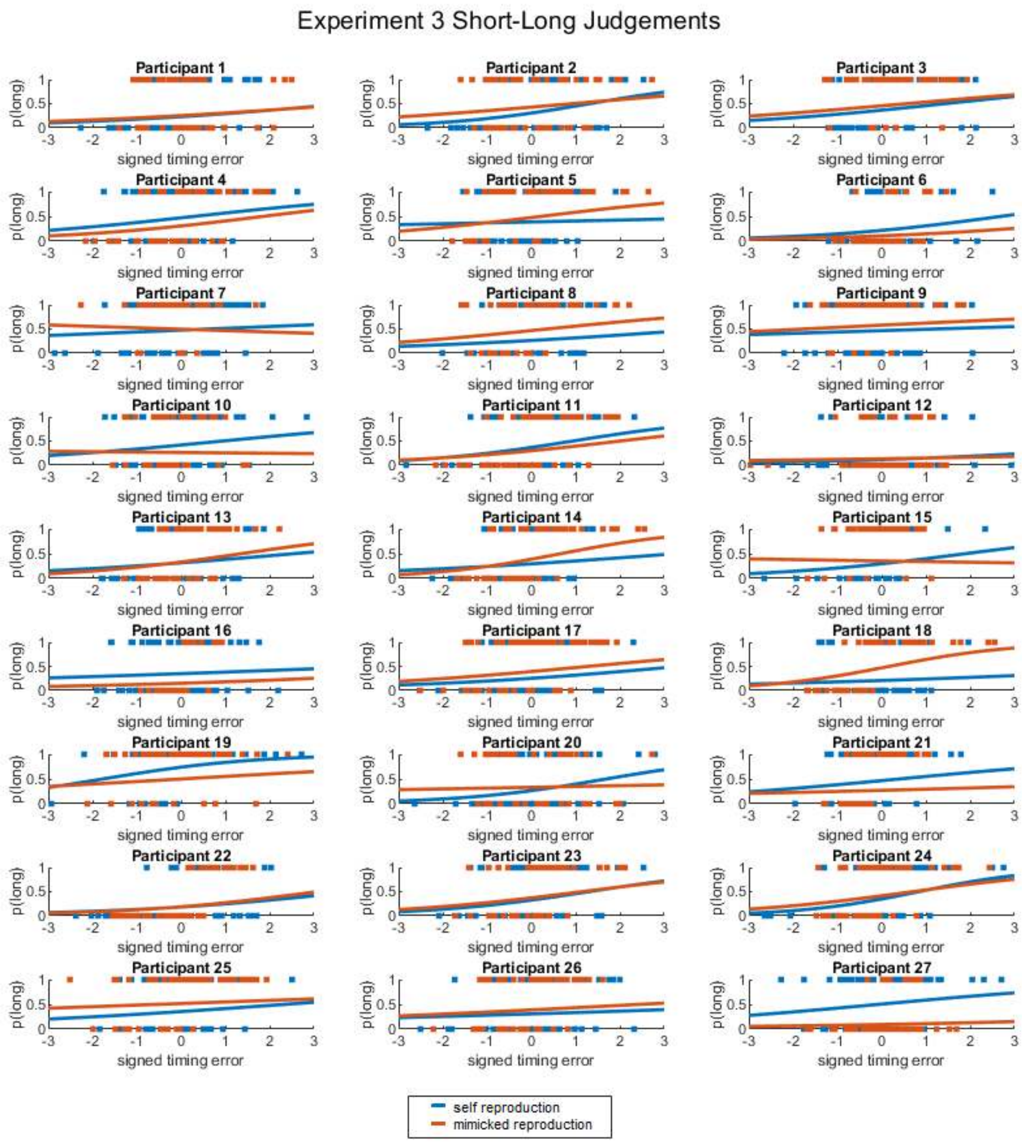


Figure S1.10.

The relationship between absolute timing error and confidence judgements across experimental conditions for per participant in Experiment 4.

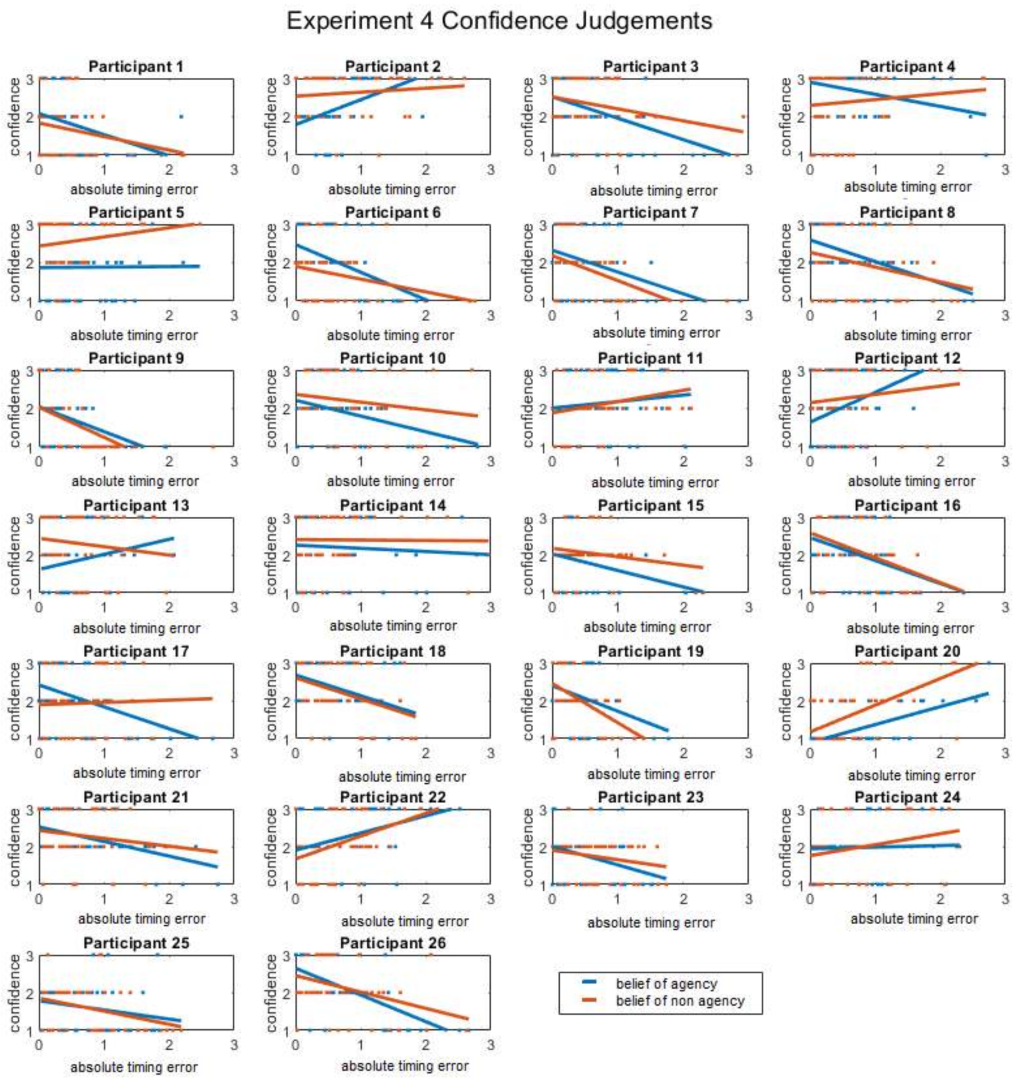


Figure S1.11.

The relationship between signed timing error and short-long judgements across experimental conditions for per participant in Experiment 4.

