THE EFFECT OF SOCIAL INFLUENCE ON TEMPORAL AWARENESS

by

TUTKU ÖZTEL

Submitted in partial fulfillment of the requirements for the degree of Master of Arts in Psychology in the Graduate School of Social Sciences and Humanities of Koç University

July, 2019

Koç University

Graduate School of Social Sciences and Humanities

This is to certify that I have examined this copy of a master's thesis by

TUTKU ÖZTEL

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

and that any or all revisions required by the final

examining committee have been made.

Committee Members:

Prof. Fuat Balcı

Asst. Prof. Terry Eskenazi

Asst. Prof. Çağlar Akçay

H - Bahçekapılı

Asst. Prof. Hasan Galip Bahçekapılı

Prof. Laurence Conty

STATEMENT OF AUTHORSHIP

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

Signed Tutku Öztel

Abstract

A key aspect of metacognition is the ability to monitor performance (Flavell, 1979; Fleming & Dolan, 2012; Metcalfe & Shimamura, 1994). A recent line of work has shown that error monitoring ability captures not only the magnitude but also the direction of errors in the context of timing behavior, thereby pointing at the informationally rich metric composition of error monitoring (Akdoğan & Balcı, 2017; Duyan & Balcı, 2018; 2019). Previous studies have shown the enhancing effect of the feeling of being watched on physical state monitoring performance (Baltazar, 2014; Hazem, George, Baltazar & Conty, 2017). In the light of this work, we hypothesized that participants would best monitor their temporal errors in the being watched condition. Moreover, based on classical social facilitation theory (Zajonc, 1965; Cottrell, 1972), we further hypothesized the mastery levels would moderate this effect. Consistent with earlier work, in all four experiments, we found that participants can monitor the magnitude and direction of their timing errors irrespective of different response formats. We found the enhancing effect of being watched on temporal error monitoring performance only when the observer was a real person (instead of a picture) and watching the participants' performance in Experiment 3. However, we failed to replicate this finding in the subsequent experiment. We did not find the moderating effect of mastery level. Briefly, our results demonstrate that metric error monitoring is a very robust phenomenon, which is not necessarily sensitive to social presence or being watched.

Keywords: metacognition, temporal error monitoring, feeling of being watched, social facilitation

ÖZET

Üstbilişin en temel yönlerinden biri performans izleme (kendi performansının farkında olma) yetisinden ileri gelir (Flavell, 1979; Fleming & Dolan, 2012; Metcalfe & Shimamura, 1994). Yapılan son çalışmalar, hata izleme yetisinin zamansal bağlamda yapılan hataların yalnızca miktarını değil, aynı zamanda yönünü de kapsadığını göstermektedir (Akdoğan & Balcı, 2017; Duyan & Balcı, 2018; 2019). Bu bulgular, hata izleme yetisinin metrik açıdan zenginliğine işaret etmektedir (Akdoğan & Balcı, 2017; Duyan & Balcı, 2018; 2019). Önceki çalışmalar, izleniyor olma hissinin insanların fiziksel durumları hakkında çıkarımlar yapabilme yetilerini iyileştirdiğini göstermektedir (Baltazar, 2014; Hazem, George, Baltazar & Conty, 2017). Bu bilgi ışığında, katılımcıların zamansal hatalarını, başkaları tarafından izlenildikleri koşulda en iyi şekilde raporlayabileceklerini öngördük. Bunun yanında sosyal iyileştirme teorisi çerçevesinde (Zajonc, 1965; Cottrell, 1972), başkaları tarafından izleniyor olma hissinin zamansal hata izleme yetisi üzerindeki etkisinin, katılımcılar arası farklı görev performansı derecelerine göre değişeceğini öngördük. Yaptığımız dört deneyde, önceki çalışmalarla tutarlı olarak katılımcıların zamansal hatalarının miktarını ve yönünü izleyebildiklerini bulduk. Öte yandan, performansın başkaları tarafından izleniyor olmasının zamansal hata izleme performansı üzerindeki iyileştirici etkisini yalnızca gözlemcinin gerçek insan (gözlemciyi temsil eden fotoğraf uyaranları yerine) olduğu Deney 3'te bulduk. Ancak yaptığımız son deneyde (Deney 4) bu etkiyi replike edemedik. Bulgularımız, zamansal hata izlemenin sosyal uyaranlara hassas olmayan bir fenomen olduğuna işaret etmektedir.

Anahtar kelimeler: üstbiliş, zamansal hata izleme, izleniyor olma hissi, sosyal iyileştirme

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my thesis co-advisors Prof. Fuat Balcı and Asst. Prof. Terry Eskenazi for always supporting me on my way of progress and teaching me that it was neither swift nor easy. Genuine thanks to Prof. Fuat Balcı for his patience while training me and always encouraging me for bravery to try again when I make mistakes.

Additionally, I would like to thank the rest of my committee members Hasan Galip Bahçekapılı, Çağlar Akçay and Laurence Conty for their intellectual contributions with their valuable feedback on my thesis.

I want to thank all my dearest friends in Koç University Psychology graduate program, especially Bihter Akyol, Esin Türkakın, Anıl Şafak Kaçar, Sümeyye Koşkulu, Zeynep Aslan, Demet Özer, Hazal Kartalkanat and Arın Korkmaz for always supporting me, making me laugh when I felt upset and stressed, and being sisters (and brothers) to me. I want to especially thank Bihter, Esin and Mertcan, from whom I learned a lot, for their intellectual contributions to this thesis. Without you, this journey would be much harder for me. Many thanks to Alper Mert for his help with data collection. I also want to thank my dear friend Semih Cihan for cheering me up with his jokes, keeping my mood up with his good company and for always saying “yes” whenever I asked him out for a drink.

Genuine thanks go to my dearest one Zeynep Demirbilek for all the fun we have had together since 2013, the first year of ours as undergraduate students. My dear cousins Aslı, Zeynep and Ekin, thanks for helping me to prepare my stimuli in Autocad and being more than sisters to me.

Lastly, I want to thank my parents and grandmother for their warmth and sharing their wisdom with me in every period of my life. I love you beyond words.

I dedicate my masters thesis to Assoc. Prof. Engin Arık who has been a mentor to me with his wise advices and great guidance since the beginning of my undergraduate years.

Table of Contents

List of Figures

1. Introduction

1.1. General Overview of Metacognition and Temporal Error Monitoring

Metacognition is defined as “cognition about cognition” (Flavell, 1976; Fleming, 2012; Overgaard & Sandberg, 2012) that includes the individuals’ ability to monitor their performance even in the absence of an external feedback (Akdoğan & Balcı, 2017; Yeung & Summerfield, 2012; Fleming & Daw, 2017). Metacognition has been proposed to play an important role in learning, problem solving, cognitive control, decision making, planning, social interactions, and action (Flavell, 1976, 1979; Brown, 1987; Schraw, Krippen, & Hartley, 2006; Fleming, 2012, 2014; Akdogan & Balcı, 2017; Yeung & Summerfield, 2012; Frith, 2012). For instance, error monitoring aspect of metacognitive knowledge is known to be utilized for adapting/correcting future behavior to minimize subsequent errors (Orr & Carrasco, 2011). Thus, the metacognitive ability overall appears to be fundamental for healthy cognition and adaptive planning and action control.

Studies that investigate metacognition in humans typically test participants at two points of interest; a) first-order task performance and b) confidence rating. For instance, in the first-order task of a two-alternative forced choice (2AFC) paradigm participants are asked to differentiate the target stimulus from noise (i.e., target irrelevant stimuli), and in the second-order task, participants are asked to report their confidence level regarding their own first order task performance. The latter aspect of performance constitutes metacognitive evaluation (e.g., Maniscalco, 2012; Yeung & Summerfield, 2012). However, this classical binary approach to evaluate the metacognitive abilities limits capturing its full information content as well as the ecological validity of the studies. In many real life situations, the decisions we make and errors we commit carry much more than binary information. For instance, we often make decisions based on a continuous dimension in which case one can talk about the “magnitude of errors”, which can be critical to “monitor” for adaptability (e.g., in motor planning). Only recently though, error monitoring and cognitive abilities in the metric domains (such as time, space and number) have started to be investigated (Akdoğan & Balcı, 2017; Duyan & Balcı, 2018; 2019; under review; Kononowicz et al., 2018). Interval timing with its statistical features indeed constitutes a fertile ground for the study metric error monitoring.

Interval timing is one of the most prominent abilities that are crucial for a variety of functions including action control, decision making, learning and signaling (Buhusi & Meck, 2005). Although humans and other vertebrate species are capable of estimating a given time interval (i.e., target duration) with high accuracy, there is a margin of error in their temporal estimations that is manifested as trial-to-trial metric variability in timing behavior (Akdoğan & Balcı, 2017; Çavdaroğlu, Zeki & Balcı, 2014; Gibbon, 1977; Gibbon, Church & Meck, 1984). Thus, keeping track of magnitude and the direction of the errors in a temporal setting can be critical for many domains (e.g., temporal risk assessment; Balcı et al., 2011). In this respect, a recent study by Akdoğan and Balcı (2017) has shown that humans are aware of the magnitude and direction of errors in their timing behavior (i.e., temporal reproductions - see also Kononowicz et al., 2018; Doenyas et al., 2019). This ability has been referred to as "temporal error monitoring" and "metric error monitoring" after the studies that showed that the same ability generalizes to spatial and numerical productions and estimations (Duyan & Balcı, 2018; 2019; under review). Together, these studies indicate the ability of humans to keep track of errors in their timing behavior and constitute a conceptual contribution to the literature by highlighting an informationally rich metric composition of error monitoring.

1.2. Social Influences on First and Second Order Task Performances

Metacognitive evaluations, as also in the case of first-order decisions (e.g., Germar et al., 2014), are subject to social influences (Eskenazi et al., 2016; Jacquot et al., 2001; De Carvalho & Yuwaza, 2001). In this respect, implicit as well as explicit social feedback (i.e., agreement or disagreement) have been found to influence the metacognitive evaluations (e.g., Eskenazi et al., 2016; Jacquot et al., 2015; Pescetelli et al., 2016). For example, Eskenazi and colleagues (2016) have found that unreliable social cues (i.e., feedback about first-order task performance) had a detrimental effect on metacognitive accuracy even though participants were told not to take those cues into account to guide their performance.

Social influence is not restricted to explicit social feedback. Another kind of social influence that can affect performance is the “feeling of being watched”. Because being watched involves being the object of someone’s attention (Conty, 2016), it can trigger evaluation apprehension which is defined as the worry of being negatively evaluated by others (Cottrell, 1972). Depending on the task one is doing, feeling of being watched and the associated sense of evaluation apprehension can both improve or impair first order performance. Being watched can disrupt performance if one gets extremely nervous about being poorly evaluated which may interfere with performance due to distraction. For example, Conty et al (2010) found the impairing effect of direct eye gaze, which induced a feeling of being watched on Stroop task performance. That is, the Stroop interference was most pronounced when participants felt like their performance was the subject of someone’s attention. A similar finding was recently reported with an incentive-based motor task, where participants’ performance was enhanced when they were being watched (Chib, Adachi & O’Doherty, 2018).

As the reflexive attention orientation effect of the gaze of others is well documented (e.g., Friesen & Kingstone 1998), the feeling of being watched that is induced by direct gaze can orient the attention to the self (i.e. self-focused attention), and thus, increase self awareness (Conty et al., 2016). In line with this view, Duval and Wicklund (1972) claimed that evaluation apprehension caused by social presence makes people focus more on themselves (self-awareness as mentioned in Conty et al., 2016) to avoid poor evaluations on part of the audience. Confirming Duval and Wicklund’s (1972) claim, Baltazar (2014) has shown that mere virtual social presence (i.e., averted gaze) is not enough per se to enhance the ability to monitor physical states, without inducing a feeling of being watched (i.e., via direct gaze). Additionally, a more recent study by Hazem, George, Baltazar, and Conty (2017) found that this effect was not specific to direct gaze, by showing an implicitly induced feeling of being watched without requiring direct gaze resulted in the same effect. Taking these two findings into account, it can be claimed that although direct gaze is sufficient per se (as in Baltazar et al., 2014), it is not necessary to affect performance (Hazem et al., 2017). In sum, social presence seems to affect performance through the induction of a feeling of being watched.

Although being watched has different effects on first order performance due to evaluation apprehension, it can improve performance in the second order decisions via inducing self-focused attention. There are studies showing the effect of social feedback on metacognition (e.g., Eskenazi et al., 2016); however, the effect of the feeling of being watched on performance monitoring in the metric domain has not been explored. In an attempt to fill this gap in the literature, in the current study, we investigated if the effect of the feeling of being watched on physical state monitoring could also apply to the monitoring of cognitive states (i.e. metacognition), specifically manifested in the form of metric error monitoring. Regarding this, in light of previous work (see: Baltazar et al., 2014), we hypothesized that the feeling of being watched would increase metric error monitoring performance.

We further investigated if the first order mastery level can moderate the effect of the feeling of being watched on temporal error monitoring performance. Earlier work on social facilitation suggested that the effect of evaluation apprehension that is induced by the mere presence of others/feeling of being watched on performance can be different for different task mastery levels (Zajonc, 1965; Cottell, 1972; Triplett, 1898; for review see: Aiello & Douthitt, 2001). Interval timing behavior lends itself in a conceptually complete fashion for such an analysis since the degree of the trial-to-trial variation in timing behavior, which would constitute the timing errors for metric error monitoring in individual trials, can also be used to quantify the overall mastery level of the participants. Thus, we hypothesized that the social influence would enhance temporal error monitoring performance as the timing variability decreases, and it would disrupt performance as the timing variability increases.

We conducted four different experiments to test our hypotheses. Two of these experiments used virtual social presence stimuli and two experiments used an actual social presence. In Experiment 1, we used four different types of visual stimuli (direct/averted x social/nonsocial) to examine how the effect of the feeling of being watched on temporal error monitoring performance that is induced with direct eye gaze differed from the effect of both mere social presence (i.e., averted gaze) and a nonsocial object. In Experiment 2, we discarded the averted stimuli to directly compare the effect direct social and nonsocial stimuli on temporal error monitoring performance. In Experiment 3, we compared participants' temporal error monitoring performance in a more realistic social scenario (higher ecological validity, Hamilton, 2016) where participants' performance was being watched by a female experimenter as they performed the task versus they performed the task alone. Lastly, in Experiment 4, we attempted to replicate the findings that we obtained in Experiment 3 with a male experimenter.

2. Experiment 1

Methods

2.1. Participants

Thirty-two healthy participants (25 females, 1 left handed, M_{age} = 21.8, SD_{age} = 2.88) participated in Experiment 1 in return for one extra course credit. All participants provided an informed consent form prior to testing. The experiment was approved by the local ethics committee at Koç University.

