Contextual Effects in Bayesian Inference of Interval Timing

Dissertation at the

Ludwig-Maximilians-Universität München

submitted by

Xiuna Zhu

March 2022

Supervisors:
Prof. Dr. Zhuanghua Shi (*1st Supervisor*)[1]
Prof. Dr. Hermann J. Müller (*2nd Supervisor*)[1]

[1] General and Experimental Psychology,
Ludwig-Maximilians-Universität München
Datum der mündlichen Prüfung: 11. 10. 2022

Table of Contents

1 General Introduction

1.1 Temporal context effect

Accurate timing is essential for proper actions in our daily activities. Although we have no specific sensory organ for timing, humans are well capable of time measurement. For instance, we can accurately follow musical rhythms, feel the beat and rhythm while listening to music, and swiftly avoid collisions while driving. Animals can accurately predict when the food will be supplied. Early classical internal-clock models (Gibbon et al., 1984; Treisman, 1963), such as scalar expectancy theory, use analogue of clock ticks generator and accumulator to explain many empirical findings, including the key signature of interval timing, the scalar property. That is, the variability of time estimation increases proportionally to the to-be-estimated time interval. The internal-clock models assume a pacemaker generates pulses that are transmitted to an accumulator through a switch. The switch is turned on when a to-be-timed interval starts and is turned off when the to-be-timed interval is off. A late developed attentional gating theory (Zakay & Block, 1996a, 1996b) further assumes that the switch control is influenced by attentional sharing.

Time perception in daily life is not always veridical to physical reality, even though humans are well capable of time measurement. Perception and perceptual strategies are seen as productions of evolution in a situated environment. Although this view is now a broad consensus in biology, research into the relationship between perception and the external world as early as 1860 prompted (Fechner, 1966) to pioneer the field of experimental psychology more broadly. As has been suggested in classical Perception Decision Action (PDA) loop framework (Hoffman et al., 2015), observers measure the external world with sensory organs to obtain perceptual representations of the external world. Nevertheless, the perceptual representation can be influenced by surrounding contexts. A large number of subjective distortions in time perception, including duration expansion and duration contraction, have been reported by various contexts, such as motion stimuli, arousal, internal states (Eagleman, 2008; Erickson & Erickson, 1958; Lee, 2017; Sackett et al., 2010; Teixeira et al., 2013; van Wassenhove et al., 2008). Most of us have consciously experienced situations in which time flies during a delightful event or slows down during a frightening event.

Very often, however, time distortion occurs without explicit knowledge of what contexts distort the time, given that the context is often not consciously known. One such example of time distortion is the Vierordt effect (Vierordt, 1868), which has been reported for more than a century and better known as the *central-tendency effect*: short intervals are overestimated and long intervals underestimated in a randomization test (Glasauer & Shi, 2021; Jazayeri & Shadlen, 2010; Lejeune & Wearden, 2009; Shi, Church, et al., 2013). In a typical randomized experiment of duration reproduction, participants are asked to reproduce a given duration after receiving a probe duration. The probe durations are usually randomized across trials. Though participants are only asked to reproduce the duration for a specific given duration, nevertheless, the reproduced duration is influenced by various *temporal contexts*, such as the past sampled durations (Glasauer & Shi, 2021; Vierordt, 1868), the tested range (Teghtsoonian & Teghtsoonian, 1978), and the spacing of the sampled durations (Brown et al., 2005; Penney et al., 2014; Wearden & Ferrara, 1995). It should be noted, however, the central tendency effect is not limited to time domain, it has also been found in a variety of sensory magnitude estimation, such as brightness, size of angles, and distance and angular (rotational-body) displacement (Hollingworth, 1910; Petzschner et al., 2012b; Petzschner & Glasauer, 2011; Teghtsoonian & Teghtsoonian, 1978), suggesting the central tendency effect is highly robust and applies to perception of many types of magnitude. Another closely related to the central tendency effect is the sequential dependence (Holland & Lockhead, 1968), sometimes also termed as the serial dependence (Fischer & Whitney, 2014; Kiyonaga et al., 2017). That is, the intensity of the previous stimulus can attract the perception of the current stimulus toward it. This local temporal context has been shown to be closely related to the central tendency effect in time perception (Glasauer & Shi, 2021, 2022). The central tendency effect can be partially explained by the dynamic updating of local temporal context.

In addition to the temporal contexts, *non-temporal contexts* can also distort subjective time. For examples, our estimation of elapsed time is biased by internal bodily states, including mental load (Angrilli et al., 1997), attention (Polti et al., 2018), emotional state (Stetson et al., 2007), and external stimulus properties, such as size, brightness, probability of the occurrence (Eagleman, 2008; Kanai & Watanabe, 2006). The original internal clock-models (Gibbon et al., 1984; Treisman, 1963) only assume the representation of time in memory is directly correspondent to the temporal 'ticks' accumulated in the memory. That is, the duration representation more or less veridically reflects sensory inputs. Thus, influences of non-temporal contexts are often explained in terms of the modulation of the internal clock speeds,

such as large stimuli or high arousal increasing the speed of the internal clock. Yet, merely modulation of the clock speed could not explain various findings by attentional modulation. In extension to the classical internal clock-model, attentional gating theory (Zakay & Block, 1996a, 1996b) and attentional sharing account (Fortin & Massé, 2000; Fortin & Rousseau, 1998) includes the attention and memory in the internal clock model. For example, the attentional sharing account assumes that the accumulation of internal 'ticks' can be varied by the modulation of the attention to the timing process. When the attention is allocated to other non-temporal tasks during the time encoding process, some 'ticks' may be lost due to the attentional sharing, causing an underestimation of the perceived duration. By contrast, when the attention is shared during the duration reproduction phase, some 'ticks' could be lost during the monitoring of the elapsed time, subsequently causing over-reproduction of the perceived duration.

In short, both temporal and non-temporal contexts could implicitly influence subjective time. Over the past half century, much empirical evidence has been accumulated for subjective time perception. Although the classical internal clock models (Gibbon, 1977; Gibbon et al., 1984; Treisman, 1963) and their extensions, such as attentional gating / sharing accounts (Zakay & Block, 1996a, 1996b), can qualitatively explain empirical findings, quantitative predictions of various contextual biases in time perception have only been achieved in the past decade using the probabilistic approach (Gu et al., 2016; Jazayeri & Shadlen, 2010; Shi, Church, et al., 2013; Shi & Burr, 2016), more precisely, Bayesian inference, which is detailed in the next subsection.

1.2 Probabilistic interpretation of the central tendency bias

The most prominent prediction of the classical internal clock models (Treisman, 1963; Treisman et al., 1990), including the scalar expectancy theory (Gibbon, 1977; SET, Gibbon et al., 1984), is the scalar property, also known as Weber scaling. That is, the variability of a perceived interval proportionally increases with the given interval. The SET assumes that the variability in the pacemaker across trials is the main key factor that leads to the scalar property. However, the classical models fall short in interpreting various context-induced temporal biases. It can not explain the classical Vierordt effect - an effect being observed more than 150 years ago (Glasauer & Shi, 2021; Vierordt, 1868). An early attempt is made by Helson (Helson, 1964) with his adaptation-level (AL) theory. According to this theory, Vierordt effect arises from the background context - such as the sampled intervals. The perception of a stimulus is

assimilated to the background context. In the past decade, a similar approach has been made in probabilistic Bayesian inference frameworks (Jazayeri & Shadlen, 2010; Petzschner & Glasauer, 2011; Shi, Church, et al., 2013; Shi & Burr, 2016). One common idea is that the percept is an integration of the prior knowledge (the background context in the AL theory) and the sensory inputs, the original idea came from Helmholz's theory of perception as unconscious inference. Importantly, the Bayesian approach assumes the integration is according to Bayes rules, the weights of the prior and sensory inputs depending on their correspondent reliability. Suppose the sensory input is $T_s \sim N(\mu_s, \sigma_s)$, and the prior of the sampled intervals follows $T_p \sim N(\mu_p, \sigma_p)$. According to the Bayesian inference, the optimal estimation would be:

$$T_o = w_p\mu_p + (1 - w_p)\mu_s,$$

where the weight w_p is proportional to the reliability of the prior. That is, $w_p = \frac{1/{\sigma_p}^2}{1/{\sigma_p}^2+1/{\sigma_s}^2}$. The estimate T_0 is called optimal because the variability of T_0 is the minimum among any linear combination of the prior T_p and sensory input T_p.

Under such quantitative approaches, researchers have successfully predicted the classical central tendency effect (Jazayeri & Shadlen, 2010), group differences in the central tendency effect (Cicchini et al., 2012), as well as other effects, such as time-order effects (Bausenhart et al., 2014; Dyjas et al., 2013), variations in the scalar properties (Ren et al., 2021). More recently, the static Bayesian framework has also been extended to dynamic Bayesian framework in explaining the dynamic updating of the priors in time perception (Glasauer & Shi, 2021, 2022). Dynamic Bayesian frameworks do not assume the prior is static, rather suggest that the prior is dynamically updated through trial-to-trial learning. One typical updating mechanism that can well explain such a dynamic process is the Kalman filter process (Petzschner & Glasauer, 2011), which has been recently used to explain the origin of the Vierordt effect (Glasauer & Shi, 2021). In that study, Glasauer and Shi (2021) showed that the central tendency effect depended on the inter-trial volatility, rather than the sampled distributions across the whole session. Given that the sample distributions were the same for the two volatility conditions, the traditional Bayesian integration fails to explain the difference of the central tendency in the two volatility conditions.

The Bayesian approach has a stark contrast to the classical internal clock models. The Bayesian approach usually does not assume any 'module'-like cognitive processing stages, which the internal clock models do, rather it assumes a quantitative integration process. In a

review paper, Shi et al. (2013) bridged this gap by linking the Bayesian approach to the classical internal clock models (see Figure 1, adapted from Shi et al. 2013).

As shown in Figure 1, internal clock models are composed of three stages: Clock, Memory, Decision. Shi et al. (2013) linked them to the Bayesian integration processes, respectively. Specifically, in the clock stage an accumulator receives and restores a pulse or a sequence of pulses from a pacemaker. This stage is corresponding to the likelihood of the sensory inputs in the Bayesian framework. The memory stage involves two memory systems, working memory and long-term reference memory. The accumulated pulses within the counter process the duration and store them in the working memory, then retrospective comparisons occur between stored interval representations within reference memory. In the Bayesian framework, the prior information is likely stored in the reference memory, representing the observer's beliefs about intervals in a probabilistic manner. As denoted by a dashed black arrow, prior is updated by the current posterior. The posterior distribution that expresses the probability distribution of the current estimate is derived by the combination of updated prior beliefs with likelihood (denoted by a dashed red arrow). In the decision stage, responses are made based on specific comparison rules, such as the difference between the two temporal memory sources is less than a threshold. In the Bayesian framework, this optimization is achieved by minimizing the loss function.

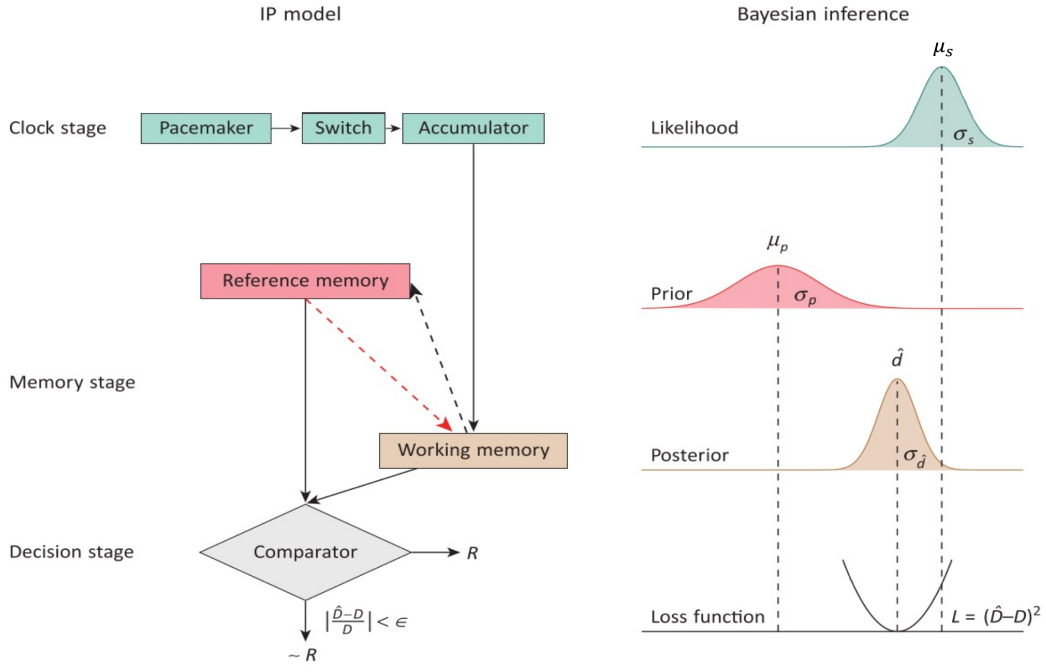

Figure 1 *Schematic representation of Bayesian inference of interval timing and the information process(IP) model (adapted from Shi, Church, et al., 2013).*

With Bayesian integration approaches, a number of studies have shown that various time distortions can be well explained as an optimal fusion of noisy sensory information with expectations based on previous experience (Cicchini et al., 2012; Roach et al., 2017; Shi, Church, et al., 2013; Shi & Burr, 2016). Yet, there are a number of issues that remain controversial, which we will explain in the following subsection.

1.3 Open issues related to temporal context effect in interval timing

Although there is now ample evidence of how the perceptual process can be well explained as an optimal integration of sensory information and prior experience shaped by contexts, it is not clear how the effect of surrounding context would take place on the prior information and sensory measurement, leading to impact on time estimation and central tendency effect bias.

1.3.1 Ensemble perception of temporal sequence

In everyday life, humans are continually confronted with piles of information from the outside world which are often redundant and similar, so our sensory system has adapted and developed an efficient system through extracting more 'intuitive' statistical properties to cope with sensory information overload and the limited capacity of working memory (Ren, 2020; Whitney & Yamanashi Leib, 2018). For instance, we can quickly judge whether a box of grapes are fresh and uniform in size at first glance without looking carefully at each grape, which reflects that we are capable of extracting statistical properties in the visual domain (Chong & Treisman, 2005). Ample studies have revealed that ensemble statistics can be rapidly acquired from a set of items through ensemble perception (Alvarez, 2011; Chen et al., 2018; Haberman & Whitney, 2012; Webster, 2014; Whitney & Yamanashi Leib, 2018). Using ensemble representations of various types of features, such as the average moving speed (Watamaniuk & Duchon, 1992), the average size (Marchant et al., 2013), and summary representation of the facial characteristics (Haberman & Whitney, 2009), is an efficient representation of a set with similar items. It should be noted, perceptual averaging is not confined to a simultaneously presented set of items (inherently 'parallel' visual domain), but also takes place for sequentially presented event, such as object weights and auditory frequencies (Curtis & Mullin, 1975; Piazza et al., 2013; Schweickert et al., 2014). However, how ensemble statistics influences the temporal processing, such as temporal-bisection task, remains open.