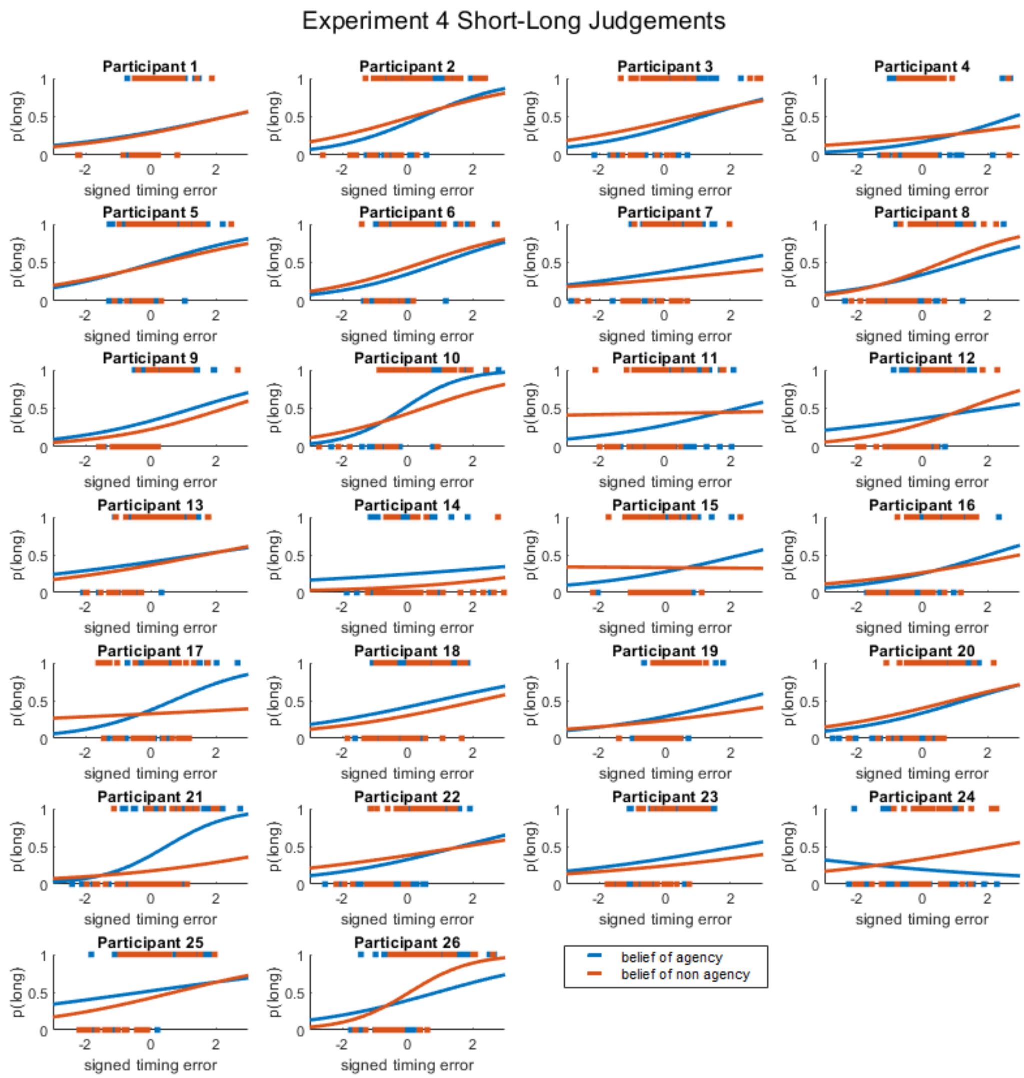


Figure S1.12.

The relationship between absolute timing error and confidence judgements across experimental conditions for per participant in Experiment 5.

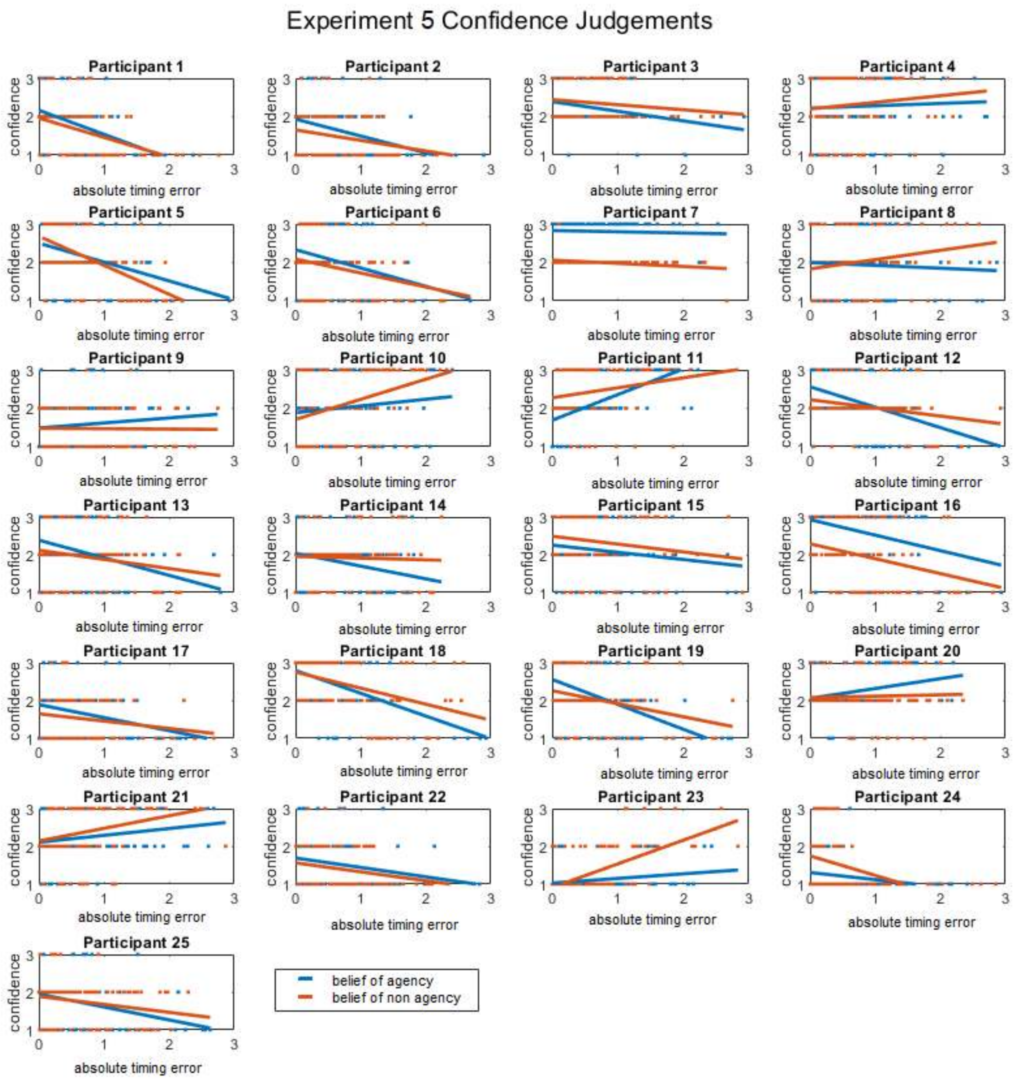


Figure S1.13.

The relationship between signed timing error and short-long judgements across experimental conditions for per participant in Experiment 5.

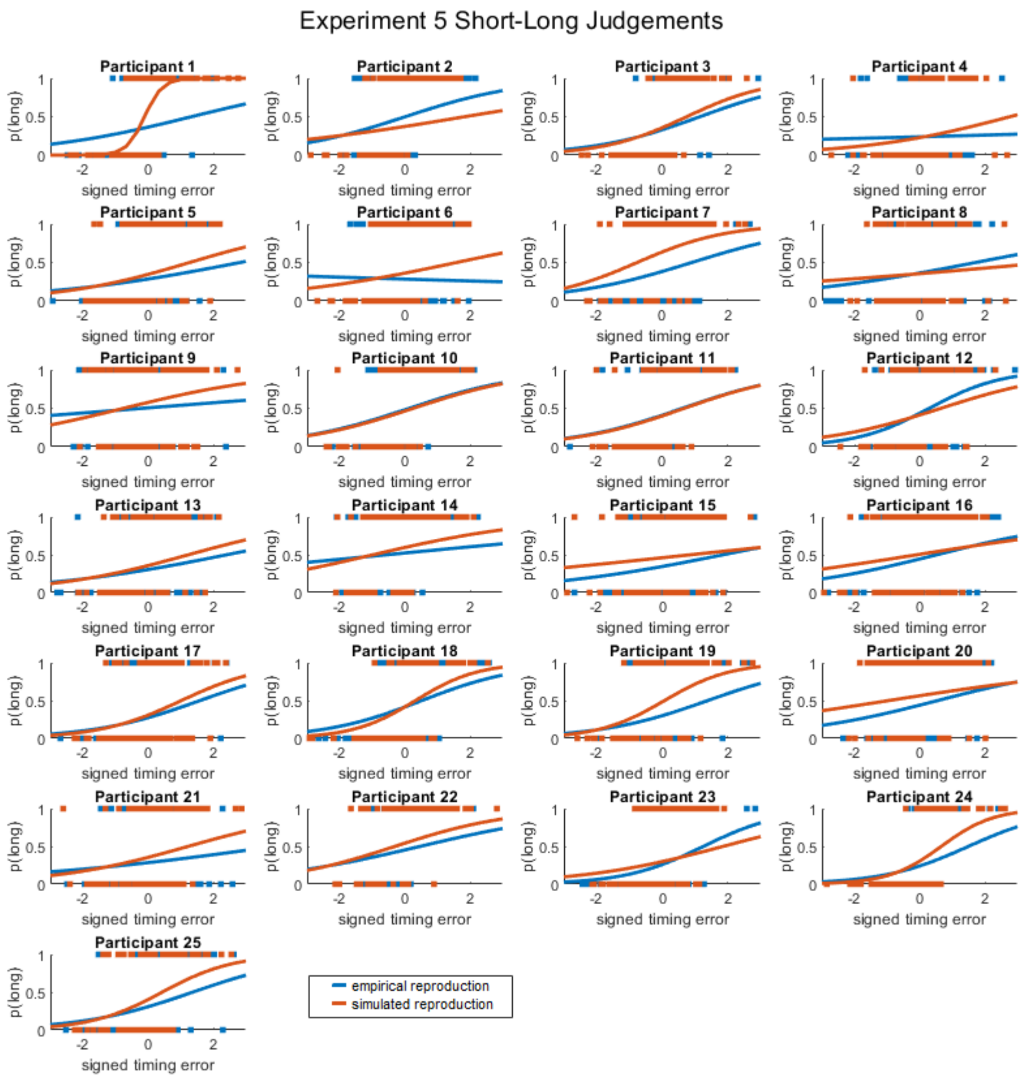


Figure S1.14.