2.2. Stimuli and Apparatus

We used Matlab with Psychtoolbox (Brainard, 1997) extension for stimulus presentation and data collection. The stimuli were presented on a 21.5-inch monitor on an IMac computer.

As social stimuli, we used eight different (four females and four males) neutral face pictures. As in study conducted by Baltazar et al (2014), pictures had three different orientation consisting of frontal, right averted and left averted (direct face to induce a feeling of being watched via eye contact, and averted faces to induce a mere social presence without necessitating a feeling of being watched). Eight different armchair pictures were used as non-social stimuli. Just like the social visual stimuli, the non-social visual stimuli had three different orientation consisting of frontal, right averted and left averted. We created the nonsocial visual stimuli from Autocad 2014. The proportion of averted and direct stimuli was the same (i.e., 48 direct, 48 averted, with equal left and right proportion for the averted stimuli). All the visual stimuli were gray scaled and the brightness and luminance were kept constant.

2.3. Procedure

Each experimental condition (social-direct, social-averted, nonsocial-direct, nonsocial-averted) was tested for 48 trials. Thus, the experiment consisted of a total of 48 x 4 = 192 trials. The left and right averted visual stimuli were presented with an equal probability (i.e., 24 trials each).

Participants were seated approximately 60 cm away from the monitor. At the beginning of the task, participants were presented with the target duration of 2300 milliseconds with a noisy patch on the center of the screen. Participants were reminded of this duration after each fifth reproduction trial. After observing the target duration, participants initiated their reproduction by pressing the space button, and they terminated their reproduction when they thought the same amount of time with the target duration had elapsed with a second press. Participants made their reproductions with the visual stimuli that are described in the Stimuli and Apparatus section. The order of the visual stimuli (i.e., experimental condition) was randomized between trials. Following the temporal reproduction, participants rated their level of confidence regarding their reproduction performance on a 100-point slider scale (-100 for low confidence and 100 for high confidence). The left-end of the slider scale was marked by "I am totally unsure" (in Turkish) and the right-end was marked by "I am totally sure" (in Turkish) for confidence judgments.

After the confidence judgements, participants indicated how short or long they thought their temporal reproduction was compared to the target duration. This error magnitude judgment was again indicated by the participants on a slider scale (-100 for too short, 0 for on target, 100 for too long). The left-end of the same scale was marked by "I reproduced too short" (in Turkish) and the right-end was marked by "I reproduced too long" (in Turkish).

The maximum and minimum values of the slider scales were visible to the participants. The confidence and the magnitude judgments were indicated via a standard computer mouse. Participants were not provided with any feedback regarding either their reproduction performance or error estimation performance. They were encouraged to use the whole scale and were asked not to count or utilize any chronometric strategy to estimate the elapsed time throughout the experiment (Rattat & Droit-Volet, 2012). Experiment 1 lasted approximately 30 minutes to complete.

Data filtering

Prior to analysis, we removed all the trials that are below -3 and above 3 z-score transformed reproduced duration (on average, approximately 1.5% of the whole dataset for all four experiments). We used the same filtering procedure in all four experiments. We conducted all our analyses in Jamovi (The jamovi project, 2019; R Core Team, 2018) using the GAMLj library (Gallucci, 2019).

2.4. Data Analysis

2.4.1. Confidence judgements as outcome variable

For the confidence judgement, prior to analysis, we z-score transformed the reproduced durations of the participants. To address our first research question if social presence/being watched can affect (non-directional) temporal error monitoring performance in terms of confidence judgements, we performed linear mixed effects analysis, using the raw confidence judgements (i.e., ranging between -100 and 100) as the outcome variable. For this, we included the absolute values of z-score transformed reproduced durations, the experimental condition (stimulus type = social or nonsocial x stimulus direction = direct or averted, as four levels of "experimental condition" factor), and their interaction term as fixed effects on the slope, participants as random effect on the intercept. We used the absolute values of z-score transformed reproduced durations to account for the non-directional errors of the participants.

To address our second research question if mastery levels can moderate the effect of social presence/being watched on temporal error monitoring performance, we included CV in our model as a continuous covariate, with all two- and three-way interaction terms loaded on slope as fixed effects. We kept the participants as random effect on the intercept. For hypothesis testing and model selection that explains confidence judgements best, we compared the Bayesian Inference Criterion (BIC) scores of all possible models (i.e., the null model with only random intercept across subjects, all models with only the main effects of the fixed factors, all models with interaction terms) using "BIC" function in RStudio (RStudio Team, 2015). This model comparison determined the best model that explained our data as the reproduction duration only model (ΔBIC = 19.13 with the second best model, which included the main effects of absolute values of z-score transformed reproduced durations, experimental condition, and CV as fixed effects).

Model 1.1.:

confidence judgements ~ absolute values of z-score reproduced durations + (1|participants)

where ~ stands for "predicted from", and 1|participants stands for "random intercept across participants".

Although the two (i.e., experimental condition and absolute values of reproduction) and three-way (i.e., CV included to the previous model) interaction models are the models that directly address our research questions, as our model selection yielded the best model that explains confidence judgements as the Model 1.1. (showing that other variables do not have an effect), we do not report their outputs in the Results section. Instead, we report them in the Supplementary Materials section.

2.4.2. Directional error magnitude judgement as outcome variable

We performed the same analysis for directional error magnitude judgements as the outcome variable, except that we used raw values (instead of absolute values) of z-score transformed reproduced durations to account for the direction of as well as the magnitude of the temporal errors. To address our first research question, we first included experimental condition and the z-score transformed reproduced durations and their interaction term as fixed effects on slope, and participants as random effect on the intercept. We included raw scores of directional error magnitude judgements (i.e., ranging between -100 and 100) as outcome variable.

As we did for confidence judgements, we included CV in our model as a continuous covariate, with all two- and three-way interaction terms loaded on slope as fixed effects to address our second research question. We kept the participants as random effect on the intercept. To select the best model that explains directional error magnitude judgements, we compared the BIC scores of all possible model (i.e., a null model with only random intercept across subjects, all models with only the main effects of the fixed factors, all models with interaction terms - CV included and CV excluded as continuous covariate) using “BIC” function in RStudio (RStudio Team, 2015). This model comparison determined the best model that explained our data as the model with three-way interaction between z-score transformed reproduced durations, experimental condition, and CV (ΔBIC = 42.76 with the second best model, which included only the z-score transformed reproduced durations as the fixed effect).

Model 1.2.:

directional error magnitude judgements ~ z-score reproduced durations * experimental condition * CV + (1|participants)

where ~ stands for “predicted from”, and 1|participants stands for “random intercept across participants”. The model depicted above consisted of all lower terms.

As Model 1.2. is the best model that explains the directional error magnitude judgements, in the Results section, we only report the results of this model. However, we also report the model with two-way (ie., experimental condition and reproduced duration) interaction in the Supplementary Materials.

2.5. Results

2.5.1. Comparison of reproduced durations and CVs across experimental conditions

In order to investigate the potential effect of experimental condition on participants’ reproduced durations, we performed a 2 x 2 Repeated Measures Analysis of Variance (ANOVA) using the type (social and nonsocial), and the direction of the stimulus (direct and averted) as independent variables on raw reproduced durations. This analysis revealed no statistically significant main effect of either stimulus type ($F(1,31) = 1.823$, $p > 0.05$, partial $\eta^2 = 0.056$; $BF_{01} = 1.927$), or stimulus direction on reproduced durations ($F(1,31) = 0.097$, $p > 0.05$, partial $\eta^2 = 0.003$; $BF_{01} = 5.091$). The two way interaction between the stimulus type and the direction on reproduced durations was also not significant ($F(1,31) = 3.3e\text{-}5$, $p > 0.05$, partial $\eta^2 = 1e\text{-}6$; $BF_{01} = 41.603$).

We performed identical analysis using CV values as dependent variable. This analysis revealed no statistically significant main effect of either stimulus type ($F(1,31) = 1.51$, $p > 0.05$, partial $\eta^2 = 0.046$; $BF_{01} = 2.78$), or stimulus direction on reproduced durations ($F(1,31)$

= 2.46, $p > 0.05$, partial $\eta^2 = 0.074$; $BF_{01} = 3.05$). The two way interaction between the stimulus type and the direction on reproduced durations was also not significant ($F(1,31) = 0.416$, $p > 0.05$, partial $\eta^2 = 0.13$; $BF_{01} = 0.084$). Thus, the experimental conditions did not have an effect on participants' mean reproductions or CVs.

2.5.2. Analysis of Metric Error Monitoring Performance

2.5.2.1. Confidence judgements as outcome variable

For the best model that explains our data (i.e., the model that includes only reproduced durations as fixed effect parameter), the main effect of the absolute values of z-score transformed reproduced durations on confidence judgements was statistically significant (ß = -15.8, SE = 0.845, $p < 0.001$), revealing a negative linear relationship between the absolute deviations of the reproduced durations and confidence judgements (see Figure 1.1).

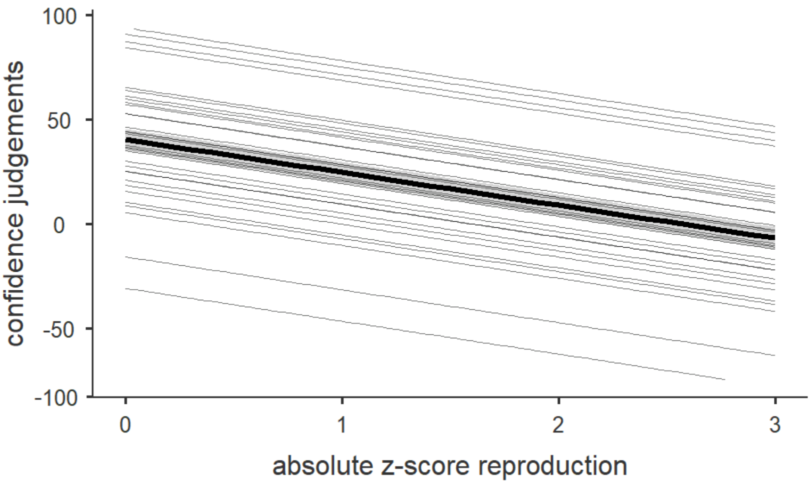


Figure 1.1: Linear relationship between the absolute deviations of the reproduced durations and the confidence judgements. Light gray lines represent individual slopes, whereas bold black line represents average slope of the sample (N=32). Error bars represent standard error (SE) of the estimation.

2.5.2.2. Directional error magnitude judgements as outcome variable

For the best model that explained our data (i.e., which included CV, and z-score transformed reproduced durations as continuous covariate, and experimental condition as categorical fixed effect, including the highest three-way, and all lower interaction terms), the main effect of z-score transformed reproduced durations on directional error magnitude judgements was statistically significant (ß = 9.398, *SE* = 0.327, $p < 0.001$). Moreover, the three-way interaction between the z-score reproduced durations, experimental condition and CV on directional error magnitude judgements was statistically significant ($F(3,6041.1) = 3.003$, $p < 0.05$). Post-hoc comparisons revealed a statistical trend for a slope difference between social direct and social averted (ß$_{social\ direct\text{-}social\ averted}$ = 17.142, *SE* = 8.979, *p* = 0.056). Figure 1.2. shows the linear relationship between the z-score transformed reproduced durations, and the directional error magnitude judgements across experimental conditions in different levels of CV.

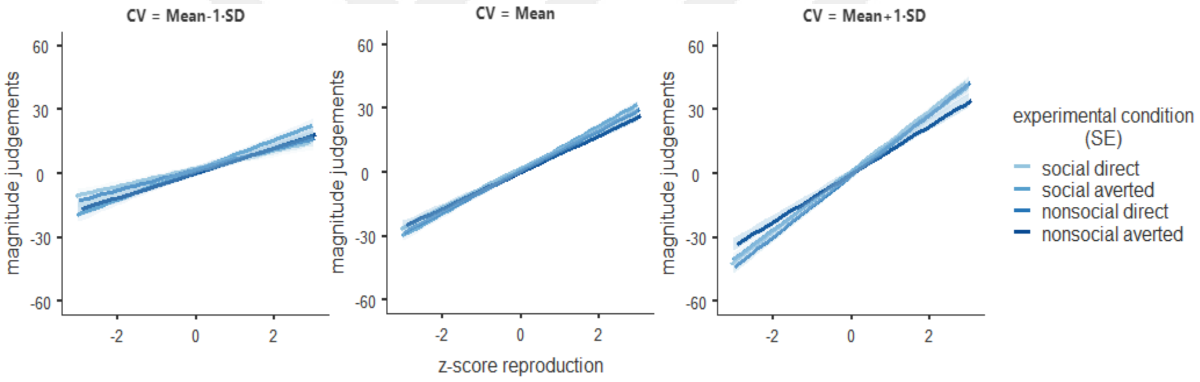


Figure 1.2.: Linear relationship between z-score transformed reproduced durations and directional error magnitude judgements across experimental conditions in three different levels of mastery. Error bars represent standard error (SE).

We also split our 4-level categorical fixed factor in a way that would address the main effects of the stimulus type (social or non-social) and direction (direct or averted; 2x2 analysis design). This way of addressing the main effects of the stimulus characteristics on our two dependent variables (confidence, and directional error magnitude judgements) yielded no significant effect of stimulus type (i.e., social or nonsocial), and stimulus direction (i.e., direct or averted; all $p > 0.05$), as the best models that explained both of the dependent variables were the models that only included reproduced durations as fixed effect.

2.6. Interim Discussion

In order to address our research questions whether being watched improves temporal error monitoring ability, and if mastery level can moderate this effect, we compared all our models that include all possible combinations of interactions and main effects of our independent variables for two outcome variables (i.e., confidence and directional error magnitude judgements) separately. As model comparison results revealed, among all possible models which included confidence judgements as outcome variable, we found only the main effect of reproduced duration on confidence judgements. Although this finding is in line with the previous work that pointed to a metric error monitoring ability (e.g., Akdoğan & Balcı, 2017; Kononowicz et al., 2018; also see: Duyan & Balcı, 2018; 2019; Doenyas et al., 2019), our social/gaze manipulation did not have an effect on this performance. Moreover, mastery levels did not moderate the effect of social presence (direct and averted gaze) conditions on confidence judgements.

As in the case of confidence judgements, we found a main effect of reproduced duration on directional error magnitude judgements, which suggests that participants could keep track of their directional temporal error magnitude. Our model yielded no significant interaction between the experimental conditions, and reproduced durations, indicating that the directional error temporal magnitude monitoring ability did not change across different experimental conditions.

In order to address our second research question regarding the moderation of the effect of being watched on error magnitude, and direction monitoring performance by the mastery level, we further investigated the post-hoc comparisons for our three-way interaction term in a way that addresses our theoretical interests. Contradicting our hypothesis, post-hoc comparisons pointed to a marginally significant trend of better error direction and magnitude estimation performance in the direct gaze condition compared to averted gaze condition as the mastery level decreased. Taken together, results that we obtained in Experiment 1 suggest that the social facilitation effect does not apply to accuracy of the confidence ratings and directional error magnitude judgements.