The temporal-bisection task was first adapted to research on animal timing (Gibbon, 1977; Gibbon et al., 1984) and later expanded to studies on human timing (Allan & Gibbon, 1991; Wearden, 1991). Initially, the focus of using temporal-bisection was on how humans and other animals make interval comparisons and researchers believed that only the sample durations matter in a temporal-bisection task. Subsequently, they found the bisection point (BP, the duration which produces 50% 'long' responses) in temporal-bisection tasks is subject to various contextual factors, including tone frequencies (Brown et al., 2005), spacing (Penney et al., 2014; Wearden & Ferrara, 1995) and skewness(Schweickert et al., 2014). For example, Brown and colleagues (2005) argued that only using the short and long standards is not sufficient to explain the contextual bias, as they showed the same duration in two different sets with the same range was perceived differently when the orders of the same duration in the two sets were different. Accordingly, Brown et al. (2005) proposed that both the relative range and the order are critical, as already shown in the previous 'Range Frequency Theory' (RFT; Parducci, 1963), and the subjective judgment of a test interval is based on the weighted average of its distance to the (short and long) standards and its rank order in the range, and they termed this as the temporal RFT (TRFT; Brown et al., 2005). Although TRFT successfully captures the biases in the temporal-bisection task, the order of the duration requires full knowledge of the whole set. To store all durations and their relative orders of a set is rather not economical, especially when the set size increases, the amount of stored information rapidly expands. It becomes impractical when the whole set is not known. One possible alternative account might be that instead of storing individual durations with their order information, observers adopt ensemble statistics of the sampled durations from a longer-term memory (Cicchini et al., 2012; Shi, Church, et al., 2013) to estimate the bisection point in the bisection task. Remembering individual items is difficult, whereas representing ensemble statistics is quick and intuitive (Alvarez, 2011; Ariely, 2001; Ren et al., 2020).

To identify the best possible account of how the ensemble context modulates performance in the temporal-bisection task, we assume that observers likely compare a perceived duration to the ensemble prior (i.e, ensemble representation of the distribution of sampled intervals), rather than storing individual items for later comparison. Ample evidence has been found that using ensemble statistical information would be efficient in dealing with a set of sample stimuli (e.g., Alvarez, 2011; Cohen et al., 2016; Ren et al., 2020; Whitney & Yamanashi Leib, 2018). Combining with the IP model (see Figure 1 left panel) introduced above, during handling a sequence of auditory stimuli, the current perceived duration is stored

in the working memory and compared to the temporal standards stored in reference memory, afterwards the sample distribution (i.e., prior) will be updated by the current estimation. The prior distribution could be treated as the temporal ensembles, providing efficient access to the past sample stimuli, rather than maintaining the full information of individual items (Ren et al., 2020; Whitney & Yamanashi Leib, 2018).

1.3.2 Temporal context under memory load

According to the SET theory and IP models (see Sec. 1.2), the integration of the sensory measure and the prior is involved with the working memory (Shi, Church, et al., 2013). Hence, the demands on working memory capacity (cognitive load) would impact sensory measurement and internal prior. Meanwhile, reproducing the perceived duration also requires continuously monitoring the passage of time. Thus, both working memory and attention are essential components for time perception. The influence of attention in subjective time has already been considered in the attention-gating account (Zakay, 1989; Zakay & Block, 1996a). It has been shown that when a person is working on a demanding task, time passes quickly, while when a person is working on an easy task, time passes slowly (Block et al., 2010). One common explanation of this phenomenon is that when attention is allocated to other non-timing tasks, such as on a demanding task, some ticks generated by the internal clock haven't been counted and passed through to the accumulator (i.e., attentional lapse). One common task to investigate this is the dual-task paradigm with one timing task and the other non-timing task (Fortin, 2003; e.g., Fortin & Rousseau, 1998). Additionally, memory representation of a time interval could be mixed with other contexts, which has been shown in various experimental paradigms, known as memory mixing effect (Gu et al., 2016; Gu & Meck, 2011; Penney et al., 2000; Stanfield-Wiswell & Wiener, 2020). It should be noted, memory mixing not only occurs when the memory trace of a previous time interval influences the processing of the current time interval (Coull et al., 2013; L. A. Jones & Wearden, 2004), but also happens in a dual-task paradigm consisting of a secondary visual working-memory task and a primary duration production-reproduction task (Macar et al., 1994). Due to additional tasks in the dual-task paradigm, the uncertainty of the sensory input increases. Subsequently, according to Bayesian inference (Jazayeri & Shadlen, 2010; Shi, Church, et al., 2013; Shi & Burr, 2016), higher cognitive load would yield a stronger central tendency bias in a timing task with a set of durations.

Some studies have revealed two separable effects of cognitive load on duration encoding and duration reproduction (Fortin, 2003; Fortin & Rousseau, 1998; see also on non-timing tasks, e.g., Glasauer et al., 2007). For example, Fortin and Rousseau (1998) used a dual-task design with the second task of a Sternberg memory task (memory set of 1, 3, or 6 digits), presented successively prior to the primary temporal reproduction task, where the latter consisted (on a given trial) of an initial duration-production phase (with two beeps demarcating a duration) followed by the duration-reproduction phase (two taps generated by the participants). The memory probe (a digit that was or was not part of the memory set) was shown either during the duration-production or the duration-reproduction phase, and the response to the memory task (either positive or negative) was to be issued pressing two different response keys, one for probes presented during the production phase or, respectively, the other for probes presented during the reproduction phase. Fortin and Rousseau found the reproduced duration to be shortened when the memory probe was presented during the production phase, but lengthened when it was shown during the reproduction phase. They took this finding to support an attention-sharing account (Fortin & Rousseau, 1998; Macar et al., 1994), according to which attentional resources are shared between the timing process and other, non-temporal cognitive processes. That is, when attention is diverted away from the primary task by other non-temporal tasks during the duration encoding phase, perceived duration will be shortened. In contrast, when the non-temporal process interferes in the duration reproduction phase, the reproduced duration will be lengthened, because of the lapse in the monitoring of the passage of time. It should be noted, however, the role of working memory on Bayesian inference of time perception has been largely neglected in the literature. Up to date, there is no consensus on the question about cognitive load influence in both the sensory estimate and the prior representation in terms of their means and variances (Jazayeri & Shadlen, 2010; Shi, Church, et al., 2013; Shi & Burr, 2016). Notably, the question how cognitive load impacts the uncertainty of the magnitude encoding and prior representation is unknown.

To explore further whether the central tendency would be differentially influenced by the cognitive load on the encoding and reproduction phases in duration judgments, in the thesis we conducted a series of experiments using the dual task to identify how cognitive load would influence internal prior and the perceived and reproduced duration in both the mean and the variability. We hypothesize that increasing cognitive load during the sensory encoding stage would not only lead to a general underestimation of the duration (Fortin & Rousseau, 1998), but also to decrease the reliability of the estimate, yielding a strong central tendency bias. By

contrast, introducing cognitive load during the reproduction stage would lengthen the reproduced duration not only to cause an overestimation of reproduction mean, and may also increase the variability of the reproduction. However, given that no additional cognitive load is imposed on the sensory encoding stage, thus the 'memory-mixing' stage would be unaffected by the cognitive load introduced during duration reproduction. When cognitive load remains high during both the duration production and reproduction phases, the underestimate from the production and the overestimate from the reproduction may cancel each other, at least to some extent and so we may not be able to observe a general bias. However, increasing the uncertainty in the sensory representation may cause a stronger central-tendency effect – a similar pattern to that has recently been reported in a non-temporal task (Allred et al., 2016).

1.3.3 Prior integration in multi-prior temporal context

Humans have developed efficient and sophisticated sensory processing systems to overcome information overload. For example, we can efficiently use ensemble summary to represent a set of items or events (Ren, 2020; Whitney & Yamanashi Leib, 2018), and use acquired contextual knowledge to compensate uncertainty during perception and improve the reliability of sensory estimation (Berniker et al., 2010; Cicchini et al., 2012; Roach et al., 2017; Shi & Burr, 2016). The learned contextual knowledge (represented as 'prior') is useful given that our world is relatively stable such that in most cases the context can quickly guide us for a better performance. According to Bayesian inference, the central tendency effect in magnitude estimates has been interpreted as the sensory estimates is assimilated to the prior in the decision stage (Fritsche et al., 2020; Jazayeri & Shadlen, 2010; Roach et al., 2017; Shi, Church, et al., 2013; Westheimer, 2008).

In time perception, the duration production-reproduction paradigm is widely adopted to study how the prior influences time judgment (Shi, Church, et al., 2013; for reviews, see Shi & Burr, 2016). For example, Jazayeri and Shadlen (2010) showed that the central tendency effect (Vieordt effect) can be explained by assuming integration of a uniform prior with the sensory input. Later Cicchini et al. (2012) suggested that a Gaussian prior could also do a good job in predicting the central tendency bias. It should be noted that most studies investigating the central tendency effect use one single prior with unimodal sensory inputs. Regarding the timing system, it remains controversial whether humans use a single central dedicated clock or distributed multi-clock systems (Buhusi & Meck, 2009). There is evidence that different modalities may have different timing systems (Buhusi & Meck, 2009; Ivry & Richardson, 2002;

Paton & Buonomano, 2018). Ample evidence suggests that multiple timing mechanisms exist in the brain across and within sensory modalities (Gau & Noppeney, 2016; Heron et al., 2012). However, up to date it is not clear if the prior representation of multimodal timing is also distributed for different sensory modalities. The question is important given that the prior representation, whether modality-specific or modality-independent, could help us understand how temporal context is represented in our brain.

Recent studies have shown that the prior representation might not be like the distributed clock systems, each having their own priors. For example, a recent study by Roach et al. (2017) showed that spatial separation or modality of two different ranges of durations for the duration reproduction task did not yield separate priors. Only when the timing tasks of two separate ranges of durations were different (e.g., one for reproduction and one for bisection), they observed separated priors. A recent study by Zang et al. (Zang et al., 2022) showed that multiple separated priors could only be developed when two ranges (short and long) were clearly separable in the range. When two ranges were overlapped, maintaining and updating two similar priors could be costly, even though two ranges of durations were modality-specific. This is consistent with the results of Roach et al.'s study (Roach et al., 2017) that observers formed a unified prior (global prior) by generalizing across the two interleaved stimulus sets instead of separate priors for each stimulus set in multi-prior context, and they did not distinguish the different stimulus sets in the random interleaving condition even if the stimuli used for the two interleaved duration distributions are clearly discriminable.

In addition to the context of sampled distribution of durations, other contexts may also influence duration judgments. For example, auditory duration is often perceived longer than the visual duration (Shi, Ganzenmüller, et al., 2013; Wearden et al., 1998). Binary categorical cue (Petzschner et al., 2012a) or categorical judgments (Luu & Stocker, 2018) may also influence final decision making. For instance, Petzschner et al.(2012b) used overlapping short and long distance in a production-reproduction task with three different experimental conditions ("blocked-ranges, no cue", "interleaved range, no cue", and "interleaved range, cue") and demonstrated two separate priors in the "interleaved range, cue" condition due to the presence of the categorical cue ('Short' vs. 'Long'). It then begs the question how those different contexts (priors) integrate together with the sensory inputs. Would it be a flat integration (i.e., combining priors and sensory inputs linearly) or combined priors and sensory inputs in a hierarchical order? If the latter is true, what kind of structure order would be? Suppose the internal prior consists of a multi-level structured prior, constituting something like

a local-global hybrid prior or a generalization of the global prior and local prior information, the question of how Bayesian optimization of time perception by the incorporation of multiple priors and sensory measurement into the time estimation process remains unknown. Furthermore, the interactive influence between priors in the multi-prior context can be posed rigorously. We aimed to extend the basic Bayesian observer model to multi-prior temporal contexts, and further uncover the rules governing how learned multiple prior knowledge are grouped together.

1.4 Aims of this thesis

The goal of the current dissertation is to investigate underlying mechanisms of temporal context effect in the framework of Bayesian Inference in subjective interval timing. To tackle this issue, classical behavioral investigations and Bayesian modeling are employed.

In Chapter 2.1, by adopting and modifying the temporal bisection paradigm of Penney's study (2014), three temporal bisection experiments were carried out to investigate the impact of ensemble statistics in time perception by manipulations of stimulus spacing, distribution means, and variances, in order to identify how the ensemble context modulates performance of the bisection task. Hierarchical Bayesian modeling was applied to fit the behavioral data to validate if the shape of the interval distribution should be taken into account.

Subsequently, in Chapter 2.2, inspired by attentional sharing account (Fortin & Rousseau, 1998; Macar et al., 1994), by adopting and modifying a dual-task duration reproduction, which consists of a secondary visual working-memory task (with low, medium, or high load) and a primary duration production-reproduction task, we asked the question about how memory load would influence the production and reproduction stages in duration reproduction experiment. In addition, taking the impact of cognitive load into consideration, a general Bayesian computational framework was proposed and developed, to validate how the sensory estimate is integrated with the prior during duration encoding and duration reproduction stages.

Finally, in Chapter 2.3, extending our Bayesian observer model to multiple prior temporal contexts by taking global prior knowledge into consideration, the rules governing prior integration were further uncovered through investigating the role of multi-prior temporal contexts in time perception, to explore hierarchical priors integration process under multi-prior context in duration production and reproduction tasks.

2 Cumulative Thesis

The current cumulative thesis is made up of three separate studies: one peer-reviewed and published paper (2.1), two manuscripts that have been submitted (2.2-2.3). These three studies are included in the following chapter, each of which is followed by a brief statement about the contributions of the authors concerned.

2.1 Temporal bisection is influenced by ensemble statistics of the stimulus set

CONTRIBUTIONS

The authors declare no competing interests. Z.S. conceptualized the research; X.Z. and C.B. performed the experiments; X.Z. and C.B. wrote the manuscript, together with H.M. and Z.S.

REFERENCE

 The official citation that should be used in referencing this article is Zhu, X., Baykan, C., Müller, H.J. et al. Temporal bisection is influenced by ensemble statistics of the stimulus set. *Atten Percept Psychophys* 83, 1201–1214(2021).https://doi.org/10.3758/s13414-020-02202-z

Temporal bisection is influenced by ensemble statistics of the stimulus set

Xiuna Zhu[1], Cemre Baykan[1], Hermann J. Müller[1], Zhuanghua Shi[1,*]

1. General and Experimental Psychology, Department of Psychology, LMU Munich, Munich, Germany

* Correspondent author

ABSTRACT

Although humans are well capable of precise time measurement, their duration judgments are nevertheless susceptible to temporal context. Previous research on temporal bisection has shown that duration comparisons are influenced by both stimulus spacing and ensemble statistics. However, theories proposed to account for bisection performance lack a plausible justification of how the effects of stimulus spacing and ensemble statistics are actually combined in temporal judgments. To explain the various contextual effects in temporal bisection, here we develop a unified *ensemble-distribution account (EDA)*, which assumes that the mean and variance of the duration set serve as a reference, rather than the short and long standards, in duration comparison. To validate this account, we conducted three experiments that varied the stimulus spacing (Experiment 1), the frequency of the probed durations (Experiment 2), and the variability of the probed durations (Experiment 3). The results revealed significant shifts of the bisection point in Experiments 1 and 2, and a change of the sensitivity of temporal judgments in Experiment 3 – which were all well predicted by EDA. In fact, comparison of EDA to the extant prior accounts showed that using ensemble statistics can parsimoniously explain various stimulus set-related factors (e.g., spacing, frequency, variance) that influence temporal judgments.

Keywords: temporal bisection, stimulus spacing, central tendency, ensemble perception

INTRODUCTION

We, humans, have the ability to perceive the passage of time relatively accurately. Our sense of time allows us to adapt to and interact with a dynamic external world. As has been suggested in classical 'internal-clock' models (Gibbon et al., 1984; Treisman, 1963), our ability to time (experienced) events in the external world is based on an internal timer. Although there is no physical timer in our brain, behavioral studies have shown the internal-clock model can explain many empirical findings and predict the key feature of time perception: the scalar property (i.e., the Weber scaling). However, more and more evidence show that, even though we are well capable of time measurement, we are still prone to biases in our timing that depend on both internal states (e.g., mental load, attention, emotional state) and external contexts (Allman et al., 2014; Allman & Meck, 2012; Pronin, 2013).