The relationship between absolute timing error and confidence judgements across experimental conditions for per participant in Experiment 6 (in lab setting).

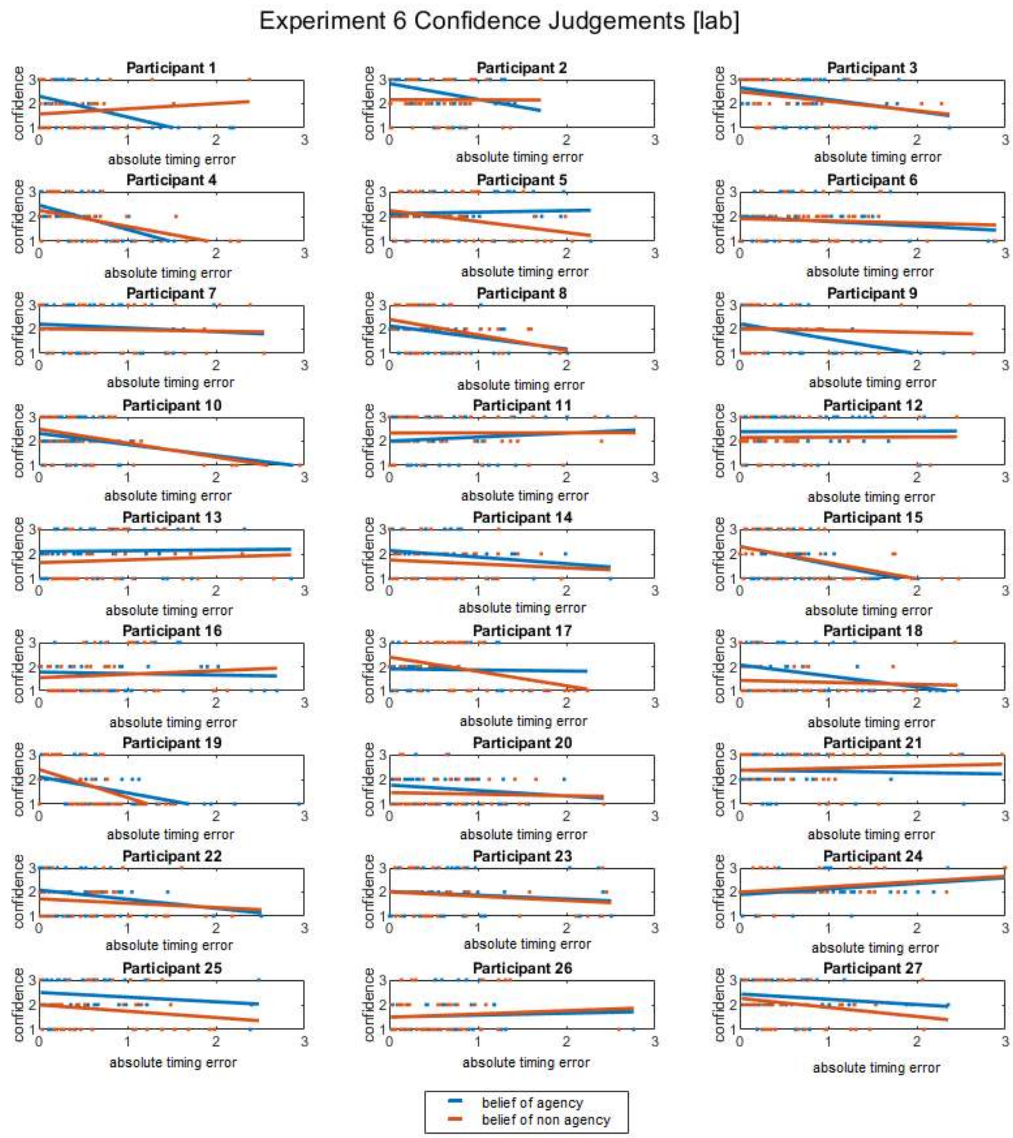


Figure S1.15.

The relationship between signed timing error and short-long judgements across experimental conditions for per participant in Experiment 6 (in lab setting).

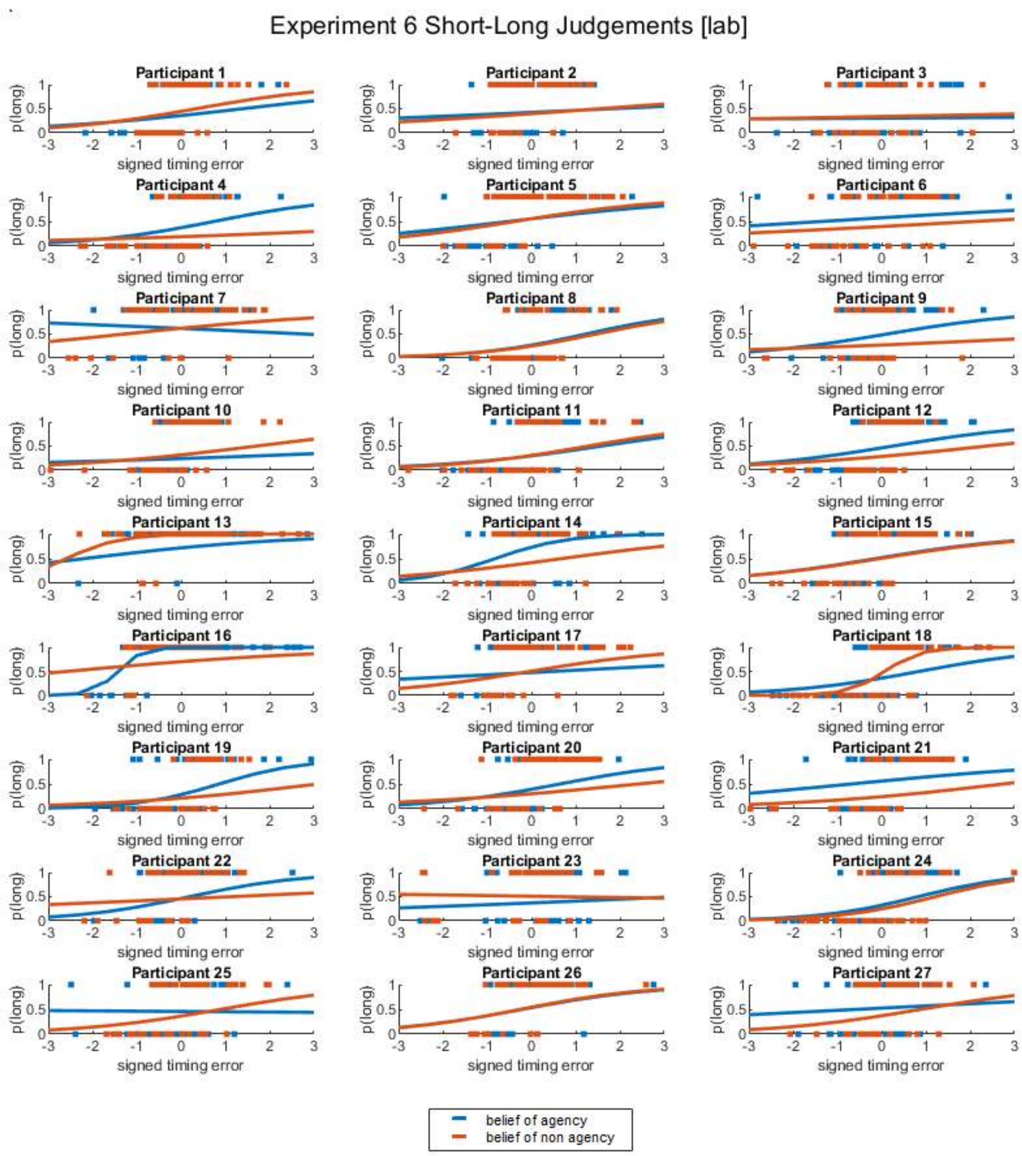


Figure S1.16.

The relationship between absolute timing error and confidence judgements across experimental conditions for per participant in Experiment 6 (in online setting).