To directly test the effect of social presence on temporal error monitoring performance, in Experiment 2, we discarded all the averted stimuli from our experimental design. Moreover, we used a response protocol that better corresponded to that used in earlier work (Akdoğan & Balcı, 2017).

3. Experiment 2

Methods

3.1. Participants

41 participants (27 females, 35 right handed, M_{age} = 20.2, SD_{age} = ± 1.24) from Koç University participated in Experiment 2 (note that this experiment was chronologically conducted after Experiment 3). All participants received one extra credit in return for their participation in the experiment. We included data from 37 participants in the analysis as the health demographics of four participants did not meet inclusion criteria (i.e., 3 of the 4 excluded participants reported use of psychiatric drugs, one was a non-Turkish native speaker).

3.2. Stimuli and Apparatus

All the stimuli and apparatus used in Experiment 2 remained the same as in Experiment 1 except the fact that all the averted stimuli (i.e., averted faces and averted chair pictures) were discarded from the experimental design. Consequently, the number of trials for per experimental condition was doubled (96 trials per condition). In addition, participants were given 5-likert scale (1-”does not apply to me at all”, 5-”completely applies to me”) The Turkish Adaptation of Brief Fear of Negative Evaluation Scale (Cronbach’s α = 0.84, Çetin, Doğan, & Sapmaz, 2010; for the original version of the scale, see: Leary, 1983) after the experiment was completed. The Brief Fear of Negative Evaluation Scale aims to measure the individual level of anxiety towards being poorly evaluated by others (Leary, 1983; Çetin, Doğan, & Sapmaz, 2010). The scale includes items such as *"Sometimes I think I am too concerned with what other people think"* with its 2nd, 7th and 10th items reversed. The 4th item was not included in the analysis (Çetin, Doğan, & Sapmaz, 2010).

3.3. Procedure

Before the experiment started, participants were provided with an informed consent form and a health screen form. All the procedure was kept the same as in Experiment 1. Different from Experiment 1, confidence and short-long judgements were collected from a mechanical keyboard, rather than a continuous slider scale (with mouse). In this respect, for the confidence judgements, 1, 2, and 3 buttons were available (1-"I am not confident at all", 2-"I am moderately confident", 3-"I am completely confident"). For the short and long judgements, participants were asked to use V and N keys respectively (V for "short", N for "long"). All the key descriptions were presented on the screen until a key press was recorded by the computer. Participants were encouraged to use the full scale for confidence ratings. The experiment took for around 40 minutes to complete.

3.4. Data Analysis

As we did in Experiment 1, first, we z-score transformed the participants' reproduced durations. Next, we concatenated the confidence, and the short-long judgements in order to investigate the error magnitude as well as the error direction by re-coding the data as the following: If the participants reported their confidence as 1 and judged their reproduced duration as shorter than the target duration, we re-coded the response as "-3" (minus sign for the direction of the perceived error, in this case, "short"). On the other hand, if the participants reported their confidence as 3 and judged their reproduced duration as longer than the target duration, we re-coded the response as "1" (positive sign for the direction of the perceived error, in this case, "longer"). This resulted in recoded data that ranged between -3 and 3. After recoding the data as described above, we performed linear mixed effects analysis to capture whole data in a single model using the re-coded confidence judgements as outcome variable.

Finally, we followed the same hypothesis testing and model selection procedure as we did for directional error magnitude judgements in Experiment 1 (i.e., using z-score transformed reproduced durations including the directionality information of the temporal error), using the rescaled confidence judgements as outcome variable. Our model comparison determined the best model that explained our data as the reproduced duration only model as depicted in Model 2 (ΔBIC = 7.31 with the second best model, which included the main effects of z-score transformed reproduced durations and experimental condition as fixed effects).

Model 2:

rescaled confidence judgements ~ z-score reproduced durations + (1|participants)

where ~ stands for "predicted from", and 1|participants stands for "random intercept across participants".

We report the results of interaction models that directly address our research questions in the Supplementary Materials. The exploratory analysis we performed using the Fear of Negative Evaluation scores can also be found in the Supplementary Materials.

3.5. Results

3.5.1. Comparison of reproduced durations and CVs across experimental conditions

To ensure that the experimental condition did not have an effect on participants' reproductions, we compared the mean reproduced durations of the participants across conditions. Paired sample t-test revealed no significant difference between the conditions in terms of the reproduced durations ($t(36) = 0.208$, $p > 0.05$, $SD = 0.1007$, $BF_{01} = 5.543$). We performed an identical analysis using CV values as dependent variables. Paired sample t-test revealed no significant difference between the conditions in terms of CV values ($t(36) = -0.614$, *p* >0.05, *SD* = 0.0422, BF_{01} = 4.744). Thus, our experimental manipulation did not affect the reproduced durations or the CV values of the participants throughout the experiment.

3.5.2. Analysis of Metric Error Monitoring Performance

For the best model that explains our data (i.e., the model that includes only reproduced durations as fixed effect parameter), the main effect of z-score transformed reproduced durations was statistically significant (ß = 0.678, *SE* = 0.0246, *p* < 0.001), revealing a positive linear relationship between the absolute deviations of the reproduced durations and confidence (and the directionality) judgements (see Figure 2).

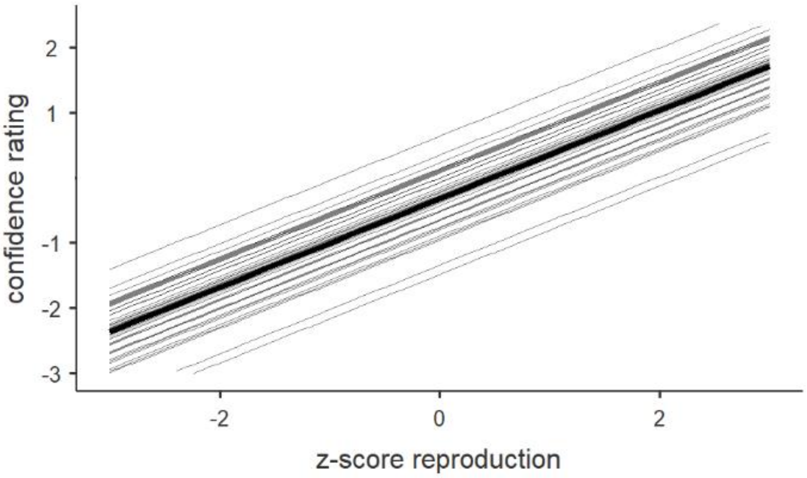


Figure 2.: Linear relationship between the z-score transformed reproduced durations and concatenated confidence, and error directionality judgements (the values on y-axis that are high in absolute magnitude represent low confidence and presumably high error magnitude). Minus and plus signs represent the reported error direction judgements (minus for “short”, plus for “long”). Light gray lines represent individual slopes, whereas bold black line represents average slope of the sample (N=37). Error bars represent standard error (SE) of the estimation.

Note that as the re-coding scheme that we followed for our analysis did not include 0, we performed identical analysis described in Data Analysis for Experiment 2 by re-coding the confidence and error directionality judgements as ranging between 1 (too short, low confidence) and 6 (too long, low confidence). Again, results revealed the best model that explains the re-coded confidence judgements as depicted in Model 2 (ΔBIC = 5.00 with the second best model, which included only the main effects of z-score reproduction, experimental condition, and CV; ß = 0.519, *SE* = 0.0192, *p* < 0.001).

3.6. Interim Discussion

To address our research questions, we followed the same model comparison procedure as in Experiment 1 to select the best model that explains our data. As model comparison results revealed, among all possible models which included confidence judgements as outcome variable, we found only the main effect of reproduced duration on confidence (which also includes the error direction) judgements. This result is a clear replication of Experiment 1 and the previous work in the literature, suggesting that participants could correctly monitor their temporal error directions and magnitude. However, although participants could correctly match their confidence and error direction judgements to their temporal errors as in Experiment 1, this ability did not improve (nor decline) in the direct gaze condition. Additionally, the mastery level of the participants had no moderating effect. Thus, as in Experiment 1, our hypotheses regarding the enhancing effect of direct gaze on temporal error monitoring performance, and the moderating effect of mastery level were refuted.

The results that indicate no effect of direct gaze condition (compared to control conditions) on temporal error monitoring performance may have been due to the lack of effectiveness of our manipulation. Taking this possibility into account, and in order to address the findings of Hazem et al. (2017) regarding the non-necessariness of the direct eye contact, in the next experiment, we manipulated the effect of being watched on temporal error monitoring performance in a more realistic scenario (i.e., which increases the ecological validity of the manipulation; Hamilton, 2016). Following these, participants completed the task under the observation of the experimenter (or alone).

4. Experiment 3

Methods

4.1. Participants

30 participants (22 female, 25 right handed, M_{age} = 22.03, SD_{age} = 3.69) from Koç University participated in this experiment for 1 extra course credit in return for their participation.

4.2. Stimuli, Apparatus and Procedure

Procedure and apparatus in the Experiment 3 remained the same as in Experiment 1. Differently from Experiment 1, in the confidence and magnitude judgement phases, the cursor appeared in a random place on the line in order to avoid biasing the participants' response tendency to the midpoint of the line. We also discarded the virtual social and non-social stimuli (i.e., face and armchair pictures) from the experimental procedure. Instead, to induce a social presence, we randomly assigned participants in social and nonsocial experimental conditions. In social condition, a female experimenter (first author) instructed the participant that she would be in the room throughout the experiment and watch the screen as they perform the task to evaluate their performance. In the non-social condition (i.e., control condition) however, the experimenter left the room after giving the instructions, and the participants performed the task on their own. The experiment lasted around 35 minutes to complete.

4.3. Data Analysis

4.3.1. Confidence judgements as outcome variable

We performed linear mixed effects analysis to capture all data in a single model. Again, we followed the same procedure as we did in Experiment 1 for the hypothesis testing and the selection of the best model that explained the variability in the confidence judgements. For confidence judgements, our model comparison determined the best model that explained confidence judgements as the model with the 2-way interaction term between the absolute z-score transformed reproduction and experimental condition (ΔBIC = 0.94 with the second best model, which only included the main effect of absolute values of z-score transformed reproduced durations as fixed effect).

Model 3.1.:

confidence ~ absolute z-score transformed reproduction * experimental condition + (1|participants).

where ~ stands for "predicted from" and 1|participants stands for "random intercept across participants". The model depicted above consisted of all lower terms. We also report the results of the three-way interaction model output in the Supplementary Materials section.

4.3.2. Directional error magnitude judgements as outcome variable

For hypothesis testing and model selection, we performed the same analysis as we did in Experiment 1 for directional error magnitude judgements as outcome variable. Our model comparison determined the best model that explained the variability in the directional error magnitude judgements as depicted in Model 3.2(ΔBIC = 8.51 with the second best model, which included only the main effects of z-score transformed reproduced durations and experimental condition as fixed effects).

Model 3.2.:

directional error magnitude judgements ~ z-score reproduced durations + (1|participants)

where ~ stands for "predicted from" and 1|participants stands for "random intercept across participants".

In addition to the best models that explained the variability in directional error magnitude judgements, we report the models that directly address our research questions in Supplementary Materials.

4.4. Results

4.4.1. Comparison of reproduced durations and CVs across experimental conditions

In order to investigate if the experimental condition had an effect on reproduced durations, we compared the mean reproduced durations of the participants across conditions. Independent samples t-test revealed no statistically significant difference between the two experimental conditions in terms of reproduced durations ($t(28) = -0.0107$, $p > 0.05$, $SD = 0.4117$; $BF_{01} = 2.904$). We performed identical analysis using CV values as dependent variable. Independent samples t-test revealed no statistically significant difference between the two experimental conditions in terms of CV values ($t(28) = -0.7499$, $p > 0.05$, $SD = 0.0491$; $BF_{01} = 2.346$). Thus, the experimental manipulation did not have an effect on participants' reproduced durations or CV values.

4.4.2. Analysis of Metric Error Monitoring Performance

4.4.2.1. Confidence judgements as outcome variable

The best model that explains the variability in confidence judgements (i.e., the model that includes absolute values of z-score transformed reproduced duration*experimental condition interaction term) revealed a significant main effect of absolute values of z-score transformed reproduced durations on confidence judgements (ß = -15.19, $SE = 0.927$, $p < 0.001$), indicating a negative linear relationship between the reproduced durations and confidence judgements. Additionally, the two-way interaction between the fixed effects (i.e., experimental condition and absolute values of z-score transformed reproductions) on confidence judgements were significant ($F(1,5918.3) = 18.25$, $p < 0.001$, ß$_{\textit{social - nonsocial}}$ = -7.92, $SE = 1.85$, $p < 0.001$). This result demonstrates a statistically significant slope difference between the two experimental conditions. In order to address which of the two experimental manipulation's effect was closer to the "ideal observer's performance", we compared the simple slopes with the ideal-performance slope, which is -66.66 (=200/3). Although the 95% CI of social condition's simple slope does not include -66.66 (indicating a significant difference between the ideal-performance slope (Simple ß$_{social}$ = -19.2, *SE* = 1.31, 95% CI = -21.7, -16.59, *p* < 0.001; Simple ß$_{nonsocial}$ = -11.2, *SE* = 1.31, 95% CI = -13.8, -8.66, *p* < 0.001), the slope difference between the social and nonsocial conditions shows that the simple slope of the social condition is significantly closer to the regression line representing the ideal performance than that of nonsocial condition's (ß$_{ideal\text{-}performance}$ - ß$_{social}$ = -47.46; ß$_{ideal\text{-}performance}$- ß$_{nonsocial}$ = -55.46). Figure 3.1. shows this linear relationship and the slope differences between the experimental conditions.

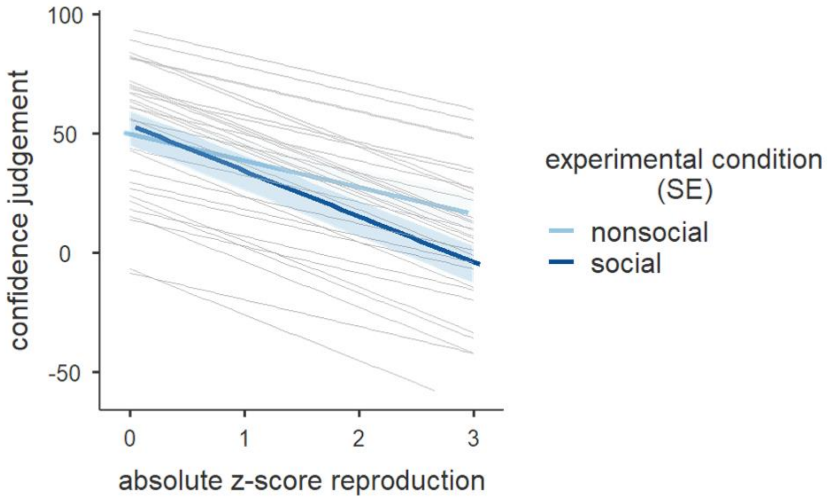


Figure 3.1.: Linear relationship between absolute deviation of reproduced durations and confidence judgements across experimental conditions. Light gray lines represent individual slopes, whereas colored bold lines represent average slope of the sample across experimental conditions (n=15). Error bars represent standard error (SE) of the estimate.