One prominent contextual bias in time perception, which has been puzzling for more than a century and half, is the central-tendency effect: duration judgments are assimilated to the center of the sample durations. Thus, for example, when asked to reproduce a series of time intervals, participants judge long durations as being shorter and short durations as being longer than they actually are. This was first, and accidentally, discovered by Karl von Vierordt (Lejeune & Wearden, 2009; Vierordt, 1868), who misused the "method of average error" that Fechner had devised by implementing randomization instead of repeated measures (Glasauer & Shi, 2018). The central-tendency effect is one classical example showing that the ensemble mean derived from long-term memory of sampled stimuli strongly influences perceptual judgments. Recent studies have suggested that ensemble statistics can be rapidly computed from a set of variant objects or a sequence of events (Alvarez, 2011; Ariely, 2001; Chen et al., 2018; Whitney & Yamanashi Leib, 2018). For example, we can quickly estimate the average size of apples in a basket, or the average tempo of a piece of music. It has been suggested that the use of ensemble statistics is beneficial by enhancing the reliability of sensory estimates (Alvarez, 2011). With regard to the central-tendency bias, this has been confirmed by Bayesian modeling: in duration judgments, for instance, the central-tendency bias is well predicted by optimal integration of the sample distribution and the sensory input (Jazayeri & Shadlen, 2010; Raviv et al., 2012; Shi et al., 2013; Shi & Burr, 2016).

While influences of ensemble statistics on time perception have been demonstrated mainly in studies of duration reproduction or temporal averaging (Acerbi et al., 2012; Burr et al., 2013; Chen et al., 2018; Ren et al., 2020; Zimmermann & Cicchini, 2020), duration context plays a critical role in duration comparison (Fründ et al., 2014; Rhodes & Di Luca, 2016;

Zimmermann & Cicchini, 2020), such as in the temporal-bisection task. In a typical temporal-bisection task, participants are given one *short* and one *long* duration as standards and they are asked to judge whether a given duration is closer to the *short* or the *long* standard (Allan & Gibbon, 1991; Raslear, 1985). Initially, researchers thought only the short and long standards matter in the temporal-bisection task, given that the task is to compare a sample duration to the both standards. However, it turned out that the sample durations themselves matter significantly (Brown et al., 2005; Penney et al., 2014; Wearden & Ferrara, 1995). For example, Wearden and Ferrara (1995) found that with the same short and long standards, a logarithmically spaced duration set, as compared to a linearly spaced set, had a lower *bisection point* (the time point that is subjectively equally distant to the short and long standard) – an effect that has been referred to as 'spacing effect'. The main account for the spacing effect holds that the temporal-bisection task does not really involve comparing a sample duration to the short and long standards, but rather to a reference point *M* somewhere near the geometric or the arithmetic mean of the two standards (Wearden & Ferrara, 1995). Brown and colleagues (2005) subsequently found that using one reference point is not sufficient to explain the contextual bias. Rather, other factors, in particular the rank of the duration in the sampled set, do also matter, that is: the same duration in two different sets (with the same short and long standards) led to different bisection points when the order (percentile) of the duration in the two sets was different. Drawing on 'Range Frequency Theory' (RFT; Parducci, 1963) to combine the two factors of the relative range and the order of the sampled durations, Brown et al. (2005) proposed that the subjective judgment of a given time interval change is based on the weighted average of its range position (i.e., how far it is from the short and long standards) and its rank order in the distribution. Although temporal RFT (TRFT; Brown et al., 2005) successfully captures the biases in the temporal-bisection task, it still lacks a plausible theoretical explanation of how we actually process ranks and allocate weights to the relative range position and the relative rank. This approach would require observers to store individual durations and their relative orders in the set, which becomes extremely difficult, if not impossible, when the set size increases, and even worse when the same duration may be perceived differently across trials.

One possible alternative, and straightforward, account might be that instead of storing individual durations to calculate their orders, observers use ensemble statistics from a longer-term memory of the sampled durations (Cicchini et al., 2012) to estimate the *M*-reference (i.e., bisection point) in the bisection task. Remembering individual items is difficult, whereas representing ensemble statistics is quick and intuitive (Alvarez, 2011; Ariely, 2001). Use of

ensemble representations has been shown for various types of features, such as the average speed of moving objects (Watamaniuk & Duchon, 1992), the average size of objects (Marchant et al., 2013), and the average emotional (facial) expression of a crowd of people (Haberman & Whitney, 2009). Given the limited capacity of attention and memory we have to consciously identify and remember the plethora of objects and events we continually encounter, ensemble representations provide us with a ready means to bolster our perceptual experience (for reviews, see Cohen et al., 2016; Whitney & Yamanashi Leib, 2018). In the temporal domain, it has been shown that people can learn ensemble statistics of time intervals up to the third central moment (i.e., mean, variance, and skewness) and use these statistics in their decision-making (Acerbi et al., 2012). Thus, in the temporal-bisection task, observers likely compare a perceived duration to the ensemble representation of the sampled distribution, rather than storing (and adjusting) individual ranks for later comparison. That is, acquiring the ensemble statistics of the test intervals, observers make bisection judgments according to the location of a given test interval within the learnt distribution. We refer to this as the 'Ensemble-Distribution Account' (EDA).

One strong difference of EDA to previous proposals (e.g., the spacing account) is that EDA takes the shape of the distribution into account in making bisection decisions. Accordingly, EDA would predict a shift of the bisection point when the shape of the distribution changes while the spacing of the probe durations remains the same. In addition, EDA would predict the variance of the ensemble statistics to influence the difficulty of temporal judgment (measured by the slope of the psychometric curve). On these grounds, we conducted three experiments to test the predictions of the ensemble-distribution account. Specifically, Experiment 1 was designed to examine for the shift of the bisection point in sets with positively skewed (PS) versus negatively skewed (NS) spacing. In this regard, EDA makes the same prediction as, and so would be indistinguishable from, the spacing and TRFT accounts. Experiment 2 further examined the bisection task with equally spaced durations under two skewed frequency-distribution sets (see Figure 1 for details): ascending frequency (AF) and descending frequency (DF). Given that the two sets have different ensemble means, we expected EDA to be able to predict, and account for, the difference in bisection points between the two conditions. Finally, Experiment 3 manipulated the variability of the sample distributions while keeping the mean of the distributions the same, by introducing a U-shaped and an inverted T-shaped set, with the former having a greater variance than the latter. According to EDA, the variance would influence the difficulty of the bisection (reflected in the JND), rather than the bisection point (reflected in the PSE). Additionally, we applied

hierarchical Bayesian modeling to the behavioral data according to various assumptions of how temporal bisection may be performed. The aim of the model fitting and comparison was to look at the data patterns obtained in the three experiments with respect to the manipulations of stimulus spacing, distribution means, and variances, to identify the best possible account of how the ensemble context modulates performance of the bisection task.

METHODS

Participants

Forty-five university students with normal hearing took part in three experiments (15 each in Experiments 1, 2, and 3; 25 females; mean age: 25.5 years). The sample size was determined based on the prior study of (Penney et al., 2014). Although they did not report effect sizes, we calculated $\eta_g = 0.27$ based on their report of a one-way ANOVA test for the auditory condition. With $\alpha = .05$, $1 - \beta = .85$, and a within-subject (repeated-measures) ANOVA design, the sample size required for replicating this effect is 9 observers. Taking a conservative approach, we opted for a sample size of 15 participants. All participants gave written consent according to the institutional guidelines prior to their participation and were paid 9 Euro per hour for their service. The study protocol was approved by the LMU Faculty of Pedagogics & Psychology Ethics Board. All participants were naive as to the purpose of the study.

Stimuli and apparatus

The experiments were conducted in a sound-reduced and moderately lit test room. Stimuli were generated by Psychtoolbox-3 (Kleiner et al., 2007) based on MATLAB R2014a (The MathWorks Inc). Auditory stimuli were generated by PsychPortAudio on a HP ProDesk computer and presented through the loudspeakers. Participants gave their responses by pressing the left or right arrow keys on the keyboard. Experimental instruction and feedback information were presented on a CRT monitor.

Procedure

We adopted the bisection task in all three experiments. Participants were familiarized with the task in two practice blocks prior to the main experiment (56 trials per block in Experiments 1 and 2; 72 trials per block in Experiment 3). The practice blocks involved the same procedure as the experiments proper, except that (i) all test intervals were uniformly

distributed (i.e., equally frequent), whereas the two distributions compared and contrasted in the formal Experiments 2 and 3 were non-uniform; and (ii) response feedback, referenced to the average of the short and long standards (see next paragraph), was provided on each trial, whereas no feedback was given in the formal experiments. Note that in Experiment 1, the intervals presented during practice had the skew inherited from the duration spacing in the formal experiment (see subsection 'Experiment 1' below for details), but the feedback reference was the same for all practice trials. Thus, if there were any effects of the training conditions in the practice blocks– in particular, assimilation of the PSE to the (common) feedback reference or sharpening of the psychometric curve – they would be the same for the two conditions that we manipulated in the experiment proper and so work against finding the predicted differential effects between the two conditions in the formal experiment. Thus, any differential effects we observe are unlikely confounded by practice effects.

A trial started with a visual fixation marker and a brief beep (20 ms, 1000 Hz, 60 dB), followed by a blank display of 500 ms, prompting participants to get ready for a new trial. Next, a white-noise stimulus (60 dB) was presented for a given duration randomly selected from a predetermined set, ranging from 400 to 1600 ms (seven test intervals in Experiments 1 and 2, eight intervals in Experiment 3; see next paragraph for details). Immediately following the offset of the white-noise stimulus, a display with a question mark ("?") was shown prompting participants to indicate whether the duration of the stimulus was closer to the *short* standard or the *long* standard, by pressing the left or the right arrow key, respectively. In the practice block, participants received feedback after their responses, that is: either "The presented interval was close to the short standard" or, respectively, "... close to the long standard", depending on whether the interval was shorter or longer than the average of the short and long standards. When the test interval coincided with the mean of the short and long standards (1000 ms in Experiment 2), a random feedback (50% "... close to the short standard" and 50% "... close to the long standard") was given. In the formal experiment, participants received no feedback regarding their responses. After a blank interval of 900 to 1100 ms, the next trial began.

Experiment 1. For better comparison, the sets of intervals introduced in Experiment 1 were similar to the set used in the (Penney et al., 2014) study, in which logarithmic spacing of durations between the short and long standards was applied. As depicted in Figure **1a**, the seven durations used in the positively skewed (PS) session were 400, 504, 636, 800, 1008, 1270, and 1600 ms, and in the negative skewed (NS) session 400, 730, 992, 1200, 1366, 1496, and 1600 ms. Each duration was repeated 48 times in the session (i.e., the durations were distributed

uniformly), and the seven durations were presented randomly intermixed within each session. Each participant completed two sessions, each consisting of six blocks of 56 trials. The order of sessions was counterbalanced (as well as possible) across participants.

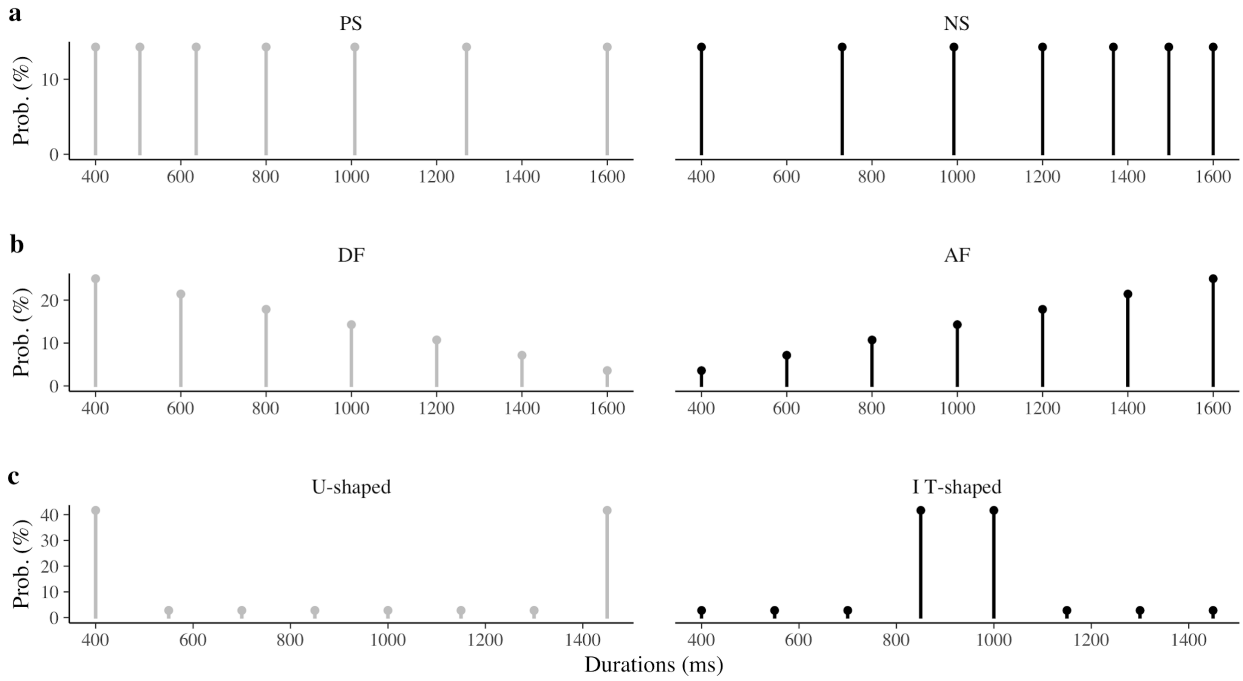

Figure 1 *Sample distributions used in Experiments 1, 2, and 3.* **(a)** *Two spacing conditions used in Experiment 1: in the positively skewed session (PS), the intervals are spaced logarithmically between 400 and 1600 ms, with a mean of 888 ms and a standard deviation (SD) of 401 ms; in the negatively skewed session (NS), the intervals follow a mirrored logarithmic spacing, with a mean of 1112 ms and a SD of 401 ms.* **(b)** *Two sample-frequency conditions for the seven equally spaced intervals (400, 600, 800, 1000, 1200, 1400, 1600 ms) used in Experiment 2: the descending frequency (DF) session has an (arithmetic) mean of 800 ms and a SD of 347 ms; the ascending frequency (AF) session has the mean of 1200 ms and a SD of 347 ms.* **(c)** *Two types of sample-frequency conditions for the eight equally spaced intervals (400, 550, 700, 850, 1000, 1150, 1300, 1450 ms) used in Experiment 3: the U-shaped session has a mean of 925 ms and a SD of 491 ms; the inverted T-shaped session has a mean of 925 ms and a SD of 175 ms.*

Experiment 2. Two types of sample-frequency distributions were tested in separate sessions: a descending sample frequency (DF) and an ascending sample frequency (AF), as depicted in Figure **1b**. In the AF session, the sample frequencies were (1/28, 2/28, 3/28, 4/28,

5/28, 6/28, 7/28) for the intervals (400, 600, 800, 1000, 1200, 1400, 1600 ms); the DF session, the sample frequencies were reversed. There were six blocks of 56 trials in each session.

Experiment 3. Eight intervals between 400 and 1450 ms (400, 550, 700, 850, 1000, 1150, 1300, and 1450 ms) were used in Experiment 3. Two types of sample frequencies were implemented: a U-shaped distribution and an inverted T-shaped distribution (see Figure **1c**). In the U-shaped sampling session, the presentation frequencies of durations (400, 550, 700, 850, 1000, 1150, 1300, and 1450 ms) were (30/72, 2/72, 2/72, 2/72, 2/72, 2/72, 2/72, 30/72), respectively; in the inverted T-shaped sampling session, the frequencies were (2/72, 2/72, 2/72, 30/72, 30/72, 2/72, 2/72, 2/72) for the same durations. The two types of sample distributions have the same arithmetic mean, but they differ in their variability. There were four blocks of 72 trials in each session. The order of the test intervals was randomized within each session, and the order of the sessions was counterbalanced (as well as possible) across participants.

Statistical Analysis

R package Quickpsy (Linares & López-Moliner, 2016) was used to fit psychometric functions and calculate the points of subjective equality (PSE, here the bisection point) and the just noticeable differences (JNDs). We used the cumulative Gaussian function as the psychometric function and the standard deviation of the estimated function as the JND (i.e., the difference between the thresholds at 50% and 75%). All statistical tests were conducted using repeated-measures ANOVAs – with additional Bayes-Factor analyses to comply with the more stringent criteria required for acceptance of the null hypothesis (Kass & Raftery, 1995; Rouder et al., 2009).