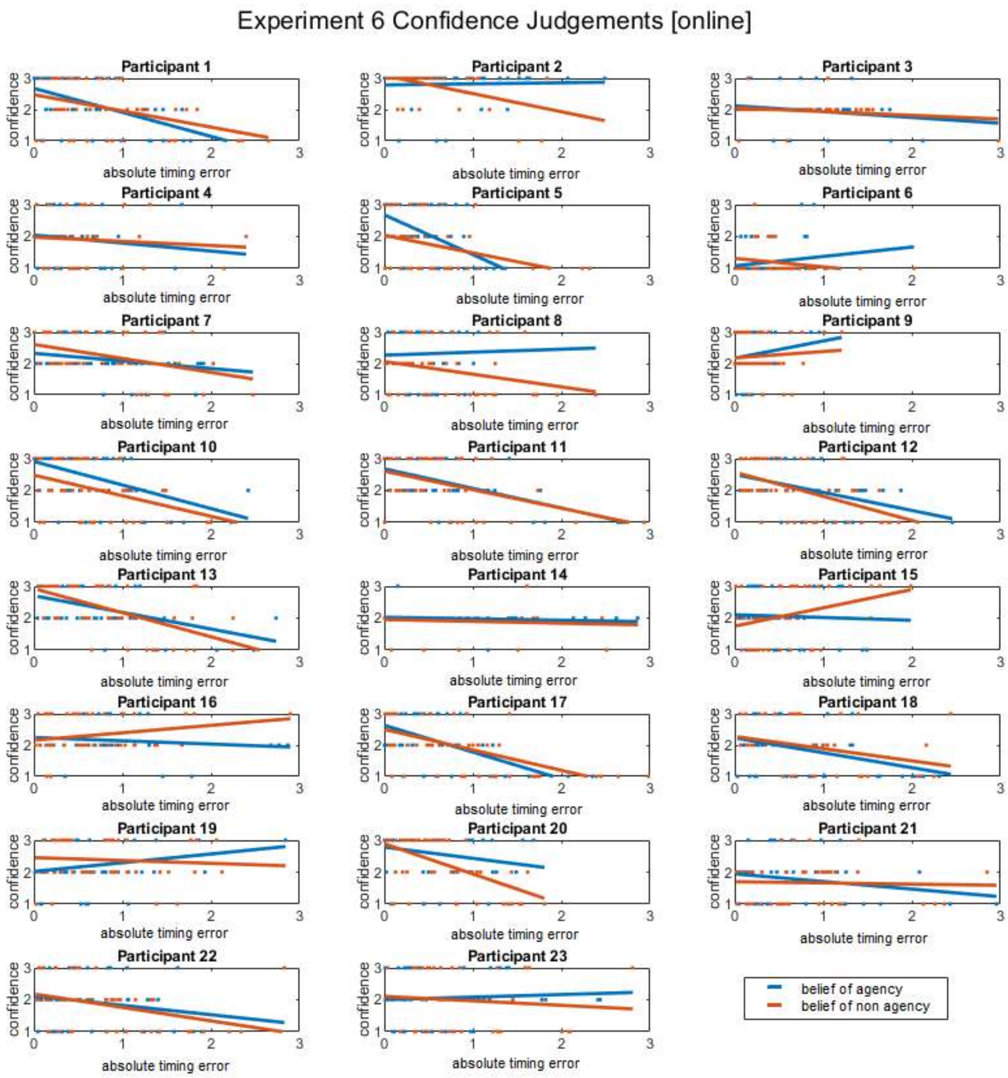
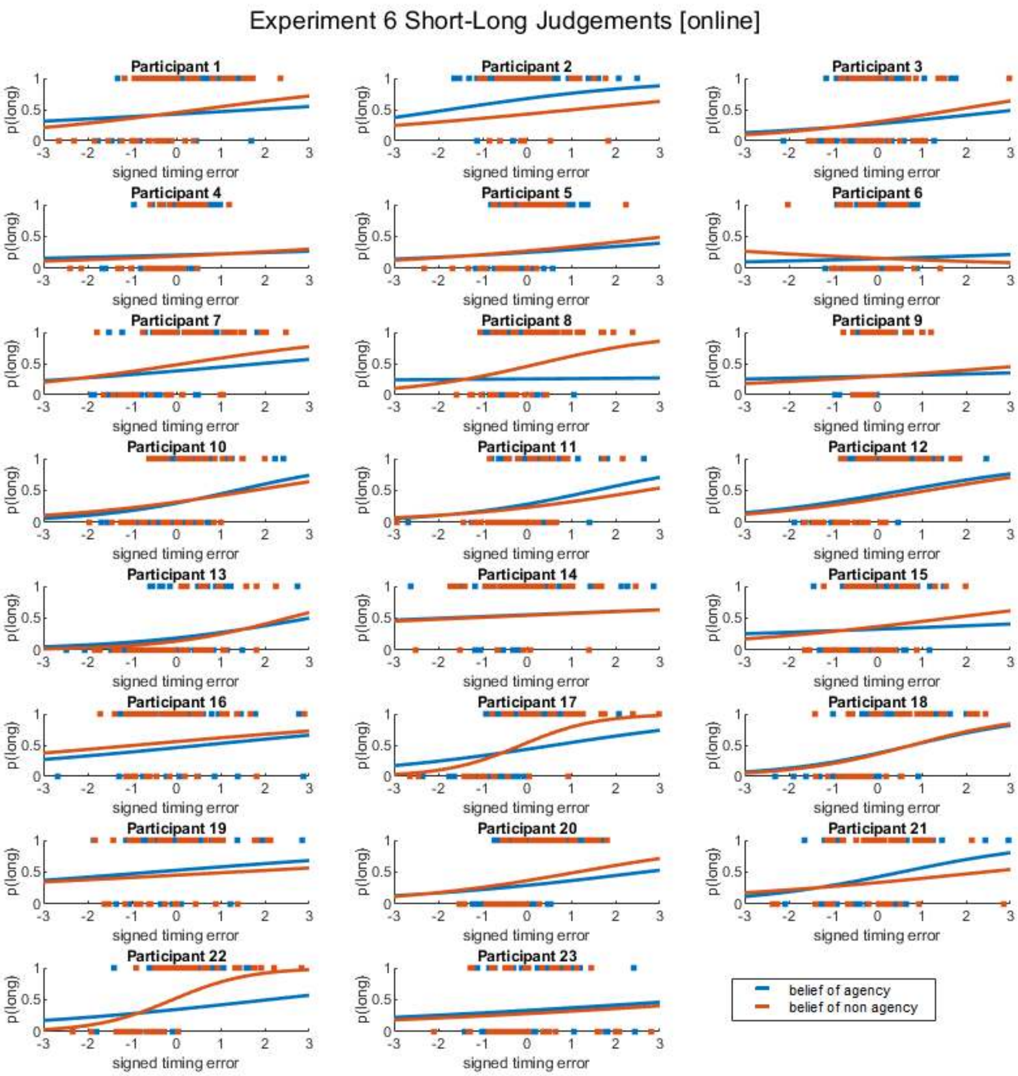


Figure S1.17.

The relationship between signed timing error and short-long judgements across experimental conditions for per participant in Experiment 6 (in online setting).



9.10. Supplementary Online Materials 2 (Chapter 4)

Table S2.1. *Model results for confidence judgments in Experiment 1*

Model Info

Info	
AIC	10206.10990
BIC	10271.24970
LogLikel.	-5093.05493
R-squared Marginal	0.01826
R-squared Conditional	0.16487

Fixed Effects Parameter Estimates

				95% Confidence Interval				
Names	Effect	ß	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	2.21469	0.06285	2.09152	2.33787	22.58	35.2402	< .00001
absolute timing error	absolute timing error	-0.1086	0.01693	-0.14176	-0.0754	4966.63	-6.4127	< .00001
target duration1	short - long	0.05874	0.03025	-5.520e−4	0.11804	4963.46	1.9417	0.05223
experimental condition1	self reproduction - observation	0.19245	0.03033	0.13300	0.25190	4964.66	6.3450	< .00001
target duration1 ✻ experimental condition1	short - long ✻ self reproduction - observation	-0.1541	0.06053	-0.27275	-0.0355	4963.62	-2.5462	0.01092

Fixed Effects Parameter Estimates

Names	Effect	ß	SE	95% Confidence Interval Lower	Upper	df	t	p
absolute timing error ✻ target duration1	absolute timing error ✻ short - long	-0.0154	0.03365	-0.08134	0.05058	4963.83	-0.4570	0.64767
absolute timing error ✻ experimental condition1	absolute timing error ✻ self reproduction - observation	-0.0923	0.03381	-0.15860	-0.0261	4965.92	-2.7312	0.00633
absolute timing error ✻ target duration1 ✻ experimental condition1	absolute timing error ✻ short - long ✻ self reproduction - observation	0.16365	0.06736	0.03163	0.29567	4964.17	2.4296	0.01515

Simple effects of absolute timing error : Parameter estimates

Moderator levels				95% Confidence Interval				
target duration	experimental condition	ß	SE	Lower	Upper	df	t	p
short	self reproduction	-0.1215	0.03504	-0.19022	-0.05282	4966	-3.4677	0.00053
	observation	-0.1110	0.03264	-0.17501	-0.04701	4965	-3.4006	0.00068
long	self reproduction	-0.1880	0.03505	-0.25668	-0.11925	4966	-5.3627	< .00001

Simple effects of absolute timing error : Parameter estimates

Moderator levels				95% Confidence Interval				
target duration	experimental condition	ß	SE	Lower	Upper	df	t	p
	observation	-0.0138	0.03216	-0.07685	0.04925	4964	-0.4292	0.66781

Table S2.2. *Model results for short-long judgments in Experiment 1*

Model Info

Info	Value	Comment
LogLikel.	-3061.4268	Unconditional Log-Likelihood
-2*LogLikel.	6122.8536	Unconditional absolute deviance
Deviance	6049.4332	Conditional relative deviance
R-squared	0.1571	Marginal
R-squared	0.2151	Conditional
AIC	6140.8500	Less is better
BIC	6199.4795	Less is better
Residual DF	4975.0000	
Chi-squared/DF	1.0040	Overdispersion indicator