4.4.2.2. Directional error magnitude judgements as outcome variable

For the best model that explains the variance in directional error magnitude judgements (i.e., the model that includes only reproduced durations as fixed effect parameter), the main effect of z-score transformed reproduced durations was statistically significant (ß = 7.804, *SE* = 0.329 *p* < 0.001), revealing a positive linear relationship between reproduced durations and directional error magnitude judgements (see Figure 3.2).

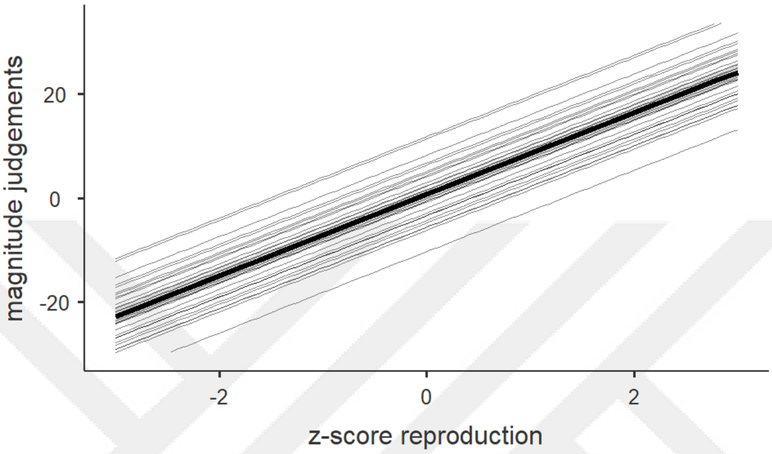


Figure 3.2.: Linear relationship between the z-score transformed reproduced durations and the directional error magnitude judgements. Light gray lines represent individual slopes, whereas bold black line represents average slope of the sample (N=30). Error bars represent standard error (SE) of the estimation.

4.5. Interim Discussion

In Experiment 3, we followed the same model comparison procedures as in the two experiments we conducted to select the model that explained our data for two of the outcome variables (i.e., confidence ratings and directional error magnitude judgements), separately. For the best model that explains our data for the confidence judgements, we found that participants in the social (i.e., experimenter observes the participants as they perform) condition were better at tracking their temporal errors than participants in the nonsocial (i.e., alone) condition. However, it is worth noting that this result is specific to confidence ratings. This result confirms our hypothesis that participants in the social condition would be more accurate to monitor their temporal errors compared to nonsocial condition. Moreover, we obtained the effect regardless of the possible biasing effect of our experimental manipulation on confidence judgements, as we found no main effect of the experimental condition. This finding holds crucial importance as it points out that the better match between the confidence judgements and temporal errors in the social condition compared to alone-condition was not led by the overall increase in the raw confidence judgements.

For the directional error magnitude judgements, however, we only found the main effect of reproduced durations. This result indicates that the error direction and magnitude monitoring performance was the same across the two experimental conditions (social and nonsocial). Moreover, as in the first two experiments we conducted, we did not find the moderating effect of mastery levels on either of our outcome variables (i.e., confidence ratings and directional error magnitude judgements), which refutes our hypothesis derived from the social facilitation theory predictions (being watched would improve performance for the high mastery group, and hinder it for the low mastery group, Zajonc, 1965; Cottrell, 1972). Taken together, social facilitation seems to not apply to temporal error monitoring performance.

In order to see if the results that we obtained in Experiment 3 are replicable, we ran a last study where we kept all the experimental procedure the same as in Experiment 3. The only difference was that the observer in the social condition was a male experimenter.

5. Experiment 4

Method

5.1. Participants

34 participants (25 females, 31 right handed, M_{age} = 21.2, SD_{age} = ± 2.9) from Koç University participated in Experiment 4. All participants received 1 extra credit in return for their participation in the experiment. As the number of reported over-confidence ratings (i.e., reporting 100% confidence regarding the accuracy of the temporal reproduction in the given

trial) of 4 participants exceeded half of the experimental trials, we included data from 30 participants in the analysis (15 participants for each condition).

5.2. Stimuli, Apparatus and Procedure

The stimuli, apparatus and procedure of Experiment 4 was kept the same as in Experiment 3 except that, the observer in the social condition was a male experimenter.

5.3. Data Analysis

5.3.1. Confidence judgements as outcome variable

We followed the same data analysis procedure for the confidence judgements as in Experiment 3. Model comparison determined the best model that explained the variance in the confidence judgement as the model that included the interaction between CV and absolute values of z-score reproduced durations (ΔBIC = 28.15 with the second best model, which included the three-way interaction between experimental condition, absolute values of z-score transformed reproduced durations and CV as fixed effects).

Model 4.1.:

confidence judgements ~ absolute z-score reproduction * CV + (1|participants)

where ~ stands for "predicted from" and 1|participants stands for "random intercept across participants". The model depicted above consisted of all lower terms.

We report the results of models that directly addresses our research questions in the Supplementary Materials.

5.3.2. Directional error judgements as outcome variable

We followed the same data analysis procedure for the directional temporal error magnitude judgements as in Experiment 3. Model comparison determined the best model that explained the variability in the magnitude judgement as the model that included only the main effect of z-score reproduced durations (ΔBIC = 6.83 with the second best model, which included the two-way interaction between experimental condition and of z-score transformed reproduced durations as fixed effects).

Model 4.2:

directional error magnitude judgements ~ z-score reproduced durations * CV + (1|participants)

where ~ stands for "predicted from" and 1|participants stands for "random intercept across participants". The model depicted above consisted of all lower terms.

As in the case of confidence judgements, the two (i.e., experimental condition and z-score reproduction) and three-way (i.e., CV included to the previous model) interaction models are the models that directly address our research questions are available in the Supplementary Materials.

5.4. Results

5.4.1. Comparison of reproduced durations and CVs across experimental conditions

In order to investigate if the experimental condition had an effect on reproduced durations, we compared the mean reproduced durations of the participants across conditions. Independent samples t-test revealed no statistically significant difference between the two experimental conditions in terms of reproduced durations ($t(28) = 1.1747$, $p > 0.05$, $SD = 0.3806$; $BF_{01} = 1.73$). We performed identical analysis using CV values as dependent variable. Independent samples t-test revealed no statistically significant difference between the two experimental conditions in terms of CV values ($t(28) = -1.0126$, $p > 0.05$, $SD = 0.0910$; $BF_{01} = 1.97$). Thus, the experimental manipulation did not have an effect on participants' reproduced durations or CV values.

5.4.2. Analysis of Metric Error Monitoring Performance

5.4.2.1. Confidence judgements as outcome variable

The best model that explains the variance in confidence judgements (i.e., the model that includes reproduced duration*CV interaction term) revealed a significant main effect of absolute values of z-score transformed reproduced durations on confidence judgements (ß = -8.51, *SE* = 0.880, *p* < 0.001), indicating that as absolute values of z-score transformed reproduced durations increases, confidence judgements decreases. Moreover, we found a statistically significant main effect of CV on confidence judgements (ß = -156.1, *SE* = 51.697, *p* < 0.01), which indicates that as CV increases, confidence levels decrease. Lastly, the two-way interaction between the fixed effects (i.e., CV and absolute values of z-score transformed reproductions) on confidence judgements was statistically significant (*F*(1,5884.1) = 44.48, *p*<0.001, ß= -64.9, *SE* = 9.731, *p* < 0.001), indicating that as CV increases, the slope for the relationship between absolute values of z-score transformed reproduced durations, and confidence judgements becomes steeper in the negative direction. Figure 4.1. shows this linear relationship and the slope differences across the levels of CV.

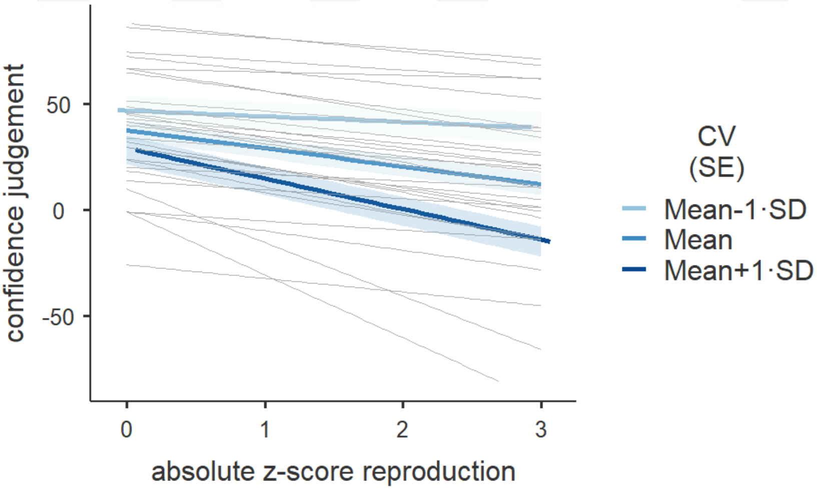
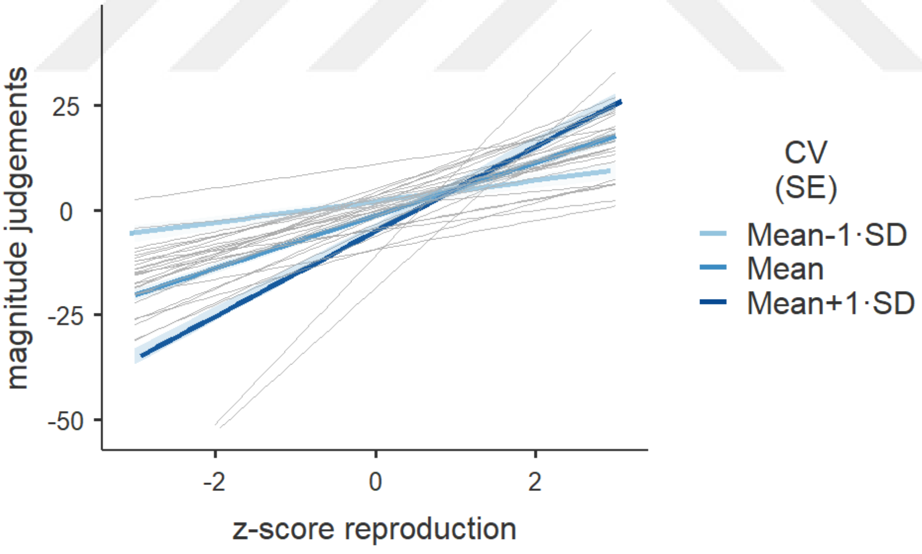


Figure 4.1.: Linear relationship between absolute deviation of reproduced durations and confidence judgements across levels of CV. Light gray lines represent individual slopes, whereas colored bold lines represent average slope of the sample across levels of CV. Error bars represent standard error (SE) of the estimate.

5.4.2.2. Directional error magnitude judgements as outcome variable

The best model that explains the variance in directional error magnitude judgements (i.e., the model that includes reproduced duration*CV interaction term) revealed a significant main effect of z-score transformed reproduced durations on directional error magnitude judgements (ß = 6.33, *SE* = 0.327, *p* < 0.001), indicating a positive linear relationship between the reproduced durations, and the directional error magnitude judgements[1]. Lastly, the two-way interaction between the fixed effects (i.e., CV and z-score transformed reproductions) on directional error magnitude judgements was statistically significant (*F*(1,5884.0) = 138.3, *p*<0.001, ß = 42.49, *SE* = 3.613, *p* < 0.001), indicating that as CV increases, the relationship between the z-score transformed reproduced durations and directional error magnitude judgements became more prominent. Figure 4.2. shows this linear relationship across the levels of CV.



[1] Note: There was a statistically significant negative relationship between CV and directional error magnitude judgements (ß = -35.15, *SE* = 10.776, *p* < 0.005) indicating that, increase in CV resulted in bias on directional error magnitude judgements towards negative direction (i.e., participants tended to report that their reproduction was "shorter" than the target duration).

Figure 4.2.: The interaction between reproduced durations and CV on directional error magnitude judgements. Light gray lines represent individual slopes, colored bold lines represent average slope of the sample (whereas colored bold lines represent the mean slope across levels of CV.) (N = 30). Error bars represent standard error (SE) of the estimate.

5.5. Interim Discussion

In Experiment 4, we followed the same model comparison procedures as in three experiments we conducted to select the model that explained our data for two of the outcome variables (i.e., confidence ratings and directional error magnitude judgements), separately. For the best model that explains our data for the confidence judgements, we found that as participants' CV levels (which is an index of mastery level in the timing context) increase, the slope for the relationship between absolute values of z-score transformed reproduced durations, and confidence judgements becomes steeper in the negative direction. However, we did not find the moderating effect of the experimental condition (i.e., social and nonsocial) on the relationship between the absolute deviation of the reproduced durations, and the confidence rating. These results also applied to the directional error magnitude judgements, as we found a statistically significant interaction between CV and the z-score transformed reproduced durations.

Taken together, we could not replicate the findings that we obtained in Experiment 3 (where we found an enhancing effect of social condition) for the confidence judgements. We also could not find the moderating effect of experimental condition on the relationship between the reproduced durations and directional error magnitude judgements.

General Discussion

This study aimed to investigate the effect of being watched on error monitoring performance in a temporal reproduction task, and how this effect would change for different task mastery levels. In light of previous findings (Baltazar et al., 2014; Hazem et al., 2017), we hypothesized that participants would be better at monitoring their timing errors when they are being watched. Moreover, we expected that participants with high mastery level would monitor their temporal errors better when they were being watched than participants with low mastery level would when being watched (Zajonc, 1965; Cottrell, 1972). We used coefficient of variation scores (CV) of participants' reproductions as an index of first order mastery level as it represents timing uncertainty in the literature (e.g., Çavdaroğlu, Zeki, & Balcı, 2014; Gibbon, 1977).