RESULTS

Spacing effect in distributions with different ensemble means (Experiment 1)

The psychometric functions, depicting the relation between the proportion of 'long' responses and the test durations, are illustrated in Figure **2a.** By visual inspection, participants made more 'long' responses in the PS session. This was confirmed by an analysis of the PSEs: as depicted in Figure **2b**, the mean PSEs (±standard error, SE) for the PS and NS sessions were 842 (±38) ms and 934 (±40) ms, respectively. A repeated-measures ANOVA revealed this effect of the spacing condition to be significant, $F(1, 14) = 21.141$, $p < .001$, $\eta_g = .0943$,

BF = 47.4. Thus, Experiment 1, which used the similar stimulus settings and procedure as Penney et al. (2014), replicated their spacing effect, as expected (Brown et al., 2005; Penney et al., 2014).

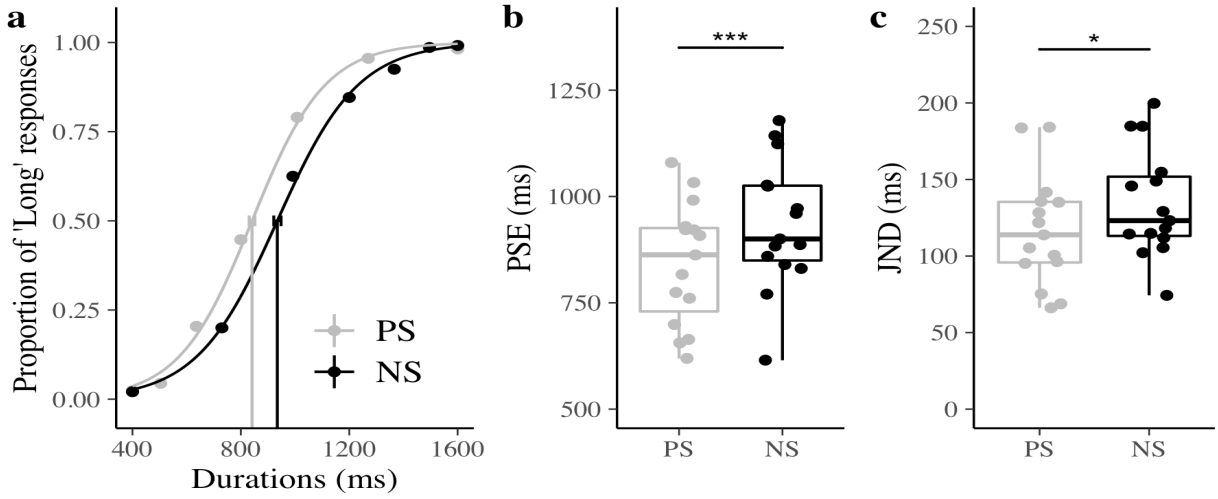

Figure 2 *Results of Experiment 1.* **a.** *Bisection functions (proportions of 'long' responses plotted against the comparison durations, and fitted psychometric curves) averaged across 15 participants for the two, positively (PS) and negatively skewed (NS), stimulus-spacing conditions.* **b.** *Boxplots of PSE of the duration judgments for the PS and NS sessions (*** p<.001). The dots depict individual PSEs estimated from individual participants. Lower and upper tips of the vertical lines correspond to minimum and maximum values, the box represents the interquartile range (between 25 and 75%) and the horizontal line represents the median.* **c.** *Boxplots of JND of the duration judgments for the PS and NS sessions (* p<.05). The dots depict individual JNDs of individual participants.*

As shown in Figure **2c**, the mean JNDs (±SE) were 116.8 (±9.7) ms for the PS and 134.1 (±9.4) ms for the NS distribution. A repeated-measures ANOVA revealed the difference to be significant, $F(1, 14) = 5.755, p = .031, \eta_g = 0.059, BF = 2.10$. The larger JND in the NS versus the PS condition is likely attributable to Weber scaling in the perceived durations. Subjective time is known to roughly follow Weber's law (i.e., it exhibits the scalar property), with longer durations showing larger variability of the subjective estimates than shorter durations (Gibbon, 1977; Wearden & Lejeune, 2008). As there were more longer durations in the NS session than in the PS session (see Figure 1a), the uncertainty in the NS session was likely higher, giving rise to the increased JNDs relative to the PS session.

Shifts in BPs are associated with ensemble mean but not with spacing information (Experiment 2)

In Experiment 2, distributions of descending frequency (DF) and ascending frequency (AF) were generated for the same set of durations with equal spacing (step of 200 ms). The arithmetic mean was 800 ms in the DF condition, versus 1200 ms in the AF condition. Figure **3a** depicts the average psychometric curves for the two (DF and AF) conditions, showing a marked shift in the location of the bisection points: the mean PSE (±SE) was 997±45 ms for the AF condition and 821±37 ms for the DF condition (see Figure **3b**). This difference was significant (repeated-measures ANOVA on the PSEs: $F(1,14) = 26.83$ $p < .001$, $\eta_g = .263$, $BF = 222.0$). That is, compared to the AF condition, the DF condition was associated with an increased probability of 'long' responses, the latter consisting of relatively more *short* intervals and thus having a relatively shorter arithmetic mean of the stimulus set (which serves as a reference for the bisection). According to EDA, temporal bisection essentially involves a comparison of a given duration to the estimate of the ensemble mean. Thus, compared to the AF set, the relatively *shorter* ensemble mean in the DF set would lead to more 'long' responses.

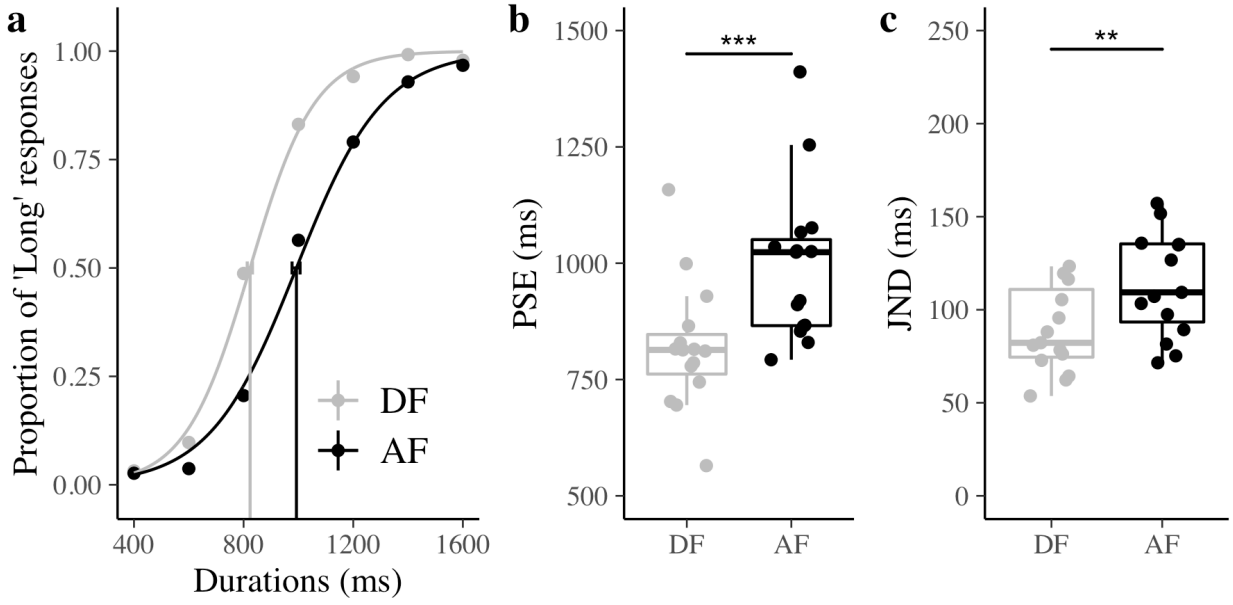

Figure 3 *Results of Experiment 2.* **a.** *Bisection functions (proportions of 'long' responses plotted against the comparison durations, and fitted psychometric curves) averaged across 15 participants for the two, descending-frequency (DF) and ascending-frequency (AF), duration-distributions conditions.* **b.** *Boxplots of PSE of the duration judgments for the DF and AF conditions (*** $p < .001$). The dots depict individual PSEs estimated from individual*

participants. **c.** *Boxplot of JND of the duration judgments for DF and AF conditions (** p < .01). The dots depict individual JNDs of individual participants.*

As shown in Figure **3c**, the mean JNDs were 108 (±22.7) ms for the DF condition and 138.8 (±28) ms for the AF condition, with the difference being significant (repeated-measures ANOVA: $F(1,14) = 15.38$, $p < .01$, $\eta_g = .027$, $BF = 17.82$). Similar to Experiment 1, the frequency distribution with more *long* durations (here the AF session) had a larger JND than the distribution with more *short* durations. Again, this was likely due to unequal Weber scaling in the two sets: the uncertainty induced by the *long* durations was more prominent in the AF (as compared to the DF) session, further pointing to the influence of ensemble statistics in duration comparison.

Sensitivity of temporal judgment is driven by ensemble variance (Experiment 3)

Experiment 3 was designed to examine whether performance on the temporal-bisection task would be affected by the variance of the contextual stimulus set. Accordingly, the distributions of the stimulus set (both with equal spacing of the interval durations) had the same mean, but they differed in their variance. Figure 4a depicts the psychometric functions averaged across 15 participants for each stimulus set condition: U-shaped and inverted T-shaped frequency distribution. Consistent with the prediction of EDA, the two psychometric curves cross each other at the 50% threshold, while having different slopes. As depicted in Figure 4b, the mean PSEs (±SE) were comparable: 863 (±31.5) ms and 864 (±26.5) ms for the U-shaped and inverted T-shaped distributions, respectively (repeated-measures ANOVA: $F(1,14) = 0.001$, $p = 0.97$, $\eta_g = .000026$, $BF = 0.331$, with the BF value providing strong evidence in favor of the null hypothesis). The absence of a difference in the PSEs between the U-shaped and inverted T-shaped distributions in this experiment, combined with findings from Experiments 1 and 2, suggests that the shift in PSEs was driven mainly by the ensemble mean, not the ensemble variance.

In contrast, as can be seen from Figure 4c, there was a marked difference in the JNDs between the two conditions, with mean JNDs of 111.75 (±11) ms for the U-shaped and 86.37 (±8) ms for the inverted T-shaped distribution. A repeated-measures ANOVA revealed the difference to be significant: $F(1, 14) = 9.171$, $p < .01$, $\eta_g = .117$, $BF = 5.96$.

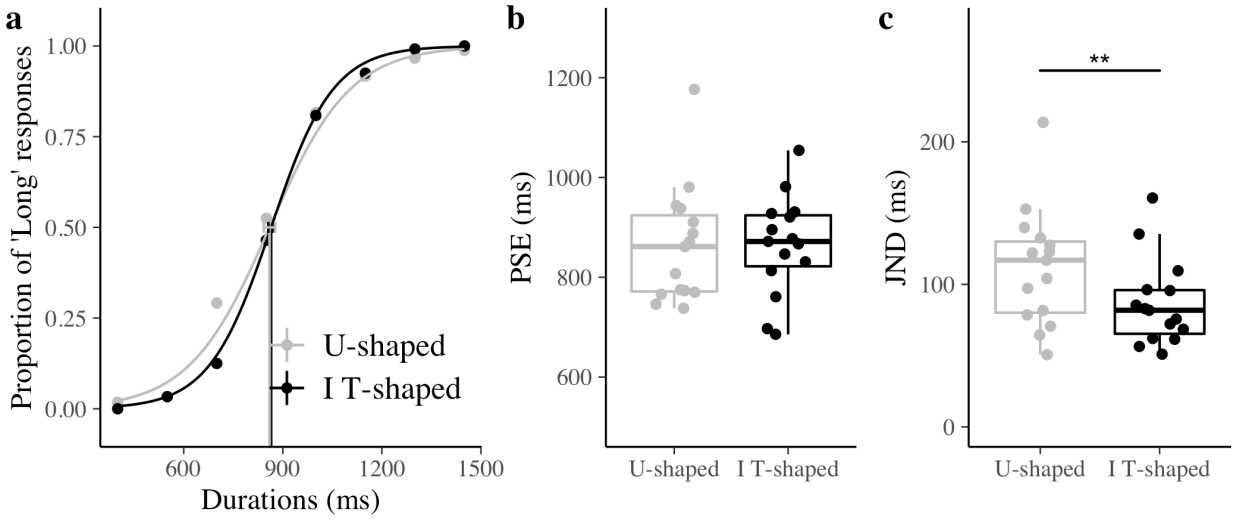

Figure 4 *Results of Experiment 3.* **a.** *Bisection functions (proportions of 'long' responses plotted against the comparison durations, and fitted psychometric curves) averaged across 15 participants for the two, U-shaped and inverted T-shaped, frequency-distribution conditions.* **b.** *Boxplots of PSE of the duration judgements for the U-shaped and inverted T-shaped frequency distributions.* **c.** Boxplots of *JND of the duration judgments for the U-shaped and inverted T-shaped frequency distributions (**p < .01).*

Following the prediction of EDA, we expected a significantly steeper slope (i.e., a smaller JND) with the inverted T-shaped, as compared to the U-shaped, frequency distribution, given that the variance was smaller in the former than in the latter. In addition to the variance of the set, the Weber scaling of the longest duration may also contribute to the large variance in the U-shape condition, similar to the differences in JNDs that we observed in Experiments 1 and 2.

MODELING

To compare different models of the temporal-bisection task, we applied Bayesian hierarchical modeling (Lee & Wagenmakers, 2014) to our behavioral data, with correspondent assumptions about how the task is performed. The framework of the hierarchical model is illustrated in Figure 5. For a given duration $X_j^{(i)}$ in condition i, we assume the bisection response follows the binomial distribution for the probability of the 'long' response $p_j^{(i)}$. The probability $p_j^{(i)}$ is determined by the ratio comparison of probe duration $X_j^{(i)}$ and the bisection point ${X_{BP}}^{(i)}$ according to the following psychometric relation:

$$log\,\frac{p_j^{(i)}}{1-p_j^{(i)}} = \alpha^{(i)} + \beta^{(i)}\left(\frac{x_j^{(i)}}{x_{BP}^{(i)}} - 1\right), \quad \text{(Eq.1)}$$

where $\alpha^{(i)}$ and $\beta^{(i)}$ are two psychometric parameters. The left side of the equation is the decision variable expressed in the log-likelihood of the two alternative ("Long" vs. "Short") responses, while the right side assumes the comparison is based on the ratio $\frac{x_j}{x_{BP}}$. The ratio comparison on a linear scale is equivalent to the subtraction comparison on the internal log-scaled representation, which conforms to Weber scaling (i.e., the scalar property). The assumption of a logarithmic scale for internal duration representation is common in theories and models of duration judgments (Petzschner et al., 2015; Ren et al., 2020; Roach et al., 2017; Wearden, 1991). The equation implies that the decision variable is a linear function of the ratio comparison. The coefficient $\beta^{(i)}$ reflects both the sensitivity of the ratio comparison and a potential central-tendency bias (which makes the slope of the psychometric function shallower) for individual participants, in a given condition i. In Bayesian hierarchical models, the parameters $\alpha^{(i)}$ and $\beta^{(i)}$ are assumed to follow Gaussian distributions with the respective hyper-parameters $(\mu_\alpha, \sigma_\alpha)$ and $(\mu_\beta, \sigma_\beta)$. In addition, EDA assumes that the distribution of decision sensitivity $\beta^{(i)}$ is Gaussian with a mean proportional to the reciprocal of the relative spread of the test durations, that is: $N\left(k \cdot \frac{\mu}{\sigma_X}, \sigma_\beta\right)$, where μ and σ_X are the mean and, respectively, the standard deviation of the test durations, and k is a scaling factor. In other words, narrower sample distributions would enhance the sensitivity of the bisection task. It should be noted that, as implemented, our EDA model uses the veridical spread of the sampled duration σ_X, disregarding the subjective Weber scaling in the perceived ensemble variability σ'_X; but the framework can be easily extended to the subjective scale.