Fixed Effects Parameter Estimates

				95% Confidence Interval			95% Exp(B) Confidence Interval			
Names	**Effect**	**ß**	**SE**	**Lower**	**Upper**	**exp(B)**	**Lower**	**Upper**	**z**	**p**
(Intercept)	(Intercept)	0.062	0.1119	-0.158	0.2811	1.064	0.854	1.33	0.5523	0.5808
target duration1	short - long	-0.032	0.0618	-0.153	0.0894	0.969	0.858	1.09	-0.513	0.6078
experimental condition1	self reproduction - observation	0.152	0.0618	0.0304	0.2728	1.164	1.031	1.31	2.4508	0.0143
signed timing error	signed timing error	0.891	0.0398	0.8156	0.9685	2.436	2.253	2.63	22.376	< .00001
target duration1 ✻ experimental condition1	short - long ✻ self reproduction - observation	5.926e-4	0.1237	-0.242	0.2430	1.001	0.785	1.28	0.0048	0.9962
target duration1 ✻ signed timing error	short - long ✻ signed timing error	-0.009	0.0787	-0.163	0.1457	0.991	0.850	1.16	-0.109	0.9130
experimental condition1 ✻ signed timing error	self reproduction - observation ✻ signed timing error	-0.083	0.0787	-0.237	0.0715	0.921	0.789	1.07	-1.052	0.2929

Fixed Effects Parameter Estimates

				95% Confidence Interval			95% Exp(B) Confidence Interval			
Names	**Effect**	**ß**	**SE**	**Lower**	**Upper**	**exp(B)**	**Lower**	**Upper**	**z**	**p**
target duration1 ✻ experimental condition1 ✻ signed timing error	short - long ✻ self reproduction - observation ✻ signed timing error	-0.356	0.1574	-0.665	-0.048	0.700	0.514	0.95	-2.263	0.0236

Simple effects of signed timing error : Parameter estimates

Moderator levels				95% Confidence Interval			95% Exp(B) Confidence Interval			
target duration	**experimental condition**	**ß**	**SE**	**Lower**	**Upper**	**exp(B)**	**Lower**	**Upper**	**z**	**p**
short	self reproduction	0.7557	0.07811	0.6026	0.9088	2.129	1.827	2.481	9.675	< .001
	observation	1.0166	0.08268	0.8545	1.1786	2.764	2.350	3.250	12.296	< .001
long	self reproduction	0.9424	0.08159	0.7825	1.1023	2.566	2.187	3.011	11.551	< .001
	observation	0.8471	0.07297	0.7041	0.9901	2.333	2.022	2.692	11.609	< .001

Table S2.3. *Model results for confidence judgments in Experiment 2*

Model Info

Info	
AIC	10341.2 1640
BIC	10406.3 1810
LogLikel.	- 5160.60 821
R-squared Marginal	0.01980
R-squared Conditional	0.18324

Fixed Effects Parameter Estimates

				95% Confidence Interval				
Names	Effect	ß	SE	Lowe r	Uppe r	df	t	p
(Intercept)	(Intercept)	2.135 54	0.067 95	2.002 36	2.268 73	22.30	31.42 65	< .00 001
absolute timing error	absolute timing error	- 0.100 9	0.017 52	- 0.135 2	- 0.066 5	4947. 25	- 5.757 1	< .00 001
target duration1	short - long	0.098 77	0.030 26	0.039 47	0.158 07	4945. 06	3.264 4	0.001 10
experimental condition1	self reproduction - observation	0.151 43	0.030 32	0.092 00	0.210 86	4945. 94	4.994 2	< .00 001
target duration1 ✻ experimental condition1	short - long ✻ self reproduction - observation	0.149 63	0.060 46	0.031 14	0.268 12	4944. 68	2.475 0	0.013 36

Fixed Effects Parameter Estimates

Names	Effect	ß	SE	95% Confidence Interval		df	t	p
				Lower	Upper			
absolute timing error ✻ target duration1	absolute timing error ✻ short - long	0.03110	0.03492	-0.0373	0.09955	4945.93	0.8906	0.37318
absolute timing error ✻ experimental condition1	absolute timing error ✻ self reproduction - observation	-0.2027	0.03506	-0.2715	-0.1341	4947.42	-5.7843	< .00001
absolute timing error ✻ target duration1 ✻ experimental condition1	absolute timing error ✻ short - long ✻ self reproduction - observation	-0.0704	0.06973	-0.2071	0.06629	4945.26	-1.0093	0.31289

Simple effects of absolute timing error : Parameter estimates

Moderator levels		ß	SE	95% Confidence Interval		df	t	p
target duration	experimental condition			Lower	Upper			
short	self reproduction	-0.20429	0.03514	-0.27319	-0.13540	4947	-5.8137	< .00001
	observation	0.03367	0.03565	-0.03623	0.10356	4946	0.9443	0.34509
long	self reproduction	-0.20021	0.03338	-0.26565	-0.13477	4946	-5.9976	< .00001

Simple effects of absolute timing error : Parameter estimates

Moderator levels				95% Confidence Interval				
target duration	experimental condition	ß	SE	Lower	Upper	df	t	p
	observation	-0.03263	0.03566	-0.10254	0.03728	4946	-0.9149	0.36027

Table S2.4. *Model results for short-long judgments in Experiment 2*

Model Info

Info	Value	Comment
LogLikel.	-3015.4381	Unconditional Log-Likelihood
-2*LogLikel.	6030.8762	Unconditional absolute deviance
Deviance	5963.1816	Conditional relative deviance
R-squared	0.1949	Marginal
R-squared	0.2382	Conditional
AIC	6048.8800	Less is better
BIC	6107.4677	Less is better
Residual DF	4956.0000	
Chi-squared/DF	1.0033	Overdispersion indicator

Fixed Effects Parameter Estimates

Names	Effect	ß	SE	exp(B)	95% Exp(B) Confidence Interval Lower	Upper	z	p
(Intercept)	(Intercept)	0.06909	0.09953	1.0715	0.8816	1.3023	0.6942	0.48756
signed timing error	signed timing error	0.96611	0.04276	2.6277	2.4165	2.8574	22.5961	< .00001
target duration1	short - long	0.08815	0.06298	1.0922	0.9653	1.2357	1.3996	0.16163
experimental condition1	self reproduction - observation	-0.3473	0.06301	0.7066	0.6245	0.7995	-5.5115	< .00001
target duration1 ✻ experimental condition1	short - long ✻ self reproduction - observation	-1.0488	0.12613	0.3503	0.2736	0.4486	-8.3157	< .00001
signed timing error ✻ target duration1	signed timing error ✻ short - long	0.05860	0.08465	1.0603	0.8982	1.2517	0.6922	0.48882
signed timing error ✻ experimental condition1	signed timing error ✻ self reproduction - observation	-0.4771	0.08471	0.6206	0.5256	0.7326	-5.6327	< .00001
signed timing error ✻ target duration1 ✻ experimental condition1	signed timing error ✻ short - long ✻ self reproduction - observation	0.10035	0.16930	1.1056	0.7934	1.5406	0.5927	0.55338

Simple effects of signed timing error : Parameter estimates

Moderator levels				95% Exp(B) Confidence Interval			
experimental condition	ß	SE	exp(B)	Lower	Upper	z	p
self reproduction	0.7275	0.05498	2.070	1.859	2.306	13.23	< .00001
observation	1.2047	0.06497	3.336	2.937	3.789	18.54	< .00001

Note. Simple effects are estimated keeping constant other independent variable(s) in the model

Table S2.5. *Model results for confidence judgments in Experiment 3*

Model Info

Info	
AIC	6049.56190
BIC	6085.47880
LogLikel.	-3018.78093
R-squared Marginal	0.01525
R-squared Conditional	0.23480

Fixed Effects Parameter Estimates

				95% Confidence Interval				
Names	Effect	ß	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	2.27952	0.07143	2.1395	2.41951	29.90	31.914	< .00001

Fixed Effects Parameter Estimates

Names	Effect	ß	SE	95% Confidence Interval Lower	95% Confidence Interval Upper	df	t	p
experimental condition1	self reproduction - mimicked reproduction	0.26225	0.04014	0.1836	0.34092	2915.18	6.534	< .00001
absolute timing error	absolute timing error	-0.0721	0.02288	-0.117	-0.0273	2918.31	-3.153	0.00163
experimental condition1 ✻ absolute timing error	self reproduction - mimicked reproduction ✻ absolute timing error	-0.2331	0.04551	-0.322	-0.1439	2916.22	-5.122	< .00001

Simple effects of absolute timing error : Parameter estimates

Moderator levels: experimental condition	ß	SE	95% Confidence Interval Lower	95% Confidence Interval Upper	df	t	p
self reproduction	-0.18871	0.02955	-0.24665	-0.1308	2919	-6.385	< .00001
mimicked reproduction	0.04443	0.03481	-0.02384	0.1127	2921	1.276	0.20202