Overall, in four experiments we conducted, our results robustly pointed at the temporal error monitoring ability. That is, with two different indices of metric error monitoring ability (i.e., confidence and directional error magnitude judgements), we found that participants could monitor their timing errors in all four experiments. This result is a clear replication of the previous findings on temporal error monitoring (Akdoğan & Balcı, 2017; Doenyas et al., 2019; Kononowicz et al., 2018). Note that the same conclusion held when we used two different response options (i.e., slider scale in Experiment 1, 3, and 4 and keyboard in Experiment 2), which again points to the robustness of the effect. Thus, our findings strongly support the existence of "metric error monitoring", in addition to previously reported error monitoring in perceptual (Fleming, Huijgen, & Dolan, 2012), and memory (e.g., Flavell & Wellman, 1975) sub-domains. This replication carries critical importance for the literature as this rather newly discovered domain of error monitoring ability (Akdoğan & Balcı, 2017) holds metric characteristics, which renders it unique compared to widely studied domains such as meta-memory and perceptual decision making. This metric characteristic has two main particular advantages for the error monitoring research: it captures the informational richness to the data, and improves the ecological validity of the results obtained by including parametric actions.

In the first experiment we conducted, we could not find the effect of direct eye gaze either on the confidence judgements or directional error magnitude judgements. That is, participants' confidence judgements and directional error magnitude judgements did not "match" better (or worse) with their actual amount of error when they were in the direct eye gaze condition, compared to the other (i.e., averted gaze and non-social) conditions. We further replicated this null finding in Experiment 2. Taken together, these two studies indicate that direct gaze which aimed to induce a feeling of being watched did not improve or disrupt the confidence ratings and directional error magnitude judgements of the participants.

The null findings in the first two experiments might have occurred due to the ineffectiveness of our manipulations with virtual social stimuli to induce a feeling of being watched on participants. In her review, Hamilton (2016) discusses that feeling of being watched induction via picture and live stimuli can have different physiological effects, where skin conductance response (SRT), as well as N170 event-related potential (ERP) amplitude increases more as a response to a live person's direct gaze compared to picture of a direct gaze (Pönkänen et al., 2011; 2012, as cited in Hamilton, 2016). In order to address this possibility, we conducted a third experiment in which a real observer watched the participants' performance throughout the test session. As a result of our more realistic manipulation, we found an enhancing effect of being watched for the match between the amount of error and confidence judgements, pointing out the higher effectiveness of the "real observer'' to induce a feeling of being watched, compared to the virtual "observer" (for a detailed discussion, see: Hamilton, 2016). However, since these results were not replicated in Experiment 4 with a male experimenter, these conclusions should be treated with caution.

Although the ineffectiveness of our visual stimuli is a possible explanation for the null results we obtained in the first two experiments, note that Baltazar et al. (2014) has found the enhancing effect of being watched with similar visual stimuli to ours (i.e., virtual averted and direct gazes) on physical state monitoring performance. That is, they found a higher correlation between the self-reports of the participants regarding their physical states (as a response to emotional stimuli) and the SCRs in the direct gaze condition compared to averted gaze, and fixation cross conditions.

As the stimuli that we used in the current study and that in Baltazar et al (2014) were equivalent, the differences in findings we obtained might have possibly stemmed from the differences in “what is being watched” instead of “how being watched is induced” among our four experiments, and that of Baltazar et al (2014). That is, in the first two experiments, we attempted to induce a feeling of being watched via a direct gaze (picture/virtual) stimulus, which appeared to be looking at the “self”, as in Baltazar et al (2014). However, in our last two experiments, there was a real observer whose gaze was directed not at the participant but at their performance by looking at the computer screen as they performed the task. In this respect, Baltazar et al. (2014) discusses that the physical state monitoring ability was enhanced via ‘’self-focused attention’’, triggered by an observer watching the “self”. Given the fact that physical states are directly related to self (Baltazar et al., 2014), self focused attention which refers to “...an awareness of self-referent information’’ (Spurr & Stopa, 2002; also see: Carver & Scheier, 1978) can enhance physical state monitoring performance, as found by Baltazar et al. (2014).

We speculate, however, that, instead of “self-referential processing” which can be supported by self-focused attention; error monitoring seems much more related to “performance-referential processing”, as it requires one to keep track of one’s own performance (Yeung & Summerfield, 2012), and not the “self”. Thus, differently from physical state monitoring that is found to be supported by self focused attention (Baltazar, 2014), “performance-focused attention” (Innes & Young, 1975) might be the supporting factor for temporal error monitoring performance. This performance focused attention can be induced by an observer who watches the performance of the participants instead of their “selves”, as we manipulated in Experiment 3 and 4. As a result, performance focused attention can enhance temporal error monitoring performance. Although the null effects of observer watching “the self” in our first two experiments and the enhancing effect of a female experimenter watching the “performance” in the third experiment seem to support this idea, we could not replicate the enhancing effect that we found in the third experiment with a male experimenter. Hence, we strongly stress that this speculation should be considered with extra caution.

The reason why we could not replicate the findings in Experiment 3 can be due to the gender difference between experimenters who observed the participants’ performance. Regarding the importance of the influence of observer’s gender on performance, one study has shown that female participants performed best in a computer-based task in the presence of a female observer, compared to when they perform the task alone, or in the presence of a male observer (Corston & Colman, 1996). When we consider the fact that the gender of the participants was not evenly distributed in the last two experiments we conducted (i.e., the number of female participants outweighed the male participants in both experiments, Experiment 3: $N_{female} = 22$, Experiment 4: $N_{female} = 25$), female participants might have performed better in the Experiment 3 where the observer was female, and thus, inflated the overall performance of the social condition. On the other hand, as the observer in Experiment 4 was male, such an inflation may not have happened and thus, the observed effect in Experiment 3 disappeared.

Note that we reanalyzed our data in the first two experiments to address this possible enhancing effect of female observer (presented as visual stimuli). Results revealed no enhancing effect of the female observer over the male observer on the temporal error monitoring performance, neither for confidence or magnitude judgements in both experiments. Again, as this null finding might be due to the fact that our visual stimuli seemed to watch the "self" and not the "performance", we suggest to run another replication experiment where another female experimenter observes participants in the social condition as in Experiment 3 and 4. This new experiment is also needed as a more controlled replication of the last two experiments we conducted in terms of the gender distribution of the participants in both experimental conditions (i.e., social and nonsocial).

Addressing our second research question regarding the social facilitation effect on temporal error monitoring performance, in our data, mastery level did not moderate the effect of being watched on temporal error monitoring performance in any of our four experiments. If anything, the post-hoc analyses of the significant three-way interaction for directional error magnitude judgements in Experiment 1 yielded a non-significant trend in the opposite direction to our predictions. That is, there was a trend implying that the effect of being watched on temporal error monitoring performance was more pronounced as mastery level decreased. However, this finding was not replicated in Experiment 4 with a male experimenter.

This conflicting result might have occurred due to a number of reasons. First, we might not have achieved enough variance for the mastery levels to demonstrate the moderating effect. Secondly, although we think that our experimental manipulations were effective to establish a feeling of being watched, they could not have established an evaluation apprehension (Cottrell, 1972). Thirdly, in Experiment 1 and Experiment 2 participants may as well have felt like they are being evaluated, but because it was the "self" that is being watched in those studies, the evaluation apprehension might have been irrelevant for error monitoring. On the other hand, in Experiment 3 and Experiment 4 we explicitly stated that the experimenter would evaluate the participants' performance, which addressed this possibility. Despite this fact, we could not find the moderating effect of mastery level in either Experiment 3 or Experiment 4. Taken together, our findings in four experiments we conducted refute our hypothesis that participants who are high in mastery level would more accurately monitor their temporal errors when they are being watched than participants who are low in mastery level (Zajonc, 1965; Cottrell, 1972). Thus, social facilitation does not apply to metric error monitoring performance.

The reason why social facilitation does not apply to error monitoring performance can be the dissociation between the first and second-order decisions (e.g., Fleming, Huijgen, & Dolan, 2012; Boldt, & Yeung, 2015). That is, the first order decisions cover "low" level, perceptual decisions, whereas second-order decisions cover "high" level decisions about the first-order decisions (e.g., deciding if the given answer was correct or not; Fleming, Dolan & Frith, 2012). Such a difference between the two types of decisions might have led the social facilitation effect to not hold for the second-order decision performance (it is crucial to note that we did not see the effect of our manipulation on the first order performance, either). Thus, performances in those two levels may not always be affected by external influences in the same way. However, exactly what difference in these two levels of performance withholds social facilitation to apply to second order performance remains to be investigated.

It is important to acknowledge the limitations of the current study. In the first two experiments, we could not make sure that the participants indeed made eye contact with the direct gaze stimuli. Thus, a more controlled experiment can be done with an eye tracker to make sure that the participants look directly into the eyes of the direct gaze stimuli, and so that the effect that is being tried to induce (i.e., feeling of being watched) can improve in strength. In our opinion, however, this limitation also holds for Baltazar et al. (2014). Another limitation is that, the cursor that participants used to indicate their confidence and directional error magnitude judgements started at the center of the slider scale. However, in Experiment 3 and Experiment 4 we randomized the location of the cursor. Such a methodological difference can also have resulted in the differences between the results obtained in these experiments (but note also the different results of Experiment 3 and 4). Lastly, the participant gender was not evenly distributed across conditions in the last two experiments, where the number of female participants outweighed the number male participants in both experiments. For this reason, we could not address the interaction between the gender of the participant and gender of the experimenter on temporal error monitoring performance, as how the responsiveness to the gender of the observer differs for participants with different genders is highlighted in the literature (e.g., Corston & Colman, 1996). Future studies should take this fact into consideration.

In addition to the limitations highlighted above, why social facilitation does not seem to apply to temporal error monitoring can be due to the individual differences (Uziel, 2007). In this respect, in his review, Uziel (2007) discusses that being watched would result in performance improving effect for individuals with positive personality traits (i.e., positive self assured), and that this effect would reverse for individuals with negative personality traits (i.e., negative apprehensive). Likewise, Double & Birney (2019) found that asking participants to report their confidence regarding their performance in a Raven's Progressive Matrices (RPM) task had improving effect on performance monitoring for participants who had high self-confidence, compared to those who had low self confidence. However, these possible moderating effects were not addressed in the current study.

To our best knowledge, although its effect on first order task performance has been investigated by previous work (e.g., Grant & Dajee, 2003; Conty et al., 2010), this study for the first time investigated the effect of feeling of being watched on error monitoring performance, and how this possible effect can change across the levels of first order mastery. In sum, our results in the four experiments we conducted indicate the robustness of metric error monitoring ability by replicating the results of previous studies that investigated this ability (Akdoğan & Balcı, 2017; Duyan & Balcı, 2018; 2019; Doenyas et al., 2019; Kononowicz et al., 2018). However, although we found the facilitating effect of social presence when the performance is being watched in Experiment 3, as we could not replicate this effect, we conclude that this robust ability of monitoring the metric errors is insensitive to being watched that is induced by social presence. It holds crucial importance for the future studies to test if this ineffectiveness is specific to being watched by testing for other social contexts such as group decision making.

Conclusion

Although its effect on the first order task has been investigated (e.g., Grant & Dajee, 2003; Zajonc, 1965; Triplett, 1898), the effect of being watched on error monitoring performance has not been investigated. In the current study, we aimed to address this gap in the literature. As a result of the third experiment we conducted, we speculated that the possible facilitating effect of being watched may not necessarily be due to "how being watched" is established (i.e., either via virtual stimuli or via a real observer as) per se (see: Hamilton, 2016). Instead, this facilitating effect of being watched might be a result of "what is being watched" (i.e., "self" or "performance"), as we found a significant effect of being watched on confidence judgements in Experiment 3. Moreover, the differences in "what is being watched" seemed to be triggering possibly different cognitive mechanisms, namely, self-focused attention and performance-focused attention in Experiment 3. However, we could not replicate this effect in Experiment 4. Thus, the possibility that "self-focused attention" and "performance-focused attention" notions may be differentiated from one another possibly due to their dependence on different cognitive mechanisms should not be over-interpreted. Taken together, with four experiments, the current study demonstrates the robustness of the metric error monitoring ability, which is not sensitive to the social influences, more specifically; being watched by others. Future studies should investigate whether this insensitivity of temporal error monitoring to social influences is specific to the "feeling of being watched" by testing for alternative social influences in such contexts as group decision making and conformity.

.

References

Aiello, J. R., & Douthitt, E. A. (2001). Social facilitation from Triplett to electronic performance monitoring. *Group Dynamics: Theory, Research, and Practice, 5*(3), 163.

Akdoğan, B., & Balcı, F. (2017). Are you early or late?: Temporal error monitoring. *Journal of Experimental Psychology: General.* Advance online publication. http://dx.doi.org/10.1037/xge0000265

Baltazar, M., Hazem, N., Vilarem, E., Beaucousin, V., Picq, J. L., & Conty, L. (2014). Eye contact elicits bodily self-awareness in human adults. *Cognition*, *133*(1), 120–127. https://doi.org/10.1016/j.cognition.2014.06.009

Boldt, A., & Yeung, N. (2015). Shared neural markers of decision confidence and error detection. *Journal of Neuroscience*, *35*(8), 3478–3484. https://doi.org/10.1523/JNEUROSCI.0797-14.2015

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision 10*(4), 433-436.

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

Carver, C. S., & Scheier, M. F. (1978). Self-focusing effects of dispositional self-consciousness, mirror presence, and audience presence. *Journal of Personality and Social Psychology*, *36*(3), 324–332. https://doi.org/10.1037/0022-3514.36.3.324

Cavdaroglu, B., Zeki, M., & Balci, F. (2014). Time-based reward maximization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1637), https://doi.org/10.1098/rstb.2012.0461

Çetin, B., Doğan, T., & Sapmaz, F. (2010). Olumsuz değerlendirilme korkusu ölçeği kısa formunun Türkçe uyarlaması: Geçerlik ve güvenirlik çalışması. *Eğitim ve Bilim, 35*(156).

Chib, V. S., Adachi, R., & O'doherty, J. P. (2018). Neural substrates of social facilitation effects on incentive-based performance. *Social Cognitive and Affective Neuroscience, 13*(4), 391-403.

Conty, L., George, N., & Hietanen, J. K. (2016). Watching eyes effects: When others meet the self. *Consciousness and Cognition*, *45*, 184-197.

Conty, L., Gimmig, D., Belletier, C., George, N., & Huguet, P. (2010). The cost of being watched: Stroop interference increases under concomitant eye contact. *Cognition, 115*(1), 133-139.

Corston, R., & Colman, A. M. (1996). Gender and social facilitation effects on computer competence and attitudes toward computers. *Journal of Educational Computing Research, 14*(2), 171-183.

Cottrell, N. B. Social facilitation. In C. G. McClintock (Ed.), *Experimental social psychology.* New York: Holt.

De Carvalho Filho, M. K., & Yuzawa, M. (2001). The effect of social influences and general metacognitive knowledge on metamemory judgments. *Contemporary Educational Psychology*, *26*(4), 571–587. https://doi.org/10.1006/ceps.2000.1077

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.

Double, K. S., & Birney, D. P. (2019). Do confidence ratings prime confidence?. *Psychonomic Bulletin & Review*, *2003,* 1-8. DOI: 10.3758/s13423-018-1553-3

Duval, S., & Wicklund, R. A. (1972). A theory of objective self awareness. Oxford, England: Academic Press.