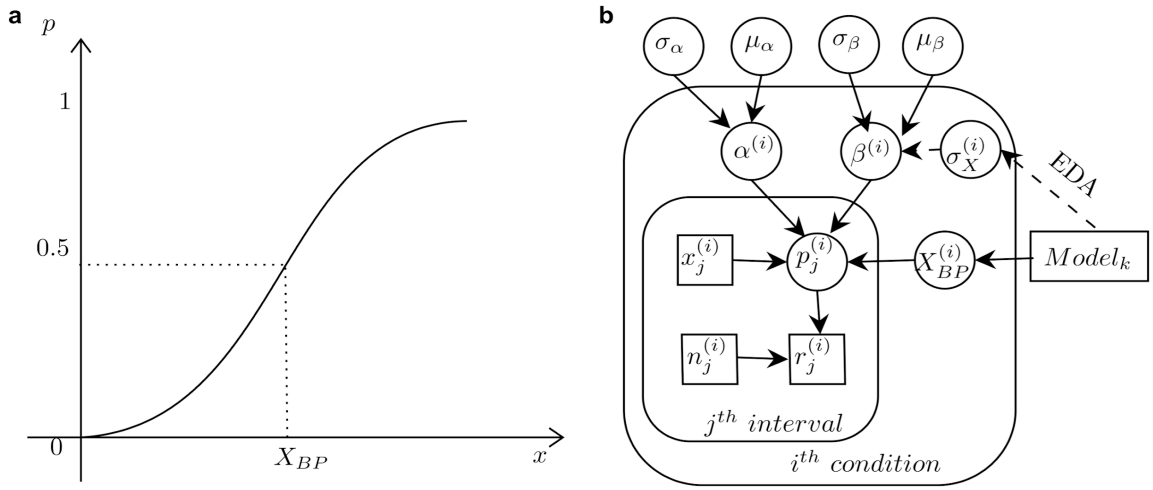

Figure 5 *The framework of the hierarchical model.* **a.** *A psychometric function with bisection point* X_{BP}. **b.** *Schematic illustration of the hierarchical model of temporal duration judgments (see text for details).*

The main difference among the various models, ranging from the simple bisection model to EDA, is how the critical reference $X_{BP}^{(i)}$ (reference *M* in Wearden & Ferrara, (1995)) is used in the bisection task (see Table 1). The simple bisection model assumes the comparison is made either between the ratios $\frac{X_S}{X}$ and $\frac{X}{X_L}$ or between probe X and the arithmetic-mean duration $\frac{X_L+X_S}{2}$; that is, essentially the comparison is made to either the geometric mean (GM) or the arithmetic mean (AM) of the standard durations. From their meta-analysis of 148 experiments, Kopec and Brody (2010) concluded that it remains controversial whether the bisection point is close to the GM or AM. The bisection point is influenced by a number of factors, including the short-long spread (i.e., the Long/Short ratio) and the probe context. The spacing account, for example, assumes the comparison reference is the arithmetic mean of the whole probe durations, each of which is equally frequent by design (Penney & Cheng, 2018). The TRFT account (Brown et al., 2005; Penney et al., 2014), on the other hand, assumes that, rather than being veridical, subjective duration is an average of two components, namely: the relative position of a sampled duration in-between the short and long standards and the ordinal position of the duration within the sample durations. The proportion of 'long' responses to a sampled duration is then based on a comparison between its calculated relative position and the short and long standards. Accordingly, the estimated temporal bisection point $X_{BP}^{(i)}$ lies roughly between the mean of the sample distribution and the arithmetic mean of the short and long standards. We simulate this using the ensemble-mean model, that is: $X_{BP}^{(i)}$ is the mean (either the GM or the AM) of the sampled distribution. Given that the mean of the sample

distribution is not a fixed parameter known to participants, but rather updated dynamically over the course of the trials, we further propose that the bisection reference ($X_{BP}^{(i)}$) is also a variable that fluctuates from trial to trial, while being centered around the geometric or the arithmetic mean of the sampled distribution. We refer to this as the two-stage ensemble-mean model (see Table 1). The EDA model goes one step further, by incorporating the variance of the distribution into bisection decisions, that is: the slope of the psychometric function (β) is inversely related to the relative spread of the ensemble distribution ($\frac{\sigma_X}{\mu}$, see Table 1). EDA predicts that increasing the variability of the test durations would decrease the bisection sensitivity when the mean of the test range is fixed. Moreover, given that EDA (see Eq. 1) conforms to Weber scaling (Jozefowiez et al., 2014; Kopec & Brody, 2010; Wearden & Lejeune, 2008), β would remain unchanged when the ratio of the mean and standard deviation of the test durations is kept constant.

Table 1. *Models and model assumptions about the comparison reference* $\boldsymbol{X_{BP}}$ *and decision sensitivity*

Models	**Reference** $\boldsymbol{X_{BP}}$	**Decision sensitivity** $\boldsymbol{\beta}$
Simple Bisection model	AM: $X_{BP} = (X_s + X_L)/2$ GM: $X_{BP} = \sqrt{X_s X_L}$	N.a.
Spacing model	AM: $X_{BP} = \frac{1}{n}\sum X_j$ GM: $X_{BP} = \sqrt[n]{\Pi X_j}$	N.a.
Ensemble-mean model	AM: $X_{BP} = \frac{1}{\Sigma f_j}\sum f_j X_j$ GM: $X_{BP} = \sqrt[\Sigma f_j]{\Pi \left(X_j\right)^{f_j}}$	constant*
Two-stage ensemble-mean model	AM: $X_{BP} \sim N(\mu, \sigma),\ \mu = \frac{1}{\Sigma f_j}\sum f_j X_j$ GM: $X_{BP} \sim N(\mu, \sigma),\ \mu = \sqrt[\Sigma f_j]{\Pi \left(X_j\right)^{f_j}}$	constant
Ensemble-distribution account (EDA)	AM: $X_{BP} \sim N(\mu, \sigma),\ \mu = \frac{1}{\Sigma f_j}\sum f_j X_j$	

	GM: $X_{BP} \sim N(\mu, \sigma),\ \mu = \sqrt[\Sigma f_j]{\Pi \left(X_j\right)^{f_j}}$	$\beta \sim N\left(k \cdot \frac{\mu}{\sigma_X}, \sigma_\beta\right)$

Note. * TRFT model (Brown et al., 2005; Penney et al., 2014) used the percentile to approximate the subjective magnitude, and then applied ratio comparison between the perceived magnitude to the short and long standards, which implicitly incorporates some degree of variance of the distribution in the bisection decision.

Based on those assumptions, we fitted the above five models to our data and estimated the corresponding psychometric functions, PSEs, and JNDs for each participant. We applied the AM and GM as $X_{BP}^{(i)}$ in all models. The simple bisection model and the spacing model lowered the PSE with the GM relative to the AM, while both the two-stage ensemble-mean model and the EDA model yielded very close predictions with both the GM and the AM (the relative mean difference between the two predictions of the PSEs $\frac{|PSE_{GM}-PSE_{AM}|}{PSE_{obs}}$was less than 2%), owing to the trial-by-trial variation rendering the small difference between the GM and AM of little effect in the hierarchical model. Overall, EDA outperformed all other models (see below), whether incorporating the AM or GM variant. Given this, here we consider only the models with the AM. To visually compare the various models with regard to their respective predictions of the PSEs and JNDs, we plotted their mean predictions with individual estimates in Figure 6. The spacing model and the ensemble-mean model made the same prediction for the spacing manipulation in Experiment 1 (PS vs. NS), given that the sampled durations were weighted equally (shown overlapped in the left panel of Figure 6). Both, however, underestimated the PSE and JND in the PS condition (more short durations), and overestimated the PSE and JND in the NS condition (more long durations). This suggests that the bisection point (BP) was assimilated to the mean of the skewed distribution, though only partially. By contrast, the two-stage ensemble-mean model and the EDA model provided a very close prediction to the observed PSEs in Experiment 1. In addition, the EDA model (but not the two-stage ensemble-mean model) predicted the mean JNDs. The main difference between the two-stage ensemble-mean (and EDA) model(s) and the ensemble-mean model is that the former assumes random trial-to-trial fluctuations of the ensemble mean, thus potentially incorporating a partial range effect in the decision (i.e., assimilation to the arithmetic mean). The ensemble means were more extreme in the DF and AF conditions in Experiment 2 as compared to Experiment 1, deviating greatly from the observed PSEs (see the middle panel in Figure 6). This suggests that the mean of the distribution was not the sole factor determining the bisection

judgments. Again, by having the reference (i.e., bisection point) vary across trials, both the two-stage ensemble-mean model and the EDA model well predict the observed PSEs. Incorporating the variance of the distribution in the bisection decision, the EDA model was able to predict the JNDs across all six different sets, indicating that the second moment of the ensemble statistics (i.e., variance) does influence performance of the bisection task.

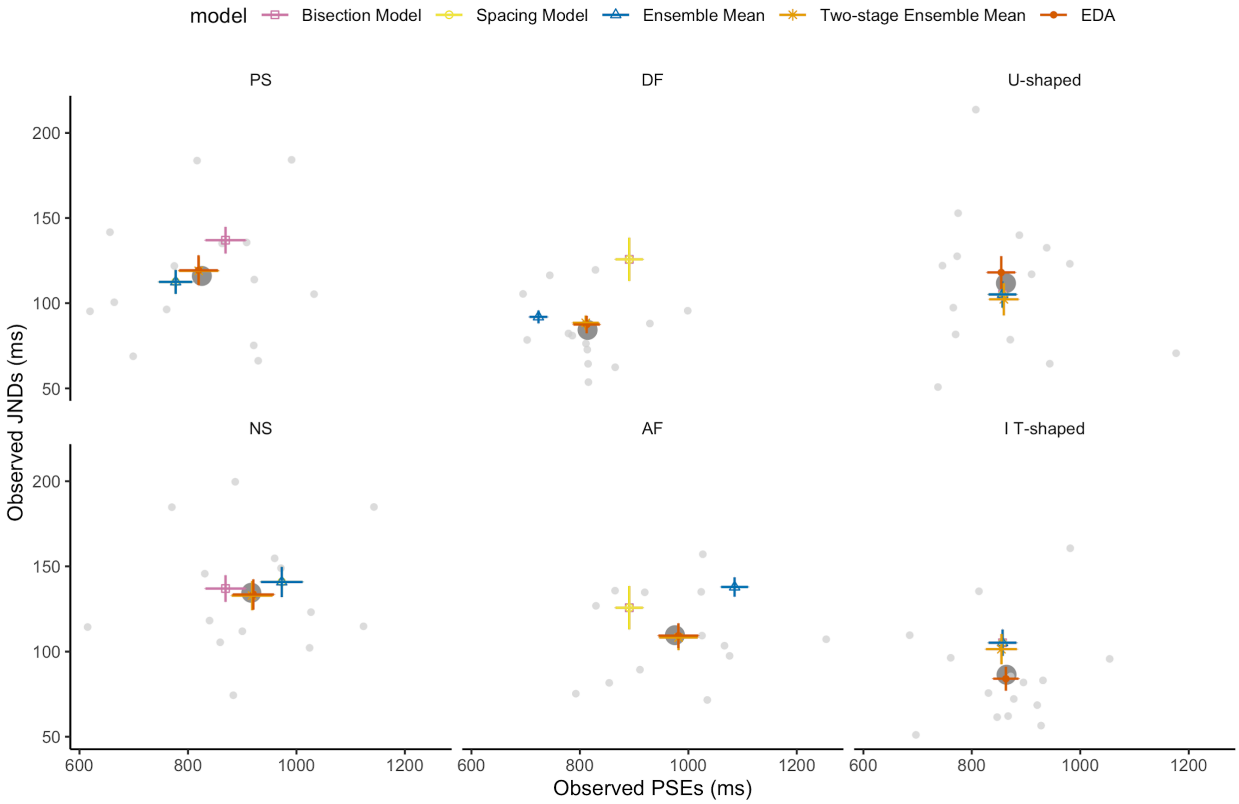

Figure 6 *Observed PSEs and JNDs and mean predictions from the models compared. PSE-JND pairs from individuals are plotted in gray dots, and their means are shown in large dark dots. The mean predictions of PSE-JND pairs and their PSEs from the five models are marked with colored error bars. Note that the predictions of the spacing model and the ensemble-mean model were the same (overlapped in the figure) for the PS and NS sets (the left panel), given that the sampled durations were equally weighted. Due to the equal spacing in the DF, AF, U-shaped, and IT-shaped sets, the bisection and the spacing models made the same predictions (overlapped in the figure). The predictions from the EDA model came the closest to the mean of the observed PSE-JND pairs across all six sets.*

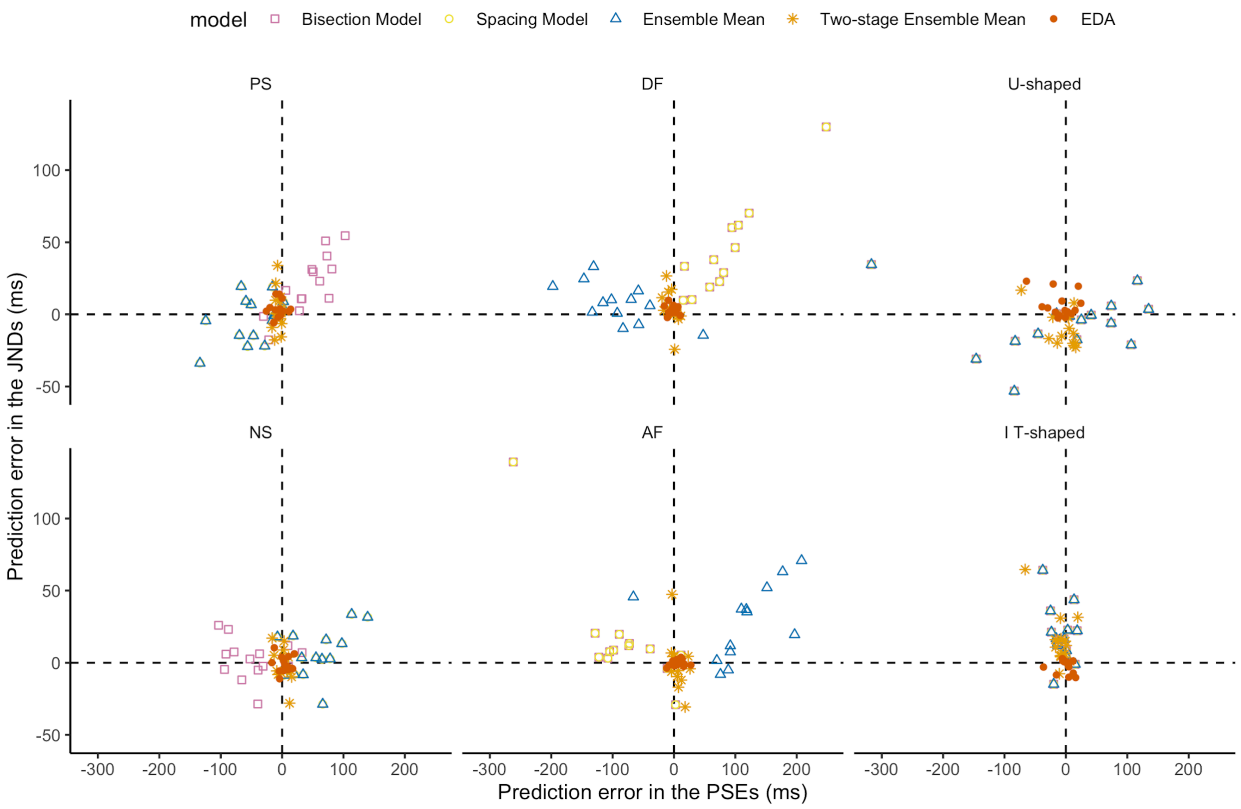

Figure 7 *Scatterplot of prediction errors in the PSEs and JNDs derived from the five models for individual observers across the six duration sets presented in Experiments 1–3 (see Figure 1). Perfect predictions are located at the center (0,0). The average prediction errors of the six conditions of the EDA model were the smallest.*

To obtain a better picture of the model predictions at the level of individual observers, we plotted the prediction errors (predicted vs. observed values) in PSEs versus JNDs in Figure 7. Each point represents the errors in PSE and JND estimation, per individual participant, derived from a specific model of the five models compared. As can be seen from Figure 7, the scatter points of the EDA model are centered nearer to the origin of the XY coordinates compared to the points of the other models, indicating that the PSEs and JNDs predicted by EDA come closest to the mean of the observed values, across all three experiments. This observation is supported by measures of the Euclidean distances of the model predictions from the observed PSE-JND pairs: the mean distances were 68.4, 68.1, 75.2, 16.7, 11.9 ms for the simple bisection, spacing, ensemble mean, two-stage, and EDA models, respectively. Formal corroboration of the superiority of EDA is provided by 'goodness-of-fit' measures of the predicted psychometric curves using the Watanabe-Akaike information criterion (WAIC). The WAIC is a measure of the quality of a hierarchical model, which takes into account the

goodness-of-fit, as measured by the likelihood, while also penalizing models with more free parameters (Vehtari et al., 2017). Lower WAIC values indicate better model performance. The mean WAICs for the predicted 'long' responses across all test durations were 77.2, 92.7, 104.6, 76.7, and 72.1 for the simple bisection, spacing, ensemble mean, two-stage, and EDA models, respectively. That is, across all conditions, the EDA model provides the best fit for both the PSEs and the JNDs, evidenced by the fact that the WAIC for the EDA model was the smallest in all cases.