Table S2.6. *Model results for short-long judgments in Experiment 3*

Model Info

Info	Value	Comment
LogLikel.	-1772.3570	Unconditional Log-Likelihood
-2*LogLikel.	3544.7141	Unconditional absolute deviance
Deviance	3453.2635	Conditional relative deviance
R-squared	0.1411	Marginal
R-squared	0.2537	Conditional
AIC	3554.7100	Less is better
BIC	3584.6449	Less is better
Residual DF	2935.0000	
Chi-squared/DF	0.9828	Overdispersion indicator

Fixed Effects Parameter Estimates

					95% Exp(B) Confidence Interval			
Names	Effect	ß	SE	exp(B)	Lower	Upper	z	p
(Intercept)	(Intercept)	0.2685	0.14189	1.3080	0.9904	1.7273	1.892	0.05847
signed timing error	signed timing error	0.8655	0.05338	2.3763	2.1402	2.6383	16.214	< .00001
experimental condition1	self reproduction - mimicked reproduction	-0.411	0.08334	0.6633	0.5633	0.7810	-4.926	< .00001

Fixed Effects Parameter Estimates

Names	Effect	ß	SE	exp(B)	95% Exp(B) Confidence Interval Lower	Upper	z	p
signed timing error ✻ experimental condition1	signed timing error ✻ self reproduction - mimicked reproduction	-0.236	0.10459	0.7902	0.6437	0.9699	-2.252	0.02434

Simple effects of signed timing error : Parameter estimates

Moderator levels				95% Exp(B) Confidence Interval			
experimental condition	ß	SE	exp(B)	Lower	Upper	z	p
self reproduction	0.7478	0.06554	2.112	1.858	2.402	11.41	< .00001
mimicked reproduction	0.9833	0.08291	2.673	2.272	3.145	11.86	< .00001

Table S2.7. *Model results for confidence judgments in Experiment 4*

Model Info

Info	
AIC	5655.39150
BIC	5690.47810
LogLikel.	-2821.69574
R-squared Marginal	0.01863

Model Info

Info	
R-squared Conditional	0.12433

Fixed Effects Parameter Estimates

				95% Confidence Interval				
Names	**Effect**	**ß**	**SE**	**Lower**	**Upper**	**df**	**t**	**p**
(Intercept)	(Intercept)	2.15342	0.05386	2.04786	2.258980	32.26	39.9834	< .00001
absolute timing error	absolute timing error	-0.1774	0.02608	-0.2285	-0.12629	2544.27	-6.8032	< .00001
experimental condition1	belief of agency – belief of non agency	0.02177	0.04445	-0.0654	0.108894	2535.37	0.4898	0.62433
absolute timing error ✻ experimental condition1	absolute timing error ✻ belief of agency – belief of non agency	-0.1085	0.05160	-0.2097	-0.00736	2536.29	-2.1027	0.03559

Simple effects of absolute timing error : Parameter estimates

Moderator levels			95% Confidence Interval				
experimental condition	ß	SE	Lower	Upper	df	t	p
belief of agency	-0.2317	0.03700	-0.3042	-0.15910	2544	-6.261	< .00001
belief of non agency	-0.1231	0.03642	-0.1946	-0.05173	2543	-3.381	0.00073

Table S2.8. *Model results for short-long judgments in Experiment 4*

Model Info

Info	Value	Comment
LogLikel.	-1368.3346	Unconditional Log-Likelihood
-2*LogLikel.	2736.6693	Unconditional absolute deviance
Deviance	2664.2704	Conditional relative deviance
R-squared	0.3433	Marginal
R-squared	0.4052	Conditional
AIC	2746.6700	Less is better
BIC	2775.9081	Less is better
Residual DF	2555.0000	
Chi-squared/DF	1.0406	Overdispersion indicator

Fixed Effects Parameter Estimates

Names	Effect	ß	SE	exp(B)	95% Exp(B) Confidence Interval Lower	Upper	z	p
(Intercept)	(Intercept)	0.41605	0.12450	1.5160	1.1877	1.935	3.3419	0.00083
signed timing error	signed timing error	1.59636	0.07667	4.9351	4.2465	5.735	20.8211	< .00001
experimental condition1	belief of agency – belief of non agency	0.04907	0.09602	1.0503	0.8701	1.268	0.5110	0.60934
signed timing error ✻ experimental condition1	signed timing error ✻ belief of agency – belief of non agency	-0.0261	0.15074	0.9742	0.7250	1.309	-0.1731	0.86254

Simple effects of signed timing error : Parameter estimates

Moderator levels: experimental condition	ß	SE	exp(B)	95% Exp(B) Confidence Interval Lower	Upper	z	p
belief of agency	1.583	0.1081	4.871	3.941	6.021	14.64	< .00001
belief of non agency	1.609	0.1069	5.000	4.055	6.165	15.06	< .00001

Table S2.9. *Model results for confidence judgments in Experiment 5*

Model Info

Info	
AIC	10080.73650
BIC	10145.81600
LogLikel.	-5030.36824
R-squared Marginal	0.01900
R-squared Conditional	0.23126

Fixed Effects Parameter Estimates

				95% Confidence Interval				
Names	**Effect**	**ß**	**SE**	**Lower**	**Upper**	**df**	**t**	**p**
(Intercept)	(Intercept)	2.07820	0.07112	1.93880	2.217591	26.45	29.2204	< .00001
reproduction type1	empirical reproduction – simulated reproduction	-0.0365	0.03031	-0.0959	0.022861	4929.81	-1.2057	0.22801
instruction type1	belief of agency – belief of non agency	0.09502	0.03028	0.03568	0.154370	4929.47	3.1383	0.00171
absolute timing error	absolute timing error	-0.1680	0.01642	-0.2001	-0.13577	4930.23	-10.230	< .00001

Fixed Effects Parameter Estimates

				95% Confidence Interval				
Names	**Effect**	**ß**	**SE**	**Lower**	**Upper**	**df**	**t**	**p**
reproduction type1 ✻ instruction type1	empirical reproduction – simulated reproduction ✻ belief of agency – belief of non agency	0.02465	0.06056	-0.0940	0.143350	4929.48	0.4071	0.68397
reproduction type1 ✻ absolute timing error	empirical reproduction – simulated reproduction ✻ absolute timing error	0.02843	0.03287	-0.0360	0.092846	4930.57	0.8649	0.38712
instruction type1 ✻ absolute timing error	belief of agency – belief of non agency ✻ absolute timing error	-0.0721	0.03279	-0.1364	-0.00784	4929.70	-2.1990	0.02792

Fixed Effects Parameter Estimates

Names	Effect	ß	SE	95% Confidence Interval Lower	95% Confidence Interval Upper	df	t	p
reproducti on type1 ✻ instruction type1 ✻ absolute timing error	empirical reproducti on – simulated reproducti on ✻ belief of agency – belief of non agency ✻ absolute timing error	0.0437 7	0.065 59	- 0.084 8	0.1723 27	4929. 80	0.667 3	0.5046 3

Simple effects of absolute timing error : Parameter estimates

Moderator levels			95% Confidence Interval				
instruction type	ß	SE	Lower	Upper	df	t	p
belief of agency	- 0.2040	0.0231 0	-0.2493	-0.15870	493 0	- 8.830	< .0000 1
belief of non agency	- 0.1319	0.0233 0	-0.1776	-0.08622	493 0	- 5.661	< .0000 1

Note. Simple effects are estimated setting higher order moderator (if any) in covariates to zero and averaging across mode factors levels (if any)

Table S2.10. *Model results for short-long judgments in Experiment 5*

Model Info

Info	Value	Comment
LogLikel.	-2768.8393	Unconditional Log-Likelihood
-2*LogLikel.	5537.6786	Unconditional absolute deviance
Deviance	5447.6653	Conditional relative deviance
R-squared	0.2773	Marginal
R-squared	0.3494	Conditional
AIC	5555.6800	Less is better
BIC	5614.2501	Less is better
Residual DF	4945.0000	
Chi-squared/DF	1.0139	Overdispersion indicator

Fixed Effects Parameter Estimates

					95% Exp(B) Confidence Interval			
Names	**Effect**	**ß**	**SE**	**exp(B)**	**Lower**	**Upper**	**z**	**p**
(Intercept)	(Intercept)	0.22915	0.12521	1.2575	0.9839	1.6073	1.8302	0.06722
reproduction type1	simulated reproduction – empirical reproduction	0.14278	0.06618	1.1535	1.0132	1.3132	2.1575	0.03096
instruction type1	belief of non agency – belief of agency	0.02045	0.06616	1.0207	0.8965	1.1620	0.3091	0.75722

Fixed Effects Parameter Estimates

					95% Exp(B) Confidence Interval			
Names	**Effect**	**ß**	**SE**	**exp(B)**	**Lower**	**Upper**	**z**	**p**
signed timing error	signed timing error	1.24836	0.04544	3.4846	3.1877	3.8092	27.4709	< .00001
reproduction type1 ✻ instruction type1	simulated reproduction - empirical reproduction ✻ belief of non agency – belief of agency	-0.3679	0.13232	0.6922	0.5341	0.8971	-2.7806	0.00543
reproduction type1 ✻ signed timing error	simulated reproduction - empirical reproduction ✻ signed timing error	0.28228	0.08891	1.3261	1.1141	1.5786	3.1747	0.00150
instruction type1 ✻ signed timing error	belief of non agency – belief of agency ✻ signed timing error	0.08323	0.08878	1.0868	0.9132	1.2934	0.9374	0.34853
reproduction type1 ✻ instruction type1 ✻ signed timing error	simulated reproduction - empirical reproduction ✻ belief of non agency – belief of agency ✻ signed timing error	0.10282	0.17764	1.1083	0.7824	1.5699	0.5788	0.56270