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.

Eskenazi, T., Montalan, B., Jacquot, A., Proust, J., Grèzes, J., & Conty, L. (2016). Social influence on metacognitive evaluations: The power of nonverbal cues. *Quarterly Journal of Experimental Psychology*, *69*(11), 2233–2247. https://doi.org/10.1080/17470218.2015.1115111

Flavell, J. (1979). Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. *American Psychologist*, *34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

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

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making : A general Bayesian framework for metacognitive computation, *Psychological Review*, *124*(1), 91–114.

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

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences. 367*(1594), 1280–1286 doi:10.1098/rstb.2012.0021

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 1–9. https://doi.org/10.3389/fnhum.2014.00443

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

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review*, *5*(3), 490-495.

Frith, C. D. (2012). The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2213–2223. https://doi.org/10.1098/rstb.2012.0123

Gallucci, M. (2019). *GAMLj: General analyses for linear models.* [jamovi module]. Retrieved from https://gamlj.github.io/.

Gibbon, J., Church, R. M. , & Meck, W. H. (1984). Scalar Timing in Memory. *Annals of the New York Academy of Sciences*, *423*, 52–77.

Germar, M., Schlemmer, A., Krug, K., Voss, A., & Mojzisch, A. (2014). Social influence and perceptual decision Making. *Personality and Social Psychology Bulletin*, *40*(2), 217–231. https://doi.org/10.1177/0146167213508985

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psychological Review*, *84*(3), 279–325. https://doi.org/10.1037/0033-295X.84.3.279

Grant, T. & Dajee, K. (2003). Types of task, types of audience, types of actor: interactions between mere presence and personality type in a simple mathematical task. *Personality and Individual Differences, 35*(3), 633-639. https://doi.org/10.1016/S0191-8869(02)00241-6

Hamilton, A. F. D. C. (2016). Gazing at me: the importance of social meaning in understanding direct-gaze cues. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1686), 20150080.

Hazem, N., George, N., Baltazar, M., & Conty, L. (2017). I know you can see me: Social attention influences bodily self-awareness. *Biological Psychology*, *124*, 21–29. https://doi.org/10.1016/j.biopsycho.2017.01.007

Innes, J. M., & Young, R. F. (1975). The effect of presence of an audience, evaluation apprehension and objective self-awareness on learning. *Journal of Experimental Social Psychology*.

Jacquot, A., Eskenazi, T., Sales-Wuillemin, E., Montalan, B., Proust, J., Grèzes, J., & Conty, L. (2015). Source unreliability decreases but does not cancel the impact of social information on metacognitive evaluations. *Frontiers in Psychology*, *6*, 1–11. https://doi.org/10.3389/fpsyg.2015.01385

Jarick, M., Laidlaw, K. E., Nasiopoulos, E., & Kingstone, A. (2016). Eye contact affects attention more than arousal as revealed by prospective time estimation. *Attention, Perception, & Psychophysics*, *78*(5), 1302-1307.

Kononowicz, T. W., Roger, C., & van Wassenhove, V. (2018). Temporal metacognition as the decoding of self-generated brain dynamics. *bioRxiv*, 206086

Leary, M. R. (1983). A brief version of the Fear of Negative Evaluation Scale. *Personality and Social Psychology Bulletin, 9*, 371-376.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Orr, J. M., & Carrasco, M. (2011). The role of the error positivity in the conscious perception of errors. *Journal of Neuroscience*, *31*(16), 5891-5892.

Overgaard, M., & Sandberg, K. (2012). Kinds of access: Different methods for report reveal different kinds of metacognitive access. *The Cognitive Neuroscience of Metacognition*, *9783642451*, 67–85. https://doi.org/10.1007/978-3-642-45190-4_4

Pescetelli, N., Rees, G., & Bahrami, B. (2016). The perceptual and social components of metacognition. *Journal of Experimental Psychology: General*, *145*(8), 949–965. https://doi.org/10.1037/xge0000180

Pönkänen, L. M., Alhoniemi, A., Leppänen, J. M., & Hietanen, J. K. (2010). Does it make a difference if I have an eye contact with you or with your picture? An ERP study. *Social Cognitive and Affective Neuroscience, 6*(4), 486-494.

Pönkänen, L. M., & Hietanen, J. K. (2012). Eye contact with neutral and smiling faces: Effects on autonomic responses and frontal EEG asymmetry. *Frontiers in Human Neuroscience, 6,* 122.

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.

Rosenberg, M. J. (1965). When dissonance fails: On eliminating evaluation apprehension from attitude measurement. *Journal of Personality and Social Psychology. 1* (1): 28–42. doi:10.1037/h0021647.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/

R Core Team (2018). *R: A Language and environment for statistical computing.* [Computer software]. Retrieved from https://cran.r-project.org/.

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*(1–2), 111–139. https://doi.org/10.1007/s11165-005-3917-8

Spurr, J. M., & Stopa, L. (2002). Self-focused attention in social phobia and social anxiety. *Clinical Psychology Review, 22*(7), 947-975.

The jamovi project (2019). *jamovi*. (Version 0.9) [Computer Software]. Retrieved from https://www.jamovi.org

Triplett, N. (1898). The dynamogenic factors in pacemaking and competition. *The American Journal of Psychology, 9*(4), 507-533.

Uziel, L. (2007). Individual differences in the social facilitation effect: A review and meta-analysis. *Journal of Research in Personality*, *41*(3), 579-601.

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. https://doi.org/10.1098/rstb.2011.0416

Yoshie, M., Nagai, Y., Critchley, H. D., & Harrison, N. A. (2016). Why I tense up when you watch me: Inferior parietal cortex mediates an audience's influence on motor performance. *Scientific Reports*, *6*, 1–11. https://doi.org/10.1038/srep19305

Zajonc, R. B. (1965). Social Facilitation. *Science*. https://doi.org/10.1126/science.149.3681.269

Supplementary Material

1. Experiment 1

Analysis of Metric Error Monitoring Performance

1.1. Confidence judgements as outcome variable

To investigate the interaction between the experimental condition, and the absolute values of z-score transformed reproduced durations on confidence judgements, we performed a linear mixed effects model (Model S1.1).

Model S1.1:

confidence judgements ~ absolute z-score reproduction * experimental condition + (1|participants)

where ~ stands for "predicted from" and 1|participants stands for "random intercept across participants". The model depicted above consisted of all lower terms.

Results revealed a statistically significant main effect of absolute values of z-score transformed reproduced durations on raw confidence judgements (ß = -15.72, *SE* = 0.845, *p* <

0.001), indicating that regardless of the experimental condition, participants could monitor the absolute magnitude of their timing errors. However, neither the main effect of the experimental condition nor the two-way interaction (experimental condition * absolute values of z-score transformed reproduced durations) were significant (all p > 0.05; see: Figure S1.1.).

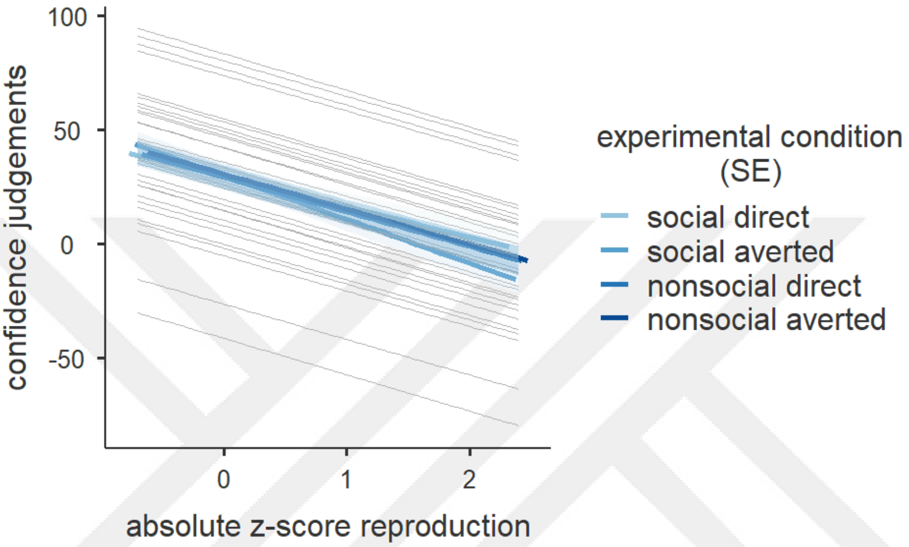


Figure S1.1: Linear relationship between the absolute z-score transformed reproduced durations and the confidence judgements across experimental conditions. Light gray lines represent individual slopes, whereas colored bold lines represent the mean slope for four experimental conditions. Error bars represent standard error (SE) of the estimation.

For the second model (i.e., CV included), our results showed that the main effect of absolute values of z-score transformed reproduced durations (ß = -15.607, *SE* = 0.845, p < 0.001) and the main effect of CV (ß = -167.34, *SE* = 40.43, p <0.001) on confidence judgement was statistically significant. The interaction between the experimental condition and the absolute values of z-score transformed reproduced durations was marginally significant (F(1,6037.6) = 2.566, p = 0.053). However, none of the other two-way and three-way interaction terms and the main effects were significant (all p > 0.05; see: Figure S1.2.).

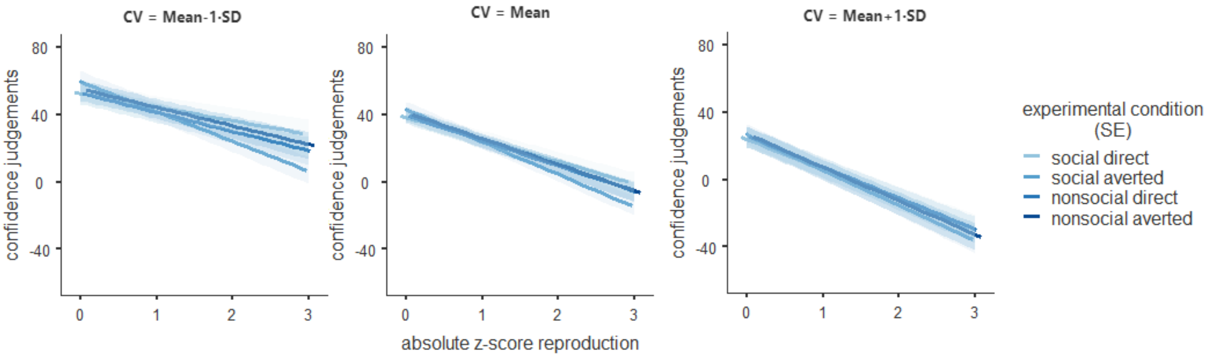


Figure S1.2: Linear relationship between absolute values of z-score transformed reproduced durations and confidence judgements across experimental conditions in three different levels of mastery levels. Low levels of CV indicate high mastery. Error bars represent standard error (SE).

1.2. Directional error magnitude judgements as outcome variable

To investigate the interaction between the experimental condition, and the z-score transformed reproduced durations on error magnitude, and direction judgements, we performed a linear mixed effects model (Model S1.2).

Model S1.2.

directional error magnitude judgements ~ z-score reproduction * experimental condition + (1|participants)

where ~ stands for "predicted from" and 1|participants stands for "random intercept across participants". The model depicted above consisted of all lower terms.

Results revealed only the main effect of z-score transformed reproduced durations as significant (ß = 9.5004, *SE*= 0.331, *p* < 0.001). The main effect of experimental condition and the interaction between the experimental condition and the z-score transformed reproduced durations were not statistically significant (all *p* > 0.05, see: Figure S2.1).

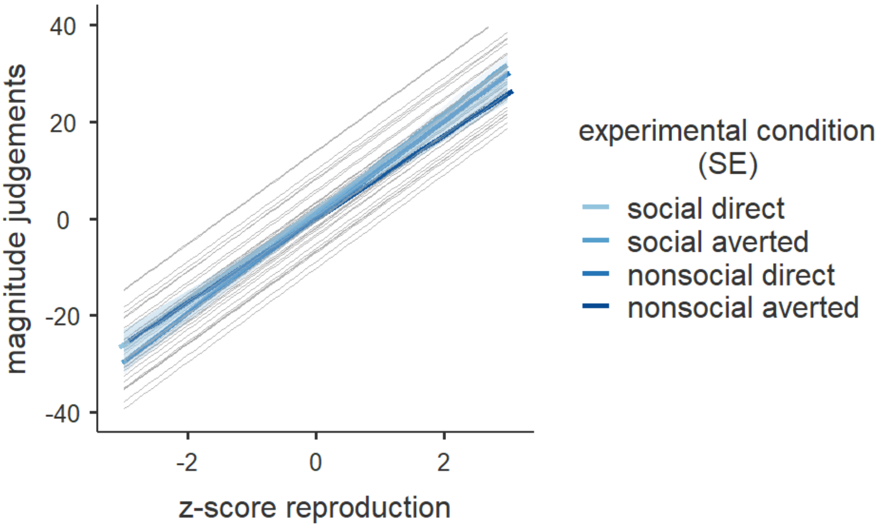


Figure S2.1: Linear relationship between the z-score transformed reproduced durations and the directional error magnitude judgements across experimental conditions. Light gray lines represent individual slopes, whereas colored bold lines represent the mean slope for four experimental conditions. Error bars represent standard error (SE) of the estimation.

2. Experiment 2

To investigate the interaction between z-score transformed reproduced durations and experimental condition on recoded confidence judgements, we performed a linear mixed effects model (Model S2).

Model S2.1:

recoded confidence judgements ~ z-score reproduction * experimental condition + (1|participants)

where ~ stands for "predicted from" and 1|participants stands for "random intercept across participants". The model depicted above consisted of all lower terms.

Results revealed a statistically significant main effect of z-score transformed reproduced durations on raw confidence judgements (ß= 0.6779, *SE* =0.0246, p < 0.001), indicating that regardless of the experimental condition, participants could monitor their temporal errors. The main effect of the experimental condition and the two-way interaction between the experimental condition and the z-score transformed reproduced durations on rescaled confidence judgements were not significant (all $p > 0.05$; see: Figure S3.1).