GENERAL DISCUSSION

In the present study, we tested six different duration sets, with set properties varying in stimulus spacing, set mean, and set variance in three experiments, to investigate whether temporal-bisection judgments would be best explained by our ensemble-distribution account (EDA). Experiment 1 demonstrated that the skewness of unequal spacing significantly shifted the bisection point, confirming previous findings (Brown et al., 2005; Penney et al., 2014; Wearden & Ferrara, 1995). Given that the short and long standards were identical in the two (positively skewed, PS, and negatively skewed, NS) sets, the finding of differential PSEs between the PS and NS conditions argues against the simple bisection account, irrespective of whether it assumes the arithmetic or the geometric mean of the standards as reference. Experiment 2 further demonstrated that the frequencies of the sampled durations greatly impacted the bisection point, even when the probe durations were equally spaced in the ascending-frequency (AF) and descending-frequency (DF) sets. The spacing account fails to predict this effect. Experiment 3 kept the spacing of the durations and the mean of the sets the same for the U-shaped and inverted T-shaped sets, but varied their variances (larger variance for the U-shaped set). The results revealed the variance of the set to influence the sensitivity of temporal bisection, reflected in differential JNDs (despite the equivalent set means). While previous accounts failed to predict the PSEs and JNDs in one set or another, the EDA model successfully accounted for the shifts of the PSEs and JNDs in all six sets examined. The results of model-fitting analyses also showed the EDA model to provide the best account of the data.

Temporal bisection and related accounts

The temporal-bisection task was first developed in research on animal timing (Gibbon, 1977; Gibbon et al., 1984) and later adapted to studies on human timing (Allan & Gibbon, 1991; Wearden, 1991). The focus of the initial studies on temporal bisection was on how humans and

other animals make interval comparisons (rather than on context-dependent manipulations of the bisection point). Accordingly, the early work implicitly assumed that the interval comparison in temporal bisection depends only on the probe stimulus (t) and the short (S) and long (L) standards (e.g., Allan & Gibbon, 1991). Later studies, however, revealed that temporal bisection is sensitive to the probe context. Logarithmic versus linear spacing of the probe durations often resulted in different bisection points, even though the short and long standards remained the same (Brown et al., 2005; Penney et al., 2014). This raised the question as to the temporal reference to which observers actually compared a given probe duration. Using various spacing manipulations to probe for shifts of bisection points, Wearden and Ferrara (1995) suggested that observers likely compare the probe duration t to a reference M, rather than with S and L, where M lies somewhere in-between the geometric and arithmetic mean of the S and L. While this M-reference proposal could qualitatively explain the shifts of the bisection points, it falls short of quantitatively predicting the influence of the sampled distribution. Subsequently, Brown and colleagues (2005) developed the TRFT model based on the 'range frequency theory', arguing that the subjective measure of a given probe duration (t') is influenced by its temporal position within the sample range and its percentile in the distribution. The TRFT model still considers temporal-bisection judgments to involve a ratio comparison between $\frac{S}{t'}$ and $\frac{t'}{L}$, where S and L are veridical. It should be noted that the TRFT model assumes that observers can retrieve the ordinal (rank) position of the probe duration (i.e., percentile), which logically requires a representation of the full ordered sequence of the durations. Building up such a representation would be very memory-intensive and thus quite unlikely with a large set of durations. By contrast, Wearden and Ferrara's M-reference model (1995) only requires an estimation of M across all trials, which is computationally less intensive. Here we adopted the M-reference approach to compare various models, assuming that temporal-bisection judgments are made by comparing the probe t to a bisection point X_{BP}, with the X_{BP} being context-sensitive. The model comparison revealed that the best fit is provided by the EDA account, which assumes that the bisection point is derived from the ensemble mean with trial-to-trial random variation. Given that the ensemble statistics are not known prior to the experiment, but rather updated dynamically from trial to trial, EDA, as well as the two-stage ensemble-mean model, can capture the dynamic adaptation of the bisection point, making them outperform the other models.

The assumption that the bisection point is sensitive to context does not rule out that subjective measures of individual probe durations themselves are modulated by the context. In fact, perceived durations are known to be subject to the central-tendency effect (Lejeune & Wearden, 2009; Shi et al., 2013). However, the central-tendency effect only assimilates individual durations toward the ensemble mean, which could degrade the sensitivity of the bisection task (given that the subjective distance to the mean is shortened), while leaving the bisection point unaffected. In other words, the bisection point is insensitive to the central-tendency effect. Given this and for the purpose of simplicity, we did not explicitly incorporate the central-tendency effect in our modeling (though it was an implicit factor influencing the general free parameter β; see the modeling section and Eq.1).

One strong prediction deriving from EDA, which sets it apart from the other accounts, is that the variability of the sampled distribution influences the sensitivity of bisection judgments, as corroborated empirically in Experiment 3. The duration set with the inverted T-shaped distribution had a lower variability compared to the U-shaped set, which led to the psychometric curve becoming steeper. The reason is that the ensemble representation of the inverted T-shaped distribution would be narrower than that of the U-shaped distribution, given that ensemble perception incorporates the external statistics. As a result, any probe durations that deviate from the ensemble mean would be more likely judged as short or long in the inverted T-shaped set than in the U-shaped set.

Ensemble perception for temporal sequences

It should be noted that, while previous research on ensemble perception has primarily focused on summary statistics of the immediately, or simultaneously, available information, such as the average size of a set of objects (Alvarez, 2011; Whitney & Yamanashi Leib, 2018), the ensemble perception we refer to here concerns the statistical summary representation that is acquired through trial history. Real-world sensory inputs relevant to our behavioral goals do not always occur all at once, and we are often exposed to changing characteristics of goal-relevant objects. For instance, linguistic research has shown that both infants and adults make use of statistical computations for deciding which series of sounds establishes a word within a nonstop flow of spoken sounds (Newport & Aslin, 2000; Saffran, 2001). In particular, people are able to track the regularities in a series, or 'group', of sound elements, which render the predictiveness of one sound element onto another. For language acquisition and the learning of new languages, such probability-based computations provide important rhythmic temporal

information (i.e., acoustic features such as duration and frequency of speech elements and summary statistics of sequences of speech sounds and silent intervals) for parsing the sequential sounds in speech. And more recently, Chen and colleagues (2018) showed that humans automatically derive the mean interval from a sequence of auditory beeps, which then cross-modally influences, or 'patterns', visual apparent motion. The present findings add to this evidence by demonstrating that temporal ensemble perception developed through trial history provides the 'reference' for duration comparison.

One might ask why we need ensemble perception in the first place, when temporal tasks, such as bisection, can be accomplished with greater precision without the influence of ensemble statistics. To answer this question, we need to consider the fundamental roles of ensemble perception. The environment we live in does not comprise random objects and events, but rather has structure and regularity (Cohen et al., 2016). We are continually confronted with abundant information which is beyond our processing capacity. Accordingly, evolutionary pressures pushed us to utilize regularity – that is, ensemble-statistical – information to overcome the capacity limit (Ariely, 2001; Whitney & Yamanashi Leib, 2018). As a result, deriving such ensemble statistics became 'intuitive' and automatic. In many situations, using intuitive ensemble perception can help us recover unattended events, spot outliers, or make predictions. For example, when listening to background music while working, one can rapidly spot a change in rhythm even if one's focus is not on music. Yet, in some instances, implicitly using ensemble statistics would give rise to unintended biases, such as the PSE/JND shifts we report here. It should be noted that in the bisection task, observers are actually explicitly told to compare the probe duration to the Short (S) and Long (L) standard. The reason why participants use the ensemble distribution as the reference, rather than S and L, is likely owing to the fact that S and L contribute to the ensemble distribution in the same way as the other durations (Wearden & Ferrara, 1995).

The role of variability in ensemble perception

Previous research on ensemble perception has largely focused on demonstrating humans' ability to accurately estimate mean values from an array of objects or sensory features (Whitney & Yamanashi Leib, 2018). However, the variability of the sampled stimuli provides useful information about the range, stimulus spacing, and exceptional cases in the data set – hence the variance statistic is a key component in ensemble perception. Although processing mean information eases the limitations of our perceptual experience by summarizing multiple

items or features to an exemplar component (Alvarez, 2011), detecting similarities or, respectively, deviations among items is not solely based on the mean information, but may also benefit from complementary measures of stimulus range and variance (Haberman et al., 2015; Michael et al., 2014; Solomon, 2010). Thus, for example, ensemble variance has been suggested to be useful for identifying potential outliers or deviations in a crowd (Whitney & Yamanashi Leib, 2018). The results of the present study show that variance information is also exploited in temporal judgments: in temporal bisection, the variance of the sample distribution provides useful information for discerning the location of a probe duration relative to the ensemble mean, thus enhancing temporal sensitivity (as evidenced by the steepness of the psychometric curve).

CONCLUSION

The ensemble context is an important determinant in time perception. The present paper reported three experiments in which we manipulated the distribution of auditory duration sets to determine factors that influence temporal-bisection performance. The results revealed the mean and variance of the stimulus set to be critical factors, producing shifts of the bisection point and, respectively, changes of the slope of the bisection curves. These findings demonstrate that the human timing mechanism involves an ensemble averaging process which works similarly to other perceptual properties in the visual and auditory domains. Moreover, we proposed an *ensemble-distribution accoun*t that explains in which way subjective judgments of time intervals vary according to the distribution summary statistics of the set mean and variance values.

OPEN PRACTICES

The data and codes for all experiments are available at https://github.com/msenselab/sets_in_bisection.

ACKNOWLEDGEMENTS

This study was supported by German Science Foundation (DFG) research grants SH 166/ 3-2 to Z.S. and DAAD scholarship 57440921 to C.B.

REFERENCES

Acerbi, L., Wolpert, D. M., & Vijayakumar, S. (2012). Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. *PLoS Computational Biology*, *8*(11), e1002771.

Allan, L. G., & Gibbon, J. (1991). Human bisection at the geometric mean. *Learning and Motivation*, *22*(1), 39–58.

Allman, M. J., & Meck, W. H. (2012). Pathophysiological distortions in time perception and timed performance. *Brain: A Journal of Neurology*, *135*(Pt 3), 656–677.

Allman, M. J., Teki, S., Griffiths, T. D., & Meck, W. H. (2014). Properties of the internal clock: first- and second-order principles of subjective time. *Annual Review of Psychology*, *65*, 743–771.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131.

Ariely, D. (2001). Seeing sets: Representation by Statistical Properties. *Psychological Science*, *12*(2), 157–162.

Brown, G. D. A., McCormack, T., Smith, M., & Stewart, N. (2005). Identification and bisection of temporal durations and tone frequencies: common models for temporal and nontemporal stimuli. *Journal of Experimental Psychology. Human Perception and Performance*, *31*(5), 919–938.

Burr, D., Della Rocca, E., & Morrone, M. C. (2013). Contextual effects in interval-duration judgements in vision, audition and touch. *Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale*, *230*(1), 87–98.

Chen, L., Zhou, X., Müller, H. J., & Shi, Z. (2018). What you see depends on what you hear: Temporal averaging and crossmodal integration. *Journal of Experimental Psychology. General*, *147*(12), 1851–1864.

Cicchini, G. M., Arrighi, R., Cecchetti, L., Giusti, M., & Burr, D. C. (2012). Optimal encoding of interval timing in expert percussionists. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(3), 1056–1060.

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the Bandwidth of Perceptual Experience? *Trends in Cognitive Sciences*, *20*(5), 324–335.

Fründ, I., Wichmann, F. A., & Macke, J. H. (2014). Quantifying the effect of intertrial dependence on perceptual decisions. *Journal of Vision*, *14*(7). https://doi.org/10.1167/14.7.9

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psychological Review*, *84*(3), 279–325.

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

Glasauer, S., & Shi, Z. (2018). 150 years of research on Vierordt's law-Fechner's fault? *bioRxiv*, 450726.

Haberman, J., Lee, P., & Whitney, D. (2015). Mixed emotions: Sensitivity to facial variance in a crowd of faces. *Journal of Vision*, *15*(4), 16.

Haberman, J., & Whitney, D. (2009). Seeing the mean: ensemble coding for sets of faces. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(3), 718–734.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*, *13*(8), 1020–1026.

Jozefowiez, J., Polack, C. W., Machado, A., & Miller, R. R. (2014). Trial frequency effects in

human temporal bisection: implications for theories of timing. *Behavioural Processes*, *101*, 81–88.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception ECVP Abstract Supplement*, *36*, 14.

Kopec, C. D., & Brody, C. D. (2010). Human performance on the temporal bisection task. *Brain and Cognition*, *74*(3), 262–272.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

Lejeune, H., & Wearden, J. H. (2009). Vierordt's The Experimental Study of the Time Sense (1868) and its legacy. *The European Journal of Cognitive Psychology*, *21*(6), 941–960.

Linares, D., & López-Moliner, J. (2016). quickpsy: An R package to fit psychometric functions for multiple groups. *The R Journal, 2016, Vol. 8, Num. 1, P. 122-131*. http://diposit.ub.edu/dspace/handle/2445/116040

Marchant, A. P., Simons, D. J., & de Fockert, J. W. (2013). Ensemble representations: effects of set size and item heterogeneity on average size perception. *Acta Psychologica*, *142*(2), 245–250.

Michael, E., de Gardelle, V., & Summerfield, C. (2014). Priming by the variability of visual information. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(21), 7873–7878.

Newport, E. L., & Aslin, R. N. (2000). Innately constrained learning: Blending old and new approaches to language acquisition. *Proceedings of the 24th Annual Boston University Conference on Language Development*, *1*, 1–21.

Parducci, A. (1963). Range-frequency compromise in judgment. *Psychological Monographs:*

General and Applied, *77*(2), 1–50.

Penney, T. B., Brown, G. D. A., & Wong, J. K. L. (2014). Stimulus spacing effects in duration perception are larger for auditory stimuli: data and a model. *Acta Psychologica*, *147*, 97–104.

Penney, T. B., & Cheng, X. (2018). Duration Bisection: A User's Guide. In *Timing and Time Perception: Procedures, Measures, & Applications* (pp. 98–127). Brill.

Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A Bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences*, 1–9.

Pronin, E. (2013). When the Mind Races: Effects of Thought Speed on Feeling and Action. *Current Directions in Psychological Science*, *22*(4), 283–288.

Raslear, T. G. (1985). Perceptual bias and response bias in temporal bisection. *Perception & Psychophysics*, *38*(3), 261–268.

Raviv, O., Ahissar, M., & Loewenstein, Y. (2012). How recent history affects perception: the normative approach and its heuristic approximation. *PLoS Computational Biology*, *8*(10), e1002731.