Simple effects of signed timing error : Parameter estimates

Moderator levels				95% Exp(B) Confidence Interval			
reproduction type	ß	SE	exp(B)	Lower	Upper	z	p
empirical reproduction	1.107	0.06364	3.026	2.671	3.428	17.40	< .00001
simulated reproduction	1.389	0.06350	4.013	3.543	4.545	21.88	< .00001

Note. Simple effects are estimated keeping constant other independent variable(s) in the model

Table S2.10.1. *Model results for confidence judgments in Experiment 6 (lab)*

Model Info

Info	
AIC	6060.73530
BIC	6096.04500
LogLikel.	-3024.36766
R-squared Marginal	0.02693
R-squared Conditional	0.11089

Fixed Effects Parameter Estimates

				95% Confidence Interval				
Names	Effect	ß	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	2.08870	0.04951	1.99167	2.18574	34.76	42.1888	< .00001

Fixed Effects Parameter Estimates

Names	Effect	ß	SE	95% Confidence Interval Lower	95% Confidence Interval Upper	df	t	p
absolute timing error	absolute timing error	-0.2130	0.02643	-0.2648	-0.1612	2643.10	-8.0571	< .00001
experimental condition1	belief of agency – belief of non agency	0.13580	0.04466	0.04827	0.22334	2633.42	3.0408	0.00238
absolute timing error ✻ experimental condition1	absolute timing error ✻ belief of agency – belief of non agency	-0.0495	0.05243	-0.1523	0.05325	2635.42	-0.9443	0.34512

Simple effects of absolute timing error : Parameter estimates

Moderator levels			95% Confidence Interval				
experimental condition	**ß**	**SE**	**Lower**	**Upper**	**df**	**t**	**p**
belief of agency	-0.2377	0.03751	-0.3113	-0.1642	2644	-6.337	< .00001
belief of non agency	-0.1882	0.03699	-0.2607	-0.1157	2641	-5.087	< .00001

Table S2.10.2. *Model results for confidence judgments in Experiment 6 (online)*

Model Info

Info	
AIC	4694.34300
BIC	4728.71080
LogLikel.	-2341.17148
R-squared Marginal	0.06013
R-squared Conditional	0.21140

Fixed Effects Parameter Estimates

				95% Confidence Interval				
Names	Effect	ß	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	2.29487	0.06489	2.16768	2.4221	26.49	35.3642	< .00001
absolute timing error	absolute timing error	-0.3298	0.02638	-0.3815	-0.278	2258.76	-12.498	< .00001
experimental condition1	belief of agency – belief of non agency	0.03989	0.04358	-0.0455	0.1253	2248.83	0.9154	0.36009
absolute timing error ✻ feedback1	absolute timing error ✻ belief of agency – belief of non agency	0.02613	0.05160	-0.0750	0.1273	2249.52	0.5064	0.61262

Simple effects of absolute timing error : Parameter estimates

Moderator levels			**95% Confidence Interval**				
experimental condition	**ß**	**SE**	**Lower**	**Upper**	**df**	**t**	**p**
belief of agency	-0.3167	0.03839	-0.3920	-0.2414	2258	-8.250	< .00001
belief of non agency	-0.3428	0.03542	-0.4123	-0.2733	2257	-9.678	< .00001

Table S2.11.1 *Model results for short-long judgments in Experiment 6 (lab)*

Model Info

Info	Value	Comment
LogLikel.	-1280.5756	Unconditional Log-Likelihood
-2*LogLikel.	2561.1511	Unconditional absolute deviance
Deviance	2459.1812	Conditional relative deviance
R-squared	0.3682	Marginal
R-squared	0.5262	Conditional
AIC	2571.1500	Less is better
BIC	2600.5759	Less is better
Residual DF	2652.0000	
Chi-squared/DF	1.1164	Overdispersion indicator

Fixed Effects Parameter Estimates

					95% Exp(B) Confidence Interval			
Names	**Effect**	**ß**	**SE**	**exp(B)**	**Lower**	**Upper**	**z**	**p**
(Intercept)	(Intercept)	0.92570	0.20909	2.5236	1.6752	3.802	4.4274	< .00001
experimental condition1	belief of non agency – belief of agency	-0.0502	0.10249	0.9510	0.7780	1.163	-0.4898	0.62429
signed timing error	signed timing error	1.87122	0.08933	6.4962	5.4528	7.739	20.9474	< .00001
experimental condition1	belief of non agency –	0.13274	0.16254	1.1420	0.8304	1.570	0.8167	0.41412

✻ signed timing error	belief of agency ✻ signed timing error

Simple effects of signed timing error : Parameter estimates

Moderator levels				95% Exp(B) Confidence Interval			
experimental condition	**ß**	**SE**	**exp(B)**	**Lower**	**Upper**	**z**	**p**
belief of agency	1.805	0.1178	6.079	4.825	7.659	15.32	< .00001
belief of non agency	1.938	0.1236	6.942	5.448	8.845	15.67	< .00001

Table S2.11.2 *Model results for short-long judgments in Experiment 6 (online)*

Model Info

Info	Value	Comment
LogLikel.	-1210.1907	Unconditional Log-Likelihood
-2*LogLikel.	2420.3813	Unconditional absolute deviance
Deviance	2340.6291	Conditional relative deviance
R-squared	0.2670	Marginal
R-squared	0.3958	Conditional
AIC	2430.3800	Less is better
BIC	2459.0212	Less is better
Residual DF	2266.0000	
Chi-squared/DF	1.2817	Overdispersion indicator

Fixed Effects Parameter Estimates

					95% Exp(B) Confidence Interval			
Names	**Effect**	**ß**	**SE**	**exp(B)**	**Lower**	**Upper**	**z**	**p**
(Intercept)	(Intercept)	0.66651	0.18257	1.947	1.3616	2.785	3.6507	0.00026
feedback1	belief of agency – belief of non agency	0.17902	0.10380	1.196	0.9759	1.466	1.7247	0.08458
signed timing error	signed timing error	1.42087	0.07961	4.141	3.5425	4.840	17.8485	< .00001

Fixed Effects Parameter Estimates

Names	Effect	ß	SE	exp(B)	95% Exp(B) Confidence Interval Lower	Upper	z	p
feedback1 ✻ signed timing error	belief of agency – belief of non agency ✻ signed timing error	0.06434	0.15358	1.066	0.7893	1.441	0.4189	0.67526

Simple effects of signed timing error : Parameter estimates

Moderator levels				95% Exp(B) Confidence Interval			
feedback	ß	SE	exp(B)	Lower	Upper	z	p
belief of agency	1.453	0.1144	4.276	3.417	5.351	12.70	< .00001
belief non agency	1.389	0.1067	4.010	3.253	4.942	13.01	< .00001

Table S2.12. The reproduction descriptives across different sources (empirical and simulated) for per participant

reproduction type	Participant ID	Mean	Median	SD	Shapiro-Wilk (W)	Shapiro-Wilk (p)
empirical	1	3.015	3.034	0.3311	0.9612	0.00516
	135	3.273	3.282	0.4432	0.9883	0.54187
	10971	3.306	3.268	0.3695	0.9733	0.04357
	61028	2.887	2.921	0.3473	0.9748	0.05858
	68643	3.271	3.189	0.8033	0.9817	0.18896
	72620	3.155	3.214	0.5407	0.9925	0.87014
	72839	3.156	3.121	0.5170	0.9785	0.10405
	76041	2.649	2.773	0.7593	0.9342	0.00009
	76780	3.268	3.312	0.6004	0.9880	0.51076
	79139	3.307	3.268	1.0211	0.9839	0.27030