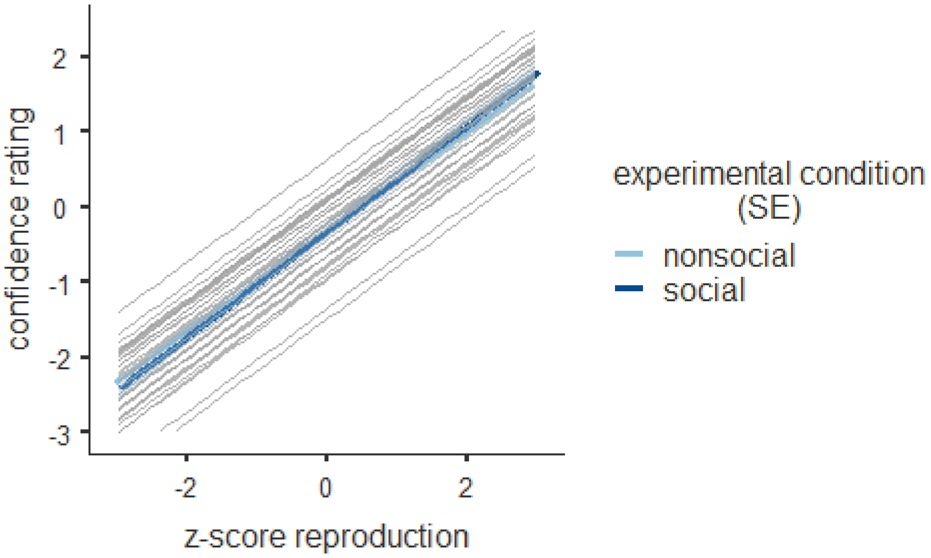


Figure S3.1: Linear relationship between the z-score transformed reproduced durations and concatenated confidence, and error directionality judgements across experimental conditions. The values on the y-axis that are high in magnitude represent low confidence and error magnitude. Minus and plus signs represent the reported error direction judgements (minus for "short", plus for "long"). Light gray lines represent individual slopes, whereas colored bold lines represent the mean slope for four experimental conditions. Error bars represent standard error (SE) of the estimation.

In order to address our second research question if levels of mastery can moderate the effect of social presence on temporal error monitoring performance, we next added CV as a continuous covariate and included all interaction terms to the model (Model S2.2, see Figure S3.2).

Model S2.2:

recoded confidence judgements ~ z-score reproduction * experimental condition * CV + (1|participants)

where ~ stands for "predicted from" and 1|participants stands for "random intercept across participants". The model depicted above consisted of all lower terms.

Again, the main effect of z-score transformed reproduced durations was significant (ß= 0.6771, *SE* =0.0246, *p* < 0.001) which indicates a positive linear relationship between the z-score transformed reproduced durations and directional error magnitude judgements. Additionally, the main effect of CV was statistically significant (ß= -2.0651, *SE* =0.685, *p* = 0.005), indicating that as timing uncertainty increases, confidence decreases. All other two and three-way interactions were nonsignificant (all *p* > 0.05, see: Figure S3.2).

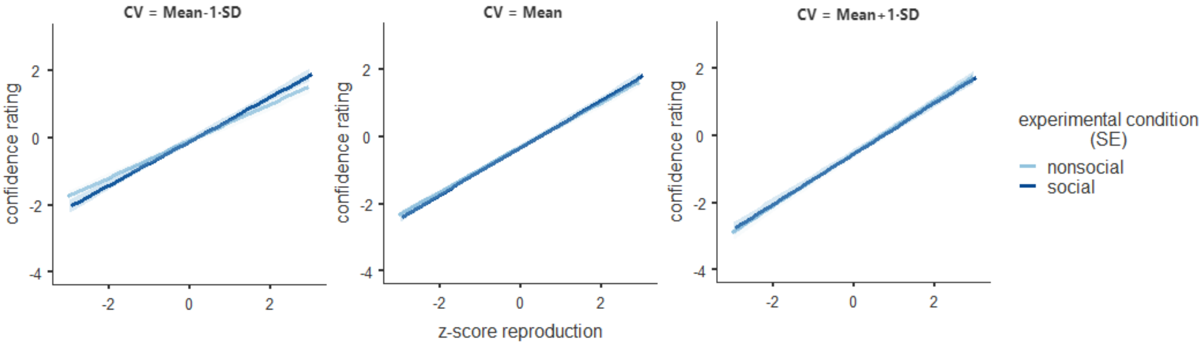


Figure S3.2.: Linear relationship between z-score transformed reproduced durations and concatenated confidence, and error directionality judgements across experimental conditions in three different levels of mastery levels. Low levels of CV indicate high mastery. Error bars represent standard error (SE).

2.2. Exploratory Analysis: Fear of Negative Evaluation Scores

Although we did not hold a specific expectation for the Fear of Negative Evaluation scores, we performed an exploratory analysis to investigate its potential effect on the recoded confidence judgements. For this reason, we investigated the interaction between the experimental condition and z-score transformed reproduced durations, CV, and mean Fear of Negative Evaluation scores on rescaled confidence judgements. To this end, we included mean Fear of Negative Evaluation scores of the participants, experimental conditions (social or nonsocial), z-score transformed reproduced durations and CV, and their interaction as fixed effects on the slope. We included participants as random effect on the intercept (Model S3).

Model S3:

recoded confidence judgements ~ z-score reproduced durations * experimental condition * CV * mean Fear of Negative Evaluation scores + (1|participants)

where ~ stands for "predicted from", and 1|participants stands for "random intercept across participants".

Results revealed a significant main effect of z-score reproduction on recoded confidence judgements (ß = 0.68, *SE* = 0.025, $p < 0.001$), suggesting that participants could monitor their temporal error magnitudes and directions. Moreover, three-way interaction between experimental condition, z-score transformed reproduced durations, and mean Fear of Negative Evaluation scores ($F(1,6801.6) = 5.123$, $p < 0.05$, ß$_{\textit{social-nonsocial*reproduction*mean anxiety scores}}$ = -0.16, *SE* = 0.069, $p < 0.05$), suggesting that as the Fear of Negative Evaluation scores increase, the relationship between the recoded confidence judgements and reproduced durations significantly decrease in the social condition compared to nonsocial condition.

3. Experiment 3

3.1. Confidence judgements as outcome variable

To investigate the interaction between absolute values of z-score transformed reproduced durations, experimental condition, and CV on confidence judgements, we performed a linear mixed effects model (Model S4.1).

Model S4.1:

confidence judgements ~ absolute z-score reproduction * experimental condition * CV + (1|participants)

where ~ stands for "predicted from" and 1|participants stands for "random intercept across participants". The model depicted above consisted of all lower terms.

Results revealed a statistically significant main effect of absolute values of z-score transformed reproduced durations (ß = -15.16, *SE* = 0.928, $p < 0.001$) on confidence judgements. Also the interaction between the experimental condition and the absolute values of z-score transformed reproduced durations on confidence judgements was statistically significant ($F(1,5917.6) = 19.47$, $p < 0.001$, ß$_{social\text{-}nonsocial}$ = -8.19, *SE* = 1.85 $p < 0.001$, see: Figure S4.1).

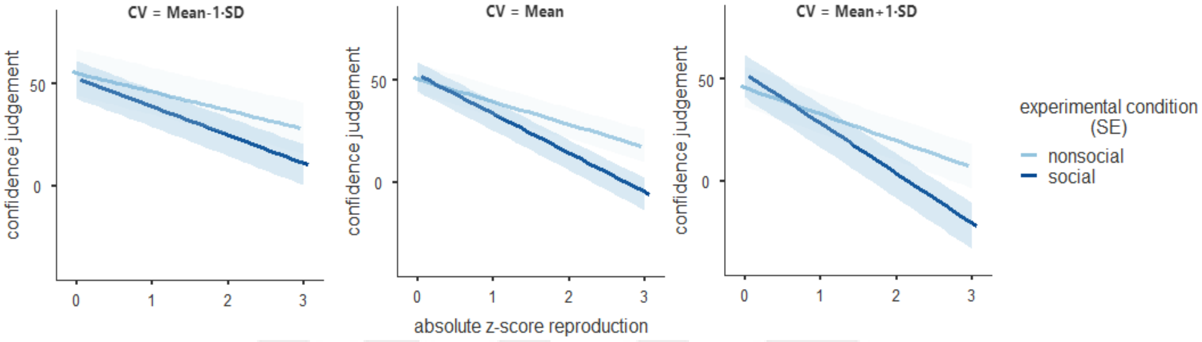


Figure S4.1 : Linear relationship between absolute values of z-score transformed reproduced durations, and confidence judgements across experimental conditions in three different levels of mastery levels. Low levels of CV indicate high mastery. Error bars represent standard error (SE).

3.2. Directional error magnitude judgements as outcome variable

To investigate the interaction between z-score transformed reproduced durations, experimental condition on directional error magnitude judgements, we performed a linear mixed effects model (Model S3.2).

Model S4.2:

directional error magnitude judgements ~ z-score reproduction * experimental condition + (1|participants)

where ~ stands for “predicted from” and 1|participants stands for “random intercept across participants”. The model depicted above consisted of all lower terms.

Results revealed a statistically significant main effect of the z-score transformed reproduced durations on directional error magnitude judgements (ß = 7.797, *SE* = 0.329, $p <$ 0.001). However, the main effect of experimental condition and the two-way interaction between the experimental condition and the z-score reproduced durations on directional error magnitude judgements was not significant (all $p > 0.05$, see: Figure S4.2.1).

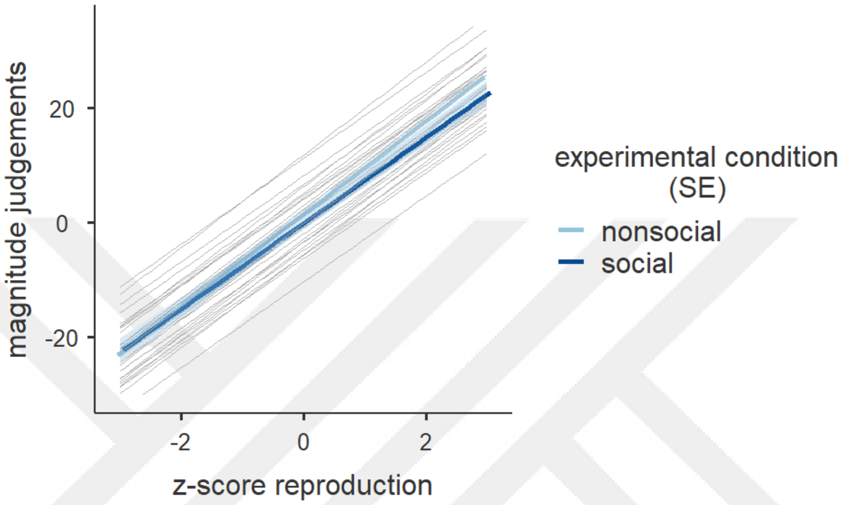


Figure S4.2.1: Linear relationship between the z-score transformed reproduced durations, and the directional error magnitude judgements across experimental conditions. Light gray lines represent individual slopes, whereas colored bold lines represent average slope of the sample across experimental conditions (n=15). Error bars represent standard error (SE) of the estimation.

Next, we added CV as a fixed effect parameter on slope to our model (Model S4.3)

Model S4.3:

directional error magnitude judgements ~ z-score reproduction * experimental condition * CV + (1|participants)

where ~ stands for "predicted from" and 1|participants stands for "random intercept across participants". The model depicted above consisted of all lower terms.

Results revealed a statistically significant main effect of the z-score transformed reproduced durations (ß = 7.837, *SE* = 0.33, $p < 0.001$). None of the other main effects and interaction terms were significant (all *p*s > 0.05, see: Figure S4.2.2).

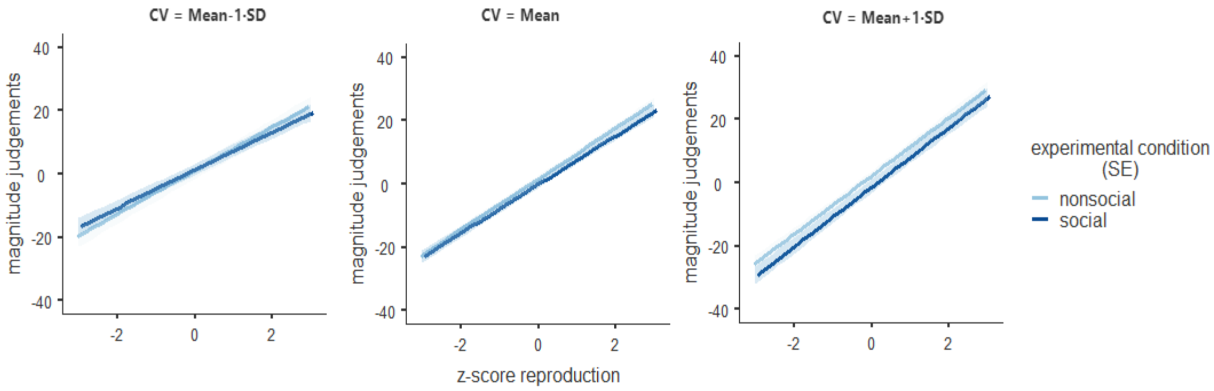


Figure S4.2.2: Linear relationship between z-score transformed reproduced durations and directional error magnitude judgements across experimental conditions in three different levels of mastery levels. Low levels of CV indicate high mastery. Error bars represent standard error (SE).

4. Experiment 4

4.1. Confidence judgements as outcome variable

We investigated the interaction between the absolute values of z-score transformed reproduced durations, and experimental condition on confidence judgements, we performed a linear mixed effects model (Model S5.1).

Model S5.1:

confidence judgements ~ absolute z-score reproduction * experimental condition + (1|participants)

where ~ stands for "predicted from" and 1|participants stands for "random intercept across participants". The model depicted above consisted of all lower terms.

Results revealed a significant main effect of absolute values of z-score reproduced duration on confidence judgements (ß= -8.54, *SE* = 0.884, $p < 0.001$). Also, the main effect of the experimental condition on confidence judgements was statistically significant (ß$_{\text{social}}$ - ß$_{\text{nonsocial}}$ = 26.74, *SE* = 9.384, $p < 0.01$). However, the interaction between the absolute values of z-score reproduced duration and experimental condition on confidence judgements was not significant ($p > 0.05$, see: Figure S5.1).

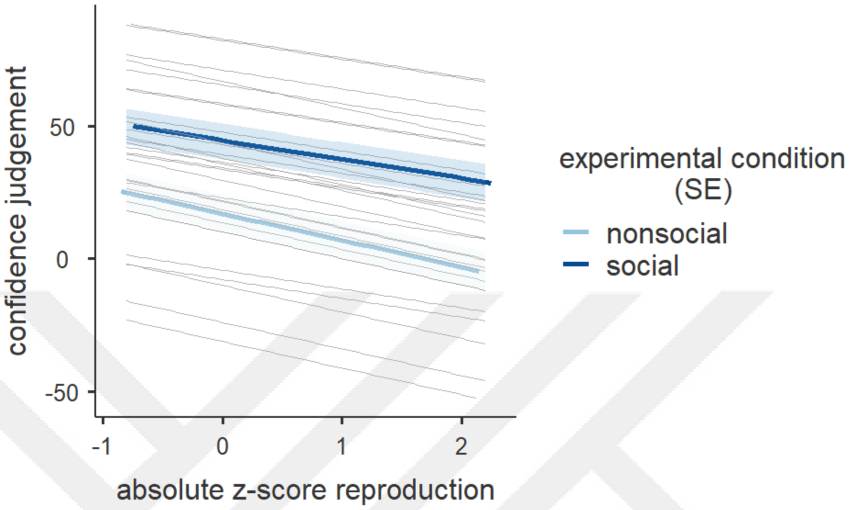


Figure S5.1: The linear relationship between absolute deviation of reproduced durations and confidence judgements across experimental conditions. Light gray lines represent individual slopes, whereas colored bold lines represent average slope of the sample across experimental conditions (n=15). Error bars represent standard error (SE) of the estimate.