Ren, Y., Müller, H. J., & Shi, Z. (2020). Ensemble perception in the time domain: evidence in favor of logarithmic encoding of time intervals. In *Scientific Report* (p. 2020.01.25.919407). https://doi.org/10.1101/2020.01.25.919407

Rhodes, D., & Di Luca, M. (2016). Temporal Regularity of the Environment Drives Time Perception. *PloS One*, *11*(7), e0159842.

Roach, N. W., McGraw, P. V., Whitaker, D. J., & Heron, J. (2017). Generalization of prior information for rapid Bayesian time estimation. *Proceedings of the National Academy of Sciences*, *114*(2), 412–417.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2),

225–237.

Saffran, J. R. (2001). Words in a sea of sounds: the output of infant statistical learning. *Cognition*, *81*(2), 149–169.

Shi, Z., & Burr, D. (2016). Predictive coding of multisensory timing. *Current Opinion in Behavioral Sciences*, *8*, 200–206.

Shi, Z., Church, R. M., & Meck, W. H. (2013). Bayesian optimization of time perception. *Trends in Cognitive Sciences*, *17*(11), 556–564.

Solomon, J. A. (2010). Visual discrimination of orientation statistics in crowded and uncrowded arrays. *Journal of Vision*, *10*(14), 19.

Treisman, M. (1963). Temporal discrimination and the indifference interval: Implications for a model of the“ internal clock.” *Psychological Monographs: General and Applied*, *77*(13), 1.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Vierordt, K. von. (1868). *Der Zeitsinn nach Versuchen [The sense of time according to research]*. H. Laupp.

Watamaniuk, S. N., & Duchon, A. (1992). The human visual system averages speed information. *Vision Research*, *32*(5), 931–941.

Wearden, J. H. (1991). Human performance on an analogue of an interval bisection task. *The Quarterly Journal of Experimental Psychology. B, Comparative and Physiological Psychology*, *43*(1), 59–81.

Wearden, J. H., & Ferrara, A. (1995). Stimulus spacing effects in temporal bisection by humans. *The Quarterly Journal of Experimental Psychology. B, Comparative and Physiological Psychology*, *48*(4), 289–310.

Wearden, J. H., & Lejeune, H. (2008). Scalar properties in human timing: Conformity and

violations. *The Quarterly Journal of Experimental Psychology*, *61*(4), 569–587.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble Perception. *Annual Review of Psychology*, *69*, 105–129.

Zimmermann, E., & Cicchini, G. M. (2020). Temporal Context affects interval timing at the perceptual level. *Scientific Reports*, *10*(1), 876.

2.2 Duration reproduction under memory pressure: Modeling the roles of visual memory load in duration encoding and reproduction

CONTRIBUTIONS

Zhu, Zang and ZS conceived and designed the experiments. Zhu and Zang collected the data. Zhu analyzed data and implemented the models. Zhu and Zang discussed the results with FA, ZS, SG and HJM. Zhu and Zang interpreted the results and wrote the paper. ZS and FA, SG and HJM commented on and revised the manuscript.

Duration reproduction under memory pressure: Modeling the roles of visual memory load in duration encoding and reproduction

Xuelian Zang[1,2#], Xiuna Zhu[2#], Fredrik Allenmark[2], Jiao Wu[1], Hermann J. Müller[2], Stefan Glasauer[3], Zhuanghua Shi[2]

1. Center for Cognition and Brain Disorders, Affiliated Hospital of Hangzhou Normal University, 310015, China.

2. General and Experimental Psychology, Department of Psychology, LMU Munich, 80802, Germany

3. Computational Neuroscience, Institute for Medical Technology, Brandenburg Technical University Cottbus-Senftenberg, 03046, Cottbus, Germany

# Note Xuelian Zang and Xiuna Zhu contributed equally to this manuscript.

ABSTRACT

Duration estimates are often biased by the sampled statistical context, yielding the classical central-tendency effect, i.e., short durations are over- and long duration underestimated. Most studies of the central-tendency bias have primarily focused on the integration of the sensory measure and the prior information, without considering any cognitive limits. Here, we investigated the impact of cognitive (visual working-memory) load on duration estimation in the duration encoding and reproduction stages. In four experiments, observers had to perform a dual, attention-sharing task: reproducing a given duration (primary) and memorizing a variable set of color patches (secondary). We found an increase in memory load (i.e., set size) during the duration-encoding stage to increase the central-tendency bias, while shortening the reproduced duration in general; in contrast, increasing the load during the reproduction stage prolonged the reproduced duration, without influencing the central tendency. By integrating an attentional-sharing account into a hierarchical Bayesian model, we were able to predict both the general over- and underestimation and the central-tendency effects observed in all four experiments. The model suggests that memory pressure during the encoding stage increases the sensory noise, which elevates the central-tendency effect. In contrast, memory pressure during the reproduction stage only influences the monitoring of elapsed time, leading to a general duration over-reproduction without impacting the central tendency.

Keywords: time perception, dual-task performance, attention-sharing, cognitive/memory load, Bayesian integration, central-tendency effect

INTRODUCTION

Accurate timing is essential for proper actions in our daily activities, such as synchronizing our body movements to a rhythm in music or pronouncing subtly different syllables such as /pa/ and /ba/ with voice-onset time (VOT) in the millisecond range. Yet, subjective time never stops surprising us. Most of us have conscious experiences of situations where time flies or time drags. More surprisingly, distortion of time often happens without us explicitly knowing it. One classical example of such implicit time distortion is the Vierordt effect (better known as the central-tendency effect), reported a century and a half ago (Vierordt, 1868), which describes the phenomenon of short intervals being overestimated and long intervals underestimated (Glasauer & Shi, 2021a, 2021b; Jazayeri & Shadlen, 2010; Lejeune & Wearden, 2009; Shi et al., 2013). Notably, central-tendency effects are ubiquitous in different types of sensory magnitude estimation (Petzschner et al., 2015), such as in spatial distance and angular (rotational-body) displacement judgments (Petzschner et al., 2012; Petzschner & Glasauer, 2011; Teghtsoonian & Teghtsoonian, 1978).

A common explanation of the central-tendency effect is that magnitude estimation is not only based on the sensory measurement but also influenced by past experience, in particular, of the range and distribution of tested intervals. According to the view of Bayesian inference, the brain integrates the sensory measure and prior knowledge together to boost the precision of the estimation – which, while being beneficial in most cases, also engenders a byproduct: a central-tendency bias (for reviews, see Petzschner et al., 2015; Shi et al., 2013). Optimal integration of the sensory input and prior knowledge depends on their respective reliability, measured by the inverse of their variance (the precision). When the sensory measurement has high precision, such as in professional drummers, there would be less influence of prior knowledge, resulting in a lesser central-tendency bias (Cicchini et al., 2012). Importantly, the integration of the sensory measure and the prior is likely involved in the working memory (WM), so that the cognitive load (the demands on WM capacity) may impact both the sensory estimate and the prior representation in terms of their means and variances. However, the role of WM on Bayesian inference of time perception has been largely neglected in the literature.

Although not focusing on Bayesian inference of time perception, Fortin and Rousseau (1998) reported two separable effects of cognitive load on duration encoding and duration reproduction, respectively. In their dual-task design, the secondary task was a Sternberg memory task with a memory set of 1, 3, or 6 digits, presented successively prior to the primary temporal reproduction task, where the latter consisted (on a given trial) of an initial duration-

production phase (with two beeps demarcating a duration) followed by the duration-reproduction phase (two taps generated by the participants). The memory probe (a digit that was or was not part of the memory set) was shown either during the duration-production or the -reproduction phase, and the response to the memory task (either positive or negative) was to be issued using the key of the first (for probes presented during the production phase) or, respectively, the second (for probes presented during the reproduction phase) tap of the temporal reproduction response (one of two keys). Fortin and Rousseau found the reproduced duration to be shortened when the memory probe was presented during the production phase, but lengthened when it was shown during the reproduction phase. They took this finding to support an attention-sharing account (Fortin & Rousseau, 1998; Macar et al., 1994), according to which attentional resources are shared between the timing process and other, non-temporal cognitive processes. When attention is diverted away from the primary task by other concurrent, non-temporal processes in the temporal encoding (i.e., production) phase, the perceived duration is shortened. On the other hand, if the non-temporal process interferes with the reproduction of a given duration, the lapse in the monitoring of the passage of time will lengthen the reproduced duration. Similar findings have been reported in other timing studies (e.g., Fortin & Couture, 2002; Fortin & Massé, 2000), as well as in a study of non-temporal magnitude estimation (Glasauer et al., 2007).

It should be noted that the above studies focused on the over- and under-estimation caused by the memory load. Thus, it remains unclear how memory load influences the uncertainty of the magnitude encoding and prior representation, as well as how it impacts the subsequent magnitude reproduction. Interestingly, a recent study (Allred et al., 2016) on working memory and spatial-length judgments suggests that high cognitive load induced by the secondary working memory task could lead to a coarser memory representation of spatial length, that is, high uncertainty, yielding a strong central-tendency effect. However, in Allred et al., (2016) design, the secondary working-memory task extended across the whole spatial-length judgment task and so did not permit dissociating between load influences on the (length) encoding vs. the (length) reproduction stages. Accordingly, whether the central tendency would be differentially influenced by the cognitive load on the encoding and reproduction phases (in duration judgments) remains unclear.

Taking together the literature reviewed above, we hypothesized that cognitive load would influence both the perceived and reproduced durations, as well as the variability of the estimates, which would further affect Bayesian inference in time estimation (Jazayeri & Shadlen, 2010; Shi et al., 2013; Shi & Burr, 2016). Specifically, we expected increasing

cognitive load during the sensory encoding stage not only to lead to a general underestimation of the duration (in line with Fortin & Rousseau, 1998), but also to decrease the reliability of the estimate. Accordingly, Bayesian integration of the sensory measure and the prior (also referred to as "memory mixing" in Gu & Meck, 2011; Jazayeri & Shadlen, 2010) would predict higher cognitive load to engender a stronger central-tendency bias. By contrast, introducing cognitive load during the reproduction stage would lengthen the reproduced duration (i.e., produce a general overestimation), and likely also increase the variability of the reproduction. However, given that no additional cognitive load is imposed on the sensory encoding stage, the reliability of the sensory estimate, and thus the Bayesian integration, would be unaffected by the cognitive load introduced during duration reproduction. When cognitive load remains high during both the duration production and reproduction phases, the underestimate from the production and the overestimate from the reproduction may cancel each other, at least to some extent and so we may not be able to observe a general bias. However, increasing the uncertainty in the sensory representation may cause a stronger central-tendency effect – a similar pattern to that recently reported in a non-temporal task (Allred et al., 2016).

Table 1 *Experimental designs with the dual tasks*

Experiment	Interleaved order of the dual task
1	Production → Reproduction → Memory phase → Memory test phase
2	Memory phase → Production → Memory test phase → Reproduction
3	Production → Memory phase → Reproduction → Memory test phase
4	Memory phase → Production → Reproduction → Memory test phase

To test these hypotheses, we adopted a dual-task paradigm, consisting of a secondary *visual* working-memory task (with low, medium, or high load) and a primary duration production-reproduction task, in four experiments (see Table 1). Importantly, we manipulated the 'spanning' of the secondary task – that is, the period over which the memory-set items had to be maintained – in relation to the primary timing task in such a way that the memory task influenced different stages (production, reproduction, or both) of the timing task. Specifically, in Experiment 1, the memory task was introduced after the timing task, providing a baseline. In Experiment 2, the memory task spanned the duration-production phase, to examine the

impact of cognitive load on duration encoding. In Experiment 3, by contrast, it spanned the duration-reproduction phase to examine the impact of cognitive load on the duration reproduction. Finally, in Experiment 4, the memory task extended over the whole timing task (i.e., both the production and reproduction phases), to examine the combined effect of a non-temporal task on the timing task, while also serving as a comparison to the study of memory load on the central tendency in a spatial task (Allred et al., 2016).

Given that the attentional sharing account (Fortin & Rousseau, 1998; Macar et al., 1994) makes clear predictions of how memory load would influence the production and reproduction stages, and Bayesian inference makes quantitative predictions of how the sensory estimate is integrated with the prior, we also developed a general computational framework for duration encoding, Bayesian integration and duration reproduction, taking into consideration possible influences of cognitive load, based on our behavioral findings. We identified the role of attentional sharing and the temporal stage of prior integration in time estimation. Specifically, owing to attentional sharing between the concurrent working-memory task and duration estimation, participants displayed underestimation with memory pressure during the production, but overestimation in the reproduction phase. These two opposing influences canceled each other, diminishing the general bias when memory pressure covers both the production and reproduction phases. Moreover, we found an increased central-tendency effect with memory pressure during the production phase, but no significant changes with memory pressure during the reproduction phase.

METHODS

Participants

Different groups, each of 16 volunteers (23-34 years old), were recruited for each experiment (9 females in Experiments 1, 2, and 4, 8 females in Experiment 3); all of them had self-reported normal or corrected-to-normal vision and normal hearing. The sample size was determined based on (Fortin & Rousseau, 1998) study, in which ten participants yielded significant under- and over-estimations. On the conservative side, we recruited 16 participants. All participants were naive as to the purpose of the experiments and received 9 Euro per hour for their service. The experiment was approved by the Ethics Committee of the Department of Psychology of LMU Munich.

Apparatus

The experiments were conducted in a sound-isolated, dimly lit cabin (5.24 cd/m^2). The visual stimuli were presented on a 21" LACIE CRT monitor, with a refresh rate of 100 Hz. The viewing distance was fixed to 57 cm (maintained by the use of a chin rest). The experimental program was developed using Matlab (Mathworks Inc.) and Psychtoolbox (Kleiner et al., 2007).

Stimuli and tasks

The experiments were dual-task experiments, consisting of a duration production-reproduction task and a visual working-memory task on each trial.

The ***duration production-reproduction task*** (see Figure 1 for an example) consisted of two phases: In the first, production phase, a grey disk (36.5 cd/m^2, 6.7° in diameter) was presented in the center of the monitor (on a dark background: 16.7 cd/m^2) for a given duration, randomly sampled from 500, 800, 1100, 1400, or 1700 ms. Participants were instructed to encode and retain the duration of the grey disk. In the following reproduction phase, participants were asked to reproduce the perceived duration of the grey disk as accurately as possible by pressing and holding the down-arrow key. The key-press triggered a visual display with a grey disk, which stayed on the screen until the key was released. Participants were asked to reproduce, as accurately as possible, the (presentation) duration of the grey disk from the production phase.

The ***visual working-memory task*** also consisted of two phases: a memory phase and a test phase (see Figure 1). During the memory phase, a number (randomly selected 1, 3, or 5) of squares were presented on an invisible circle (diameter approximately 19. 64°) on the dark background. Each square (subtending 3. 35° × 3. 35°) was filled with a color randomly selected from 180 color values uniformly distributed along a circle in CIE 1976 (L* = 70) color space (van den Berg et al., 2012; Zhang & Luck, 2008). Adjacent items were arranged equidistantly in displays with three or, respectively, five items; in one-item displays, the single item appeared always at the bottom position on the invisible circle. Participants had to encode and retain the color in which a square (at a given location) appeared (with the memory load increasing with the number of squares). In the test phase, one item location from the memory phase was selected as the location of the test (or probe) stimulus. The color of the probe square was either the same or different (randomly and equally determined) relative to the previous

(memory) item at that location. In case of a "different" probe, the color was randomly selected from the remaining 179 possible colors. [Note that the probe square was presented without the thin grey outline that framed the items in the memory phase (see Figure 1).] Participants were asked to indicate whether the probe item had the same or a different color as the previous item at that location by pressing either the left (same) or right (different) arrow key.