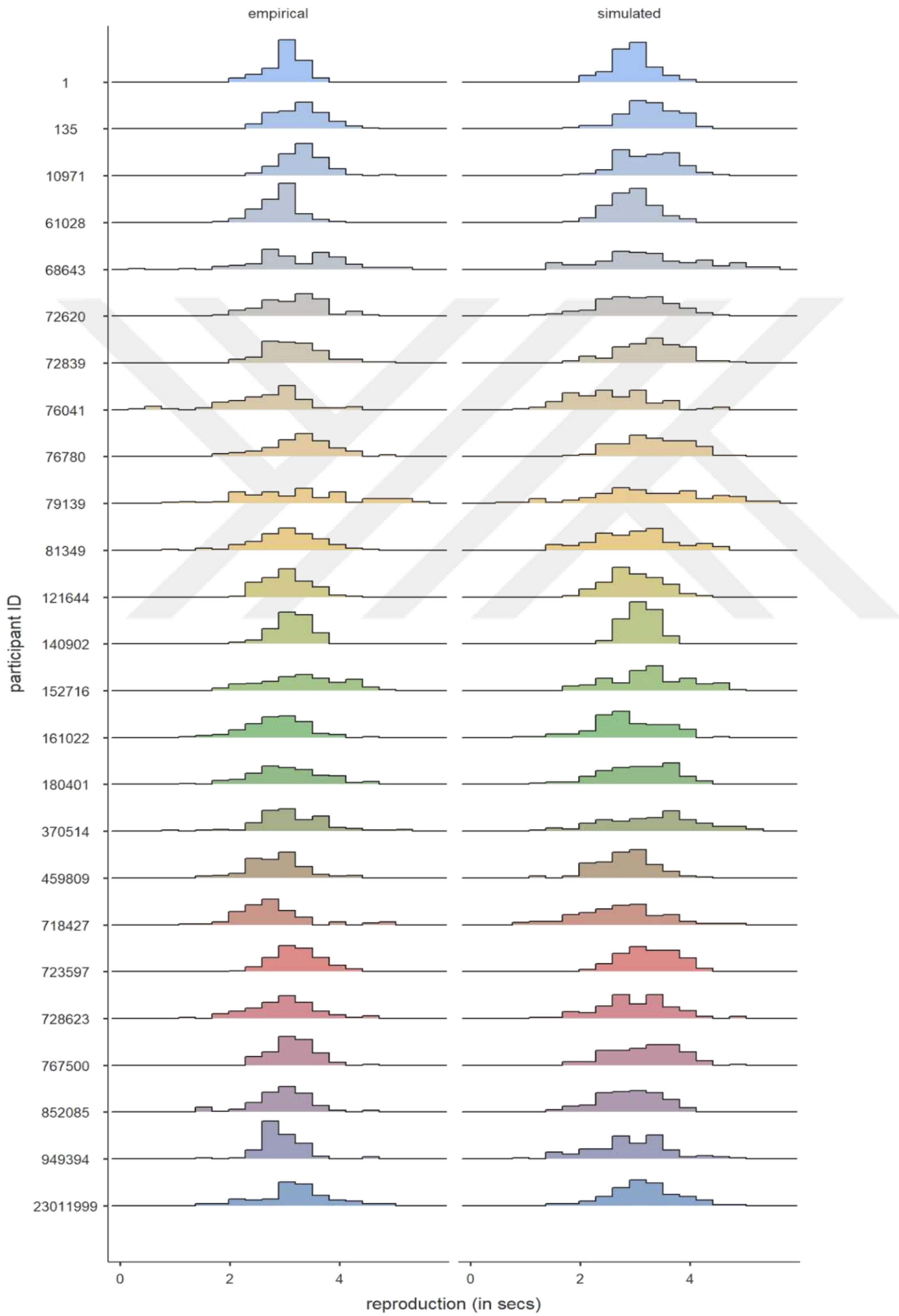
reproduction type	Participant ID	Mean	Median	SD	Shapiro-Wilk (W)	Shapiro-Wilk (p)
	81349	3.059	3.116	0.6061	0.9794	0.13195
	121644	3.023	3.004	0.4177	0.9745	0.05130
	140902	3.125	3.091	0.3128	0.9804	0.14795
	152716	3.297	3.289	0.7333	0.9826	0.21645
	161022	2.853	2.873	0.5690	0.9823	0.20423
	180401	3.046	2.988	0.6292	0.9944	0.95600
	370514	3.139	3.011	0.6899	0.9603	0.00451
	459809	2.867	2.888	0.5282	0.9782	0.09676
	718427	2.764	2.657	0.6472	0.8874	< .00001
	723597	3.226	3.206	0.4400	0.9918	0.80860
	728623	2.956	3.021	0.6204	0.9861	0.38688
	767500	3.164	3.171	0.3992	0.9894	0.64056
	852085	2.955	2.973	0.5630	0.9544	0.00186
	949394	2.941	2.870	0.4091	0.9127	< .00001
	23011999	3.178	3.210	0.6645	0.9812	0.17565
simulated	1	2.940	2.935	0.3959	0.9847	0.30450
	135	3.268	3.256	0.4903	0.9790	0.11176
	10971	3.230	3.250	0.5095	0.9905	0.70513
	61028	2.931	2.934	0.4276	0.9950	0.97399
	68643	3.262	3.153	0.9689	0.9776	0.08586
	72620	2.997	3.040	0.6611	0.9929	0.88677
	72839	3.254	3.286	0.6039	0.9875	0.47164
	76041	2.573	2.533	0.7266	0.9811	0.16586
	76780	3.326	3.284	0.6260	0.9896	0.63696
	79139	3.273	3.195	1.0664	0.9886	0.55713
	81349	2.996	3.032	0.7341	0.9888	0.56872
	121644	2.964	2.922	0.4713	0.9886	0.56233
	140902	3.091	3.081	0.2791	0.9890	0.58978
	152716	3.271	3.256	0.7275	0.9864	0.39829
	161022	2.845	2.729	0.6637	0.9850	0.32416
	180401	3.048	3.050	0.6480	0.9848	0.30833
	370514	3.315	3.395	0.8558	0.9931	0.89682
	459809	2.812	2.849	0.5281	0.9850	0.31587
	718427	2.714	2.758	0.7656	0.9946	0.96204
	723597	3.266	3.233	0.4942	0.9811	0.16107
	728623	3.018	3.026	0.6667	0.9928	0.87356

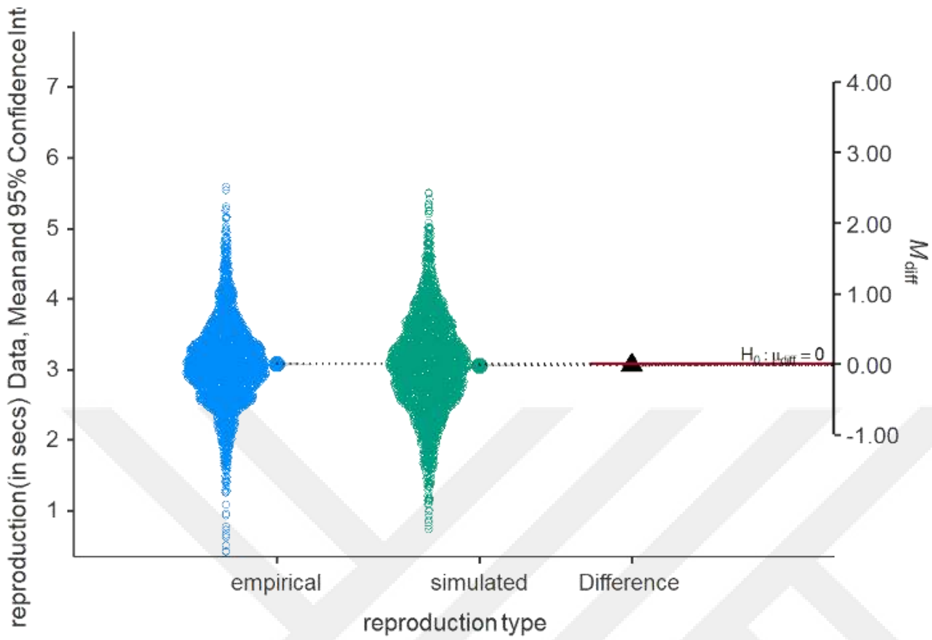
reproduction type	Participant ID	Mean	Median	SD	Shapiro-Wilk (W)	Shapiro-Wilk (p)
	767500	3.200	3.230	0.6170	0.9889	0.57815
	852085	2.886	2.904	0.5512	0.9875	0.48266
	949394	2.894	2.939	0.7306	0.9853	0.33538
	23011999	3.153	3.146	0.6429	0.9908	0.72527

Table S2.13. The comparison between the two distributions (i.e., empirical and simulated) across individual participants (where 1 and 0 in the most right column stands for alternative and null hypotheses, respectively)

participant ID	Kolmogorov-Smirnov statistics	*p*	hypothesis supported
1	0.218585859	0.01433053	**1**
135	0.076734694	0.922558618	0
10971	0.207959184	0.023411341	**1**
61028	0.155979381	0.165815869	0
68643	0.108571429	0.579394105	0
72620	0.155934343	0.170205384	0
72839	0.156666667	0.158142659	0
76041	0.194848485	0.039517449	**1**
76780	0.093939394	0.75183153	0
79139	0.076767677	0.920923104	0
81349	0.145979381	0.225647995	0
121644	0.121212121	0.436565645	0
140902	0.121212121	0.436565645	0
152716	0.063939394	0.984123954	0
161022	0.121212121	0.436565645	0
180401	0.09	0.794183771	0
370514	0.207272727	0.023579021	**1**
459809	0.13	0.343885712	0
718427	0.185714286	0.057617312	0
723597	0.138282828	0.276380257	0
728623	0.116969697	0.479277001	0
767500	0.185051546	0.060156417	0
852085	0.149144506	0.205059232	0
949394	0.194583333	0.04237887	**1**
23011999	0.113673469	0.519715894	0

Figure S2.1. The reproduction distributions across different sources (empirical and simulated) for per participant




reproduction(in secs) Data, Mean and 95% Confidence Int
7
6
5
4
3
2
1
4.00
3.00
2.00
1.00
0.00
-1.00
M diff
H0 : μdiff = 0
empirical
simulated
Difference
reproduction type