To investigate the interaction between absolute values of z-score transformed reproduced durations, experimental condition, and CV on confidence judgements, we performed a linear mixed effects model (Model S5.2).

Model S5.2:

confidence judgements ~ absolute z-score reproduction * experimental condition * CV + (1|participants)

where ~ stands for "predicted from" and 1|participants stands for "random intercept across participants". The model depicted above consisted of all lower terms.

Results revealed a significant main effect of absolute values of z-score reproduced duration on confidence judgements (ß= -8.552, *SE* = 0.913, *p* < 0.001). Also, the main effect of the experimental condition and CV on confidence judgements was statistically significant (experimental condition: ß$_{\text{social}}$ - ß$_{\text{nonsocial}}$ = 21.35, *SE* = 8.655, *p* < 0.05; CV: ß = -157.746, *SE* = 69.947, *p* < 0.05)[2]. All other interaction terms were not statistically significant (all *p* > 0.05, see: Figure S5.2).

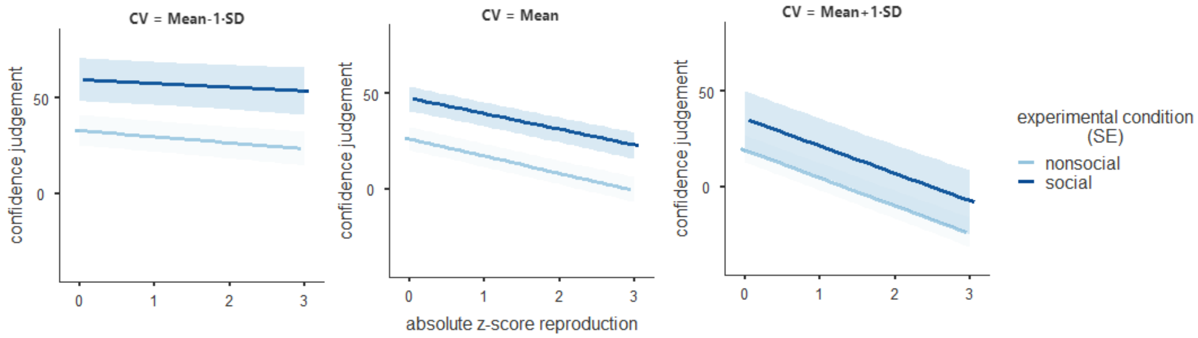


Figure S5.2: Linear relationship between absolute z-score transformed reproduced durations and confidence judgements across experimental conditions in three different levels of mastery levels. Low levels of CV indicate high mastery. Error bars represent standard error (SE).

4.2. Directional error magnitude judgements as outcome variable

Results revealed a statistically significant main effect of z-score transformed reproduced durations on directional error magnitude judgements (ß = 6.348, SE = 0.33, p < 0.001). Moreover, the interaction between the z-score transformed reproduced durations and experimental condition was statistically significant (ß$_{\text{social}}$ - ß$_{\text{nonsocial}}$ = -2.144, *SE* = 0.661, *p* < 0.001)

In order to address which of the two experimental manipulation's effect was closer to the "ideal observer's performance", we compared the simple slopes with the ideal-performance slope, which is 33.33 (=200/6). Although the 95% CI of nonsocial condition's simple slope

[2] The interaction between the absolute values of z-score reproduced duration, and CV on confidence judgements was statistically significant (ß = -66.023, *SE* = 14.94, *p* < 0.001).

does not include -33.33 (indicating a significant difference between the ideal-performance slope (Simple $ß_{social}$ = 5.28, *SE* = 0.466, 95% CI = 4.36, 6.19, $p < 0.001$; Simple $ß_{nonsocial}$ = 7.42, *SE* = 0.469, 95% CI = 6.5, 8.34, $p < 0.001$), the slope difference between the social and nonsocial conditions shows that the simple slope of the nonsocial condition is significantly closer to the regression line representing the ideal performance than that of nonsocial condition's ($ß_{ideal\text{-}performance}$ - $ß_{social}$ = 28.05; $ß_{ideal\text{-}performance}$- $ß_{nonsocial}$ = 25.91; see: Figure S5.3).

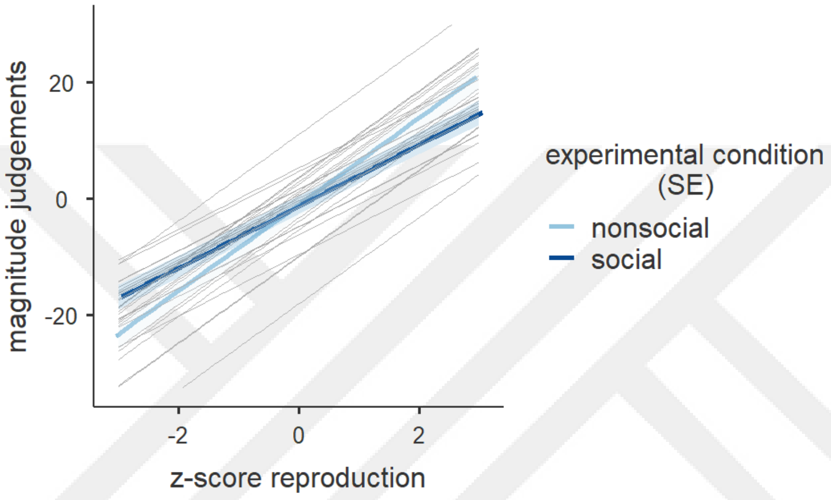


Figure S5.3: Linear relationship between the z-score transformed reproduced durations, and the directional error magnitude judgements across experimental conditions. Light gray lines represent individual slopes, whereas colored bold lines represent average slope of the sample across experimental conditions (n=15). Error bars represent standard error (SE) of the estimation.

To investigate the interaction between reproduced durations, experimental condition, and CV on directional error magnitude judgements, we performed a linear mixed effects model (Model S5.2).

Model S5.2:

directional error magnitude judgements ~ z-score reproduction * experimental condition * CV + (1|participants)

where ~ stands for "predicted from" and 1|participants stands for "random intercept across participants". The model depicted above consisted of all lower terms.

Results revealed a significant main effect of reproduced durations on directional error magnitude judgements (ß = 6.432, *SE* = 0.341, $p < 0.001$). However, the three way interaction between experimental condition, z-score reproduction and CV was not significant ($p > 0.05$, see: Figure S5.4)

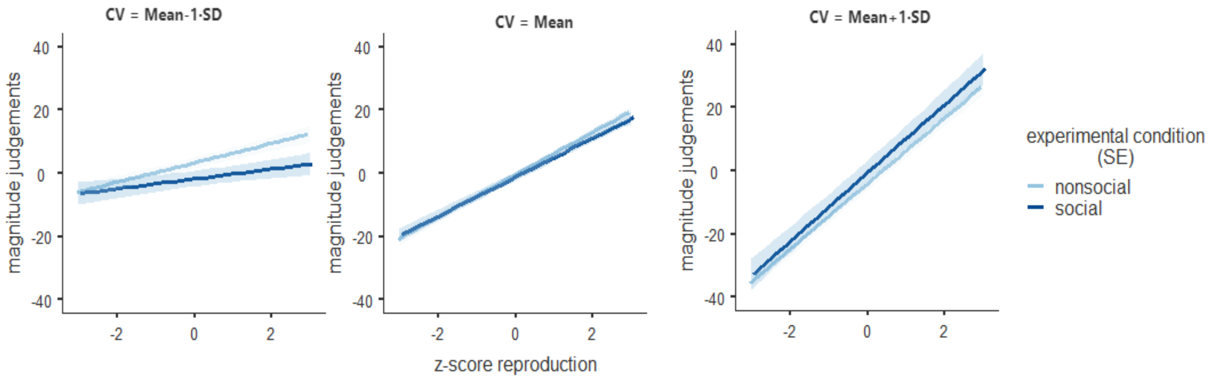
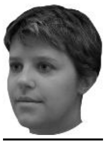
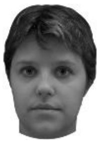
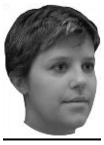
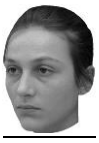
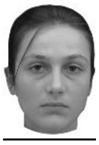
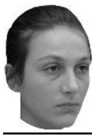
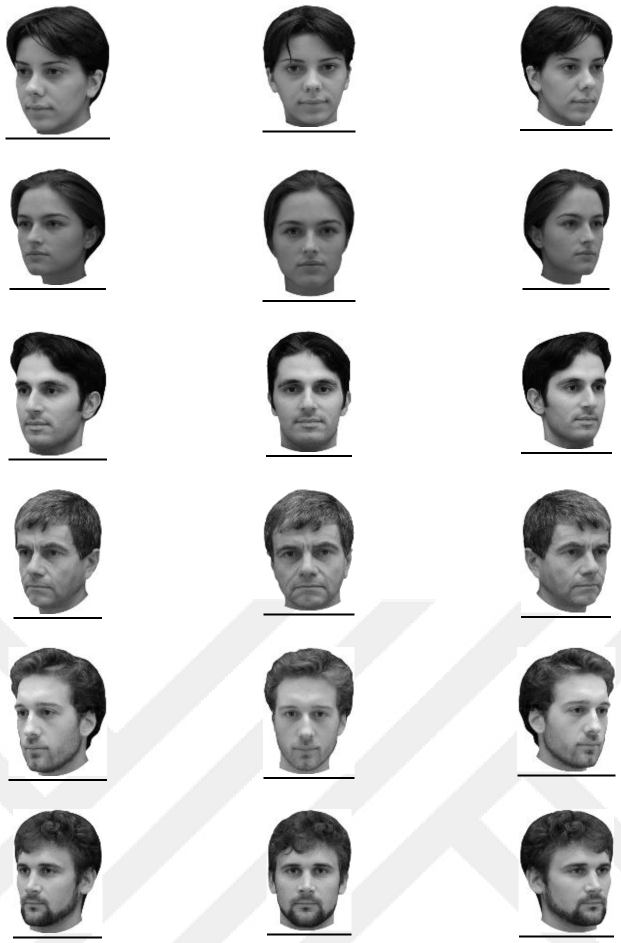


Figure S5.4: Linear relationship between z-score transformed reproduced durations and directional error magnitude judgements across experimental conditions in three different levels of mastery levels. Low levels of CV indicate high mastery. Error bars represent standard error (SE).

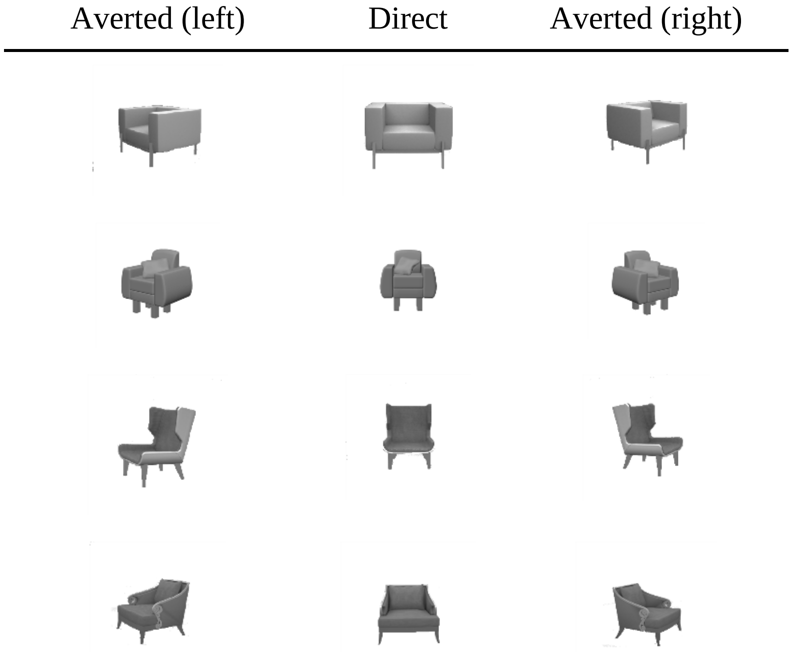
Appendices

Appendix 1: Social visual stimuli that are used in Experiment 1 (averted stimuli were discarded from the stimuli set and only stimuli in the middle column were used for Experiment 2)

Averted (left)	Direct	Averted (right)
		
		



Appendix 2: Nonsocial visual stimuli that are used in Experiment 1 (averted stimuli were discarded from the stimuli set for Experiment 2)



Appendix 3a: The original version of Fear of Negative Evaluation scale (Leary, 1983)

1. Sometimes I think I am too concerned with what other people think.

2. I worry about what kind of impression I make on people.

3. I am afraid that people will find fault with me.

4. I am concerned about other people's opinions of me.

5. When I am talking to someone, I worry about what they may be thinking of me.

6. I am afraid that others will not approve of me.

7. I am usually worried about the kind of impression I make.

8. I am frequently afraid of other people noticing my shortcomings.

9. I worry what other people with think of me even when I know it doesn't make any difference.

10. It bothers me when people form an unfavorable opinion of me.

11. Often worry that I will say or do the wrong things.

12. If I know that someone is judging me, it tends to bother me.

Appendix 3b: The Turkish standardized version of Fear of Negative Evaluation scale (as used in the current study, Çetin, Doğan, & Sapmaz, 2010)

1.Önemli olmadığını bilsem de başkalarının hakkımda ne düşündüğü beni endişelendirir.

2. İnsanların benimle ilgili olumsuz izlenimleri olduğunu bilsem bile bunu umursamam.

3. Çoğu zaman insanların benim kusurlarımı fark edeceklerinden korkarım.

5. Başkalarının beni onaylamayacağından korkarım.

6. Diğer insanların bende bir kusur bulacaklarından korkarım.

7. Diğer insanların hakkımdaki düşünceleri beni rahatsız etmez.

8. Birileriyle konuşurken benim hakkımda ne düşünecekleri ile ilgili endişelenirim.

9. Genellikle başkaları üzerinde nasıl bir izlenim bıraktığımla ilgili olarak endişe duyarım.

10. Eğer birisi benimle ilgili bir değerlendirmede bulunursa, bu beni çok fazla etkilemez.

11. Bazen diğer insanların hakkımda ne düşündükleri ile ilgili olarak fazla endişelendiğimi düşünüyorum.

12. Çoğunlukla yanlış bir şey yapacağım ya da söyleyeceğim diye endişelenirim.