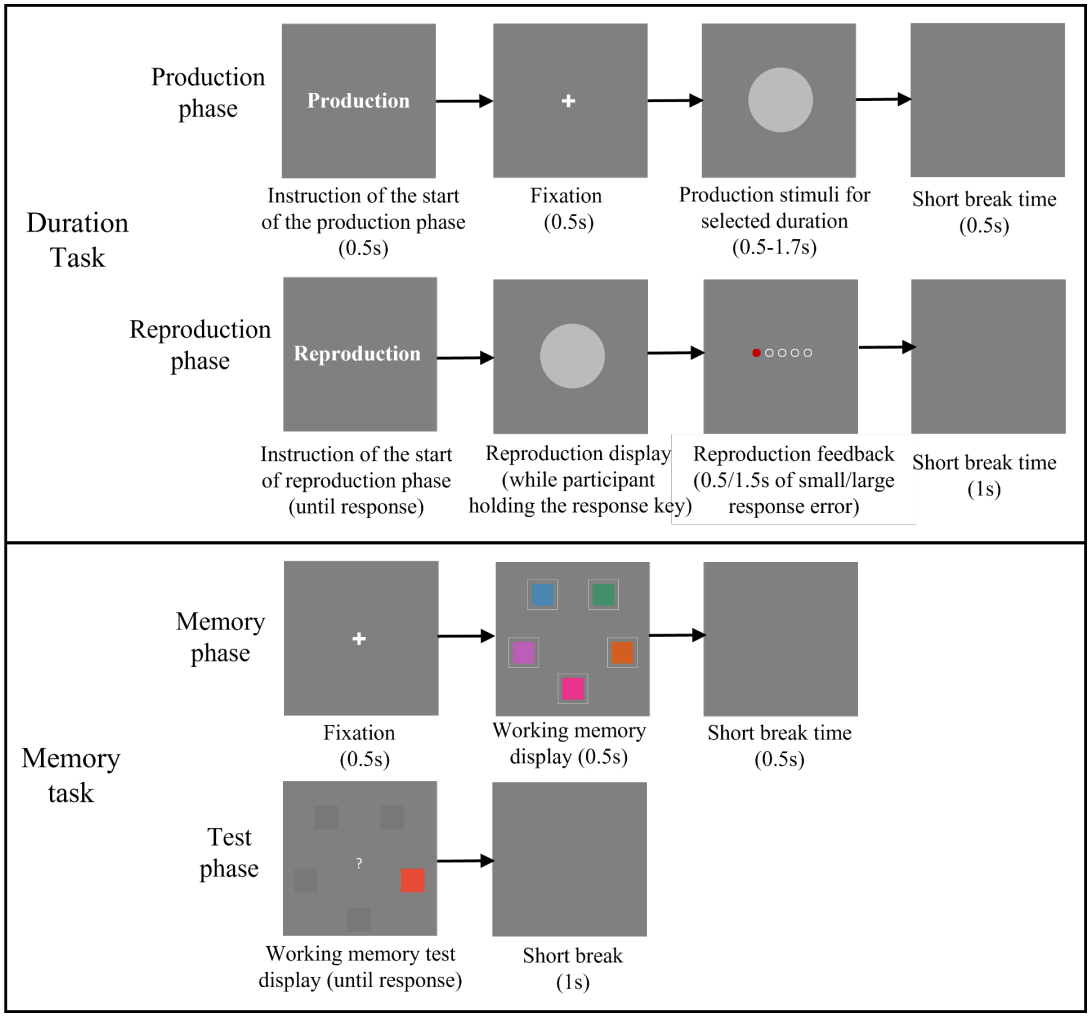

Figure 1 *Schematic illustration of the dual tasks used in Experiment 1. The duration task includes the production and reproduction phases (depicted in the upper two panels). The working memory task also includes two phases: the memory phase and the test phase (depicted in the lower two panels). In the example, the correct answer would have been "different".*

Design and procedure

Experiment 1 served as a baseline. Participants were presented with the two tasks successively: the duration task first and the memory task second. Thus, any interference between the two tasks would have been minimal (see Table 1). The experiment consisted of 18 blocks of 20 trials each. The working-memory load (1, 3, or 5 items) was fixed per block, but

randomly counter-balanced across the 18 blocks. Using a block-design for the memory load in this baseline experiment was meant to rule out any possible interference of the memory load on the updating of the duration prior across trials.

The duration task started with the cue word 'Production' on the screen for 500 ms, indicating the *production phase*. Then, a fixation dot appeared for 500 ms, followed by a grey disk in the center of the monitor; this remained visible for a given trial duration (0.5 to 1.7 s), after which the display turned back to a blank screen for 500 ms. Next, the second cue display with the word 'Reproduction' appeared, prompting the start of the *reproduction phase*. Participants followed their own pace to initiate the reproduction, which required them to press and hold down the down-arrow key. Immediately after the key press, a grey disk appeared and remained visible until the key was released (i.e., the visible disk duration served as the reproduced event). The duration of the key pressing was recorded as the reproduced duration. After the participant released the response key, a feedback display was presented showing the relative reproduction error (i.e., the ratio of the reproduction error to the given trial duration) by highlighting one of five linearly arranged dots (see Figure 1). The five dots, from the left to the right, were mapped to the relative error ranges: below -30%, between [-30%,-5%], [-5%, 5%], [5%, 30%] and greater than 30%, respectively. The three central dots were colored green, and the left- and right-most dots red, indicating how large a reproduction error was made. When the relative error was between -30% and 30%, the error feedback display was presented for 500 ms; otherwise, it was presented for 1500 ms, alerting participants that the error was too high. After a break of 1 s with a blank screen, the working-memory task started with a fixation cross presented for 500 ms, followed by a memory display containing one, three, or five colored squares visible for 500 ms (*memory phase*). After a blank screen of 500 ms, a probe display with a single color patch at one of the previous (i.e., memory-phase item) locations (*test phase*). Participants had to indicate whether this probe patch was colored the same vs. differently relative to the (coincident) item in the memory display, by pressing the left ('same') or the right ('different') arrow key. After a one second interval, the next trial began.

Experiment 2 examined whether maintaining information in working memory during the production (but not the reproduction) phase would affect the sensory duration measurement and further influence the use or updating of priors. The tasks and displays were essentially the same as in Experiment 1, except that, in the trial-event sequence, the working-memory task spanned the duration production (and not the reproduction) phase. That is, each trial started with the *memory phase*, followed by the duration *production phase*. Next, participants performed the *memory test*, before proceeding to the duration *reproduction phase* (see Table 1

and Figure 1). Given that the prior updating was not impacted by memory load in the baseline experiment (see Results of Experiment 1), we adopted a trial-wise design for the memory load in Experiments 2 to 4, effectively making the load unpredictable; accordingly, the influence of memory load (if any) would be locally trial-based for each tested duration.

Experiment 3 was essentially the same as Experiment 2, except that now the working-memory task spanned the reproduction (rather than the production) phase. That is, each trial started first with the *production phase*, followed by the *memory phase*; then, participants had to *reproduce* the given (trial) duration and finally perform the *memory test* (see Table 1 and Figure 1).

Experiment 4 examined how duration estimation would be affected by maintaining the working-memory information across both the production and reproduction phases. That is, the task sequence on each trial was: memory phase, duration-production phase, duration-reproduction phase, and finally memory-test phase (see Table 1).

MODELING OF DURATION ESTIMATION UNDER MEMORY LOAD

The duration production-reproduction task involves the component stages of duration encoding, Bayesian integration, and duration reproduction. Here we proposed a generative processing architecture and potential influences of the memory load on these stages.

1. Duration encoding

We assume that, while the 'raw' sensory measure (S) of given sample duration (D) is not influenced by cognitive load, its representation in working memory S_{wm} may be affected by the load. Given that the scalar property (i.e., Weber scaling) is the key feature of duration estimation (Gibbon et al., 1984; Shi et al., 2013), we further assume logarithmic scaling of the sensory measure (S), to simplify calculation; that is, S$\sim N(\mu_s, \sigma_s^2)$, where μ_s is the mean of the logarithmic scale representation of the given sample interval D (see Petzschner & Glasauer, 2011; Roach et al., 2017), and σ_s^2 reflects the variance of internal-measurement noise (ϵ):

$$S = \ln(D) + \epsilon\,, \qquad \text{(Eq. 1)}$$

According to the classical internal-clock and attentional-gate models (Block & Zakay, 1997; Gibbon et al., 1984), sharing of attention by a concurrent non-temporal task would lead to an extra loss of the accumulated clock ticks, as well as increasing the noise of the memory representation of the sensory measurement (S_{wm}). Here, for modeling the concurrent memory-

load effects, we assume – as a simple approximation and following the principle of 'Occam's razor' – that both the number of ticks and the noise modulation (as represented on the logarithmic scale) are *linearly* affected by the memory load.[1] That is, the memory representation is normally distributed, $S_{wm} \sim N(\mu_{wm}, \sigma_{wm}^2)$, with both the mean μ_{wm} and the variance σ_{wm}^2 influenced by the memory load linearly:

$$\mu_{wm} = \ln(D) - k_s \cdot M \ , \qquad \text{(Eq. 2)}$$

and

$$\sigma_{wm}^2 = \sigma_s^2 + l_s \cdot M + ts \cdot G, \qquad \text{(Eq. 3)}$$

where M represents the level of the working-memory load (set to 1, 2, and 3 for the low, medium, and high load, respectively[2]), and k_s and l_s are scaling factors of the mean and the variance of the memory representation, respectively. Note that based on the attentional-gate model, we specifically assume that the cognitive-load influence on duration encoding in Eqs. 2 and 3, k_s and l_s, are constrained to be non-negative ($k_s \geq 0$, $l_s \geq 0$). Given that the duration-production phase was *non*-overlapping with the secondary working-memory task in Experiments 1 and 3, k_s and l_s were set to zero for those experiments (i.e., there would be no influence of the secondary task on the production phase). In more detail, $k_s \cdot M$ represents the loss of clock ticks during the accumulation process. In addition, the cognitive load increases the variance of the memory representation in a linear fashion (Bays, 2015), which is captured by the term $l_s \cdot M$.

2. Bayesian integration

The classical central-tendency effect shown in duration reproduction can be explained by memory mixing between the internal prior of the sampled durations and the sensory measure (Acerbi et al., 2012; Gu et al., 2016; Jazayeri & Shadlen, 2010; Penney et al., 2000). Given that the sampled durations were the same across the three memory-load conditions, we assume the internal prior was the same for the different memory loads, following the normal distribution

---

[1] This 'linearity' assumption does not rule out that the true relation is more complex. In principle, the linear constraint can be relaxed by assuming independent impacts of the memory load on the memory representation. However, such an approach has more free parameters than the current model, and it does not provide additional insights of the underlying mechanism.

[2] In the memory-load task, the set sizes were 1, 3, and 5 items of the low-, medium-, and high-load conditions, respectively. It has been shown that the variance is an approximate linear function of set size (for a review, see Bays, 2015). In addition, performance accuracy in the memory task decreased approximately linearly with the set size (see Figure 2). Accordingly, here we set M as a simple linear function of the set size (M = 2 × Set Size – 1) and assume linear relations in Eqs. 2 and 3, without loss of generality.

with the mean μ_p and the variance σ_p^2 (both parameters in logarithmic space). The internal prior is then integrated with the memory representation of duration S_{wm} according to the Bayes rule, which minimizes the uncertainty of the final duration representation. According to Bayesian inference, the posterior distribution $D'_{post} \sim N\left(\mu'_{post}, {\sigma'}^2_{post}\right)$ of the final memory representation can be estimated by

$$\mu'_{post} = w_p\mu_p + \left(1 - w_p\right)\mu_{wm} \quad \text{(Eq. 4)}$$

$${\sigma'}^2_{post} = \frac{\sigma_p^2 * \sigma_{wm}^2}{\sigma_p^2 + \sigma_{wm}^2} \quad \text{(Eq. 5)}$$

The optimal weight w_p is proportional to the inverse variability of the priors ($w_p = \frac{\frac{1}{\sigma_p^2}}{\frac{1}{\sigma_p^2}+\frac{1}{\sigma_{wm}}^2}$), which indicates the relative influence of the prior. The weight plays a key role in the central-tendency effect (Jazayeri & Shadlen, 2010; Shi et al., 2013): larger values of w_p mean a stronger central-tendency effect.

3. Duration reproduction

Given that attention is required for the time-monitoring during the reproduction stage, the memory load might influence the final reproduction output. Our model assumes that the memory load influences the monitoring of the elapsed time of the reproduction, that is: there is a continuous monitoring of the elapsed time starting from the key press ($\mu_{elapsed}$) and comparison of the elapsed time with the duration held in memory (μ'_{post}). Similar to the production stage, the sensory representation of the elapsed time can be distorted by the memory load. Accordingly, the comparison is:

$$\left|\frac{\mu'_{post} - \left(\mu_{elapsed} - k_r M\right)}{\mu'_{post}}\right| < \epsilon\,, \quad \text{(Eq. 6)}$$

which is equivalent to comparing the elapsed time to the mean duration $\mu'_{post} + k_r M$. In order to compare the model calculations to the observed reproduction behavior, we transferred the logarithmic space representation in the model back to the linear space, using the lognormal distribution with mean and variance:

$$\mu_r = e^{\mu'_{post} + k_r M + \frac{{\sigma'}^2_{post}}{2}} \quad \text{(Eq. 7)}$$

$${\sigma_r}^2 = \left|e^{{\sigma'}^2_{post}} - 1\right| \cdot e^{2\left(\mu'_{post} + k_r M\right) + {\sigma'}^2_{post}} \quad \text{(Eq. 8)}$$

In addition, we assume the uncertainty of reproduction is further influenced by motor noise $D_{mn} \sim N(0, \sigma_{mn}^2)$, but with the the influence of the motor noise decreasing as the duration increases:

$$\sigma_{observed}^2 = {\sigma_r}^2 + \frac{\sigma_{mn}^2}{D} \quad \text{(Eq. 9)}$$

Thus, we have seven parameters in total in the above model framework: the memory load scaling parameters of the mean and variance of the duration encoding (k_s, l_s), the variance of sensory measurement σ_s^2, the memory-load scaling factor k_r in the reproduction phase, the motor noise ${\sigma^2}_{mn}$, and the mean and variance of the prior ($\mu_p, {\sigma_p}^2$).

Moreover, the general over-/under-estimation is usually measured by the average mean reproduction, which has a positive linear relation to an indicator called the indifferent point (IP), i.e.: the point at which the duration reproduction is veridical (Lejeune & Wearden, 2009)[3]. Based on Eqs. 2, 4, and 7, and letting the reproduced duration equal to the given duration (by the definition of IP), the log of the indifference point can be written as

$$ln\,(IP) = [k_r M\ -\ (1 - w_p)k_s M + {\sigma'^2}_{post}/2 +\ w_p\mu_p]/w_p \quad \text{(Eq. 10)}$$

As we can see from Eq. 10, the first-order impact of the memory load is the combination of the scaling factors k_s and k_r, and the second-order impact is mediated through the variance (μ_p and w_p).

4. Parameter estimation

We adopted Stan, a platform for statistical modeling and Bayesian statistical inference (Bürkner, 2016; Stan Development Team, 2020), to estimate the parameters in the hierarchical Bayesian modeling. To complete the Bayesian hierarchical model, we have used standard non-informative priors on those hyperparameters (see Figure 2). In our proposed model, the seven parameters θ =$\left(k_s, \sigma_s, l_s, \mu_p, \sigma_p, k_r, \sigma_{mn}\right)$ are sampled from their respective prior distributions p(θ) with hyperparameters. The distributional belief about parameters θ can be denoted as a conditional probability function $p(\theta|D_{observed})$. Using the Bayes rule, the posterior distribution $p(\theta|D_{observed}) \propto p(\theta) \times p(D_{observed}|\theta)$ can be derived from the prior distribution, p(θ), and a likelihood, $p(D_{observed}|\theta)$.

After compiling the model specification from Stan's probabilistic programming

[3] Suppose the duration reproduction engenders an approximate linear central tendency effect. With simple mathematical calculation, the mean reproduction bias is equivalent to w_p times the difference between the indifference point (IP) and the mean sample duration.

language to a C++ program, we used the Markov Chain Monte Carlo (MCMC) sampling method in RStan with 8000 iterations per chain (total of four chains) to estimate the parameters for individual participants by maximizing the joint posterior distribution of parameters of interest. The estimation was performed using the R-package RStan (Carpenter et al., 2017; Stan Development Team, 2018), with the data and R-code available at https://github.com/msenselab/working_memory_reproduction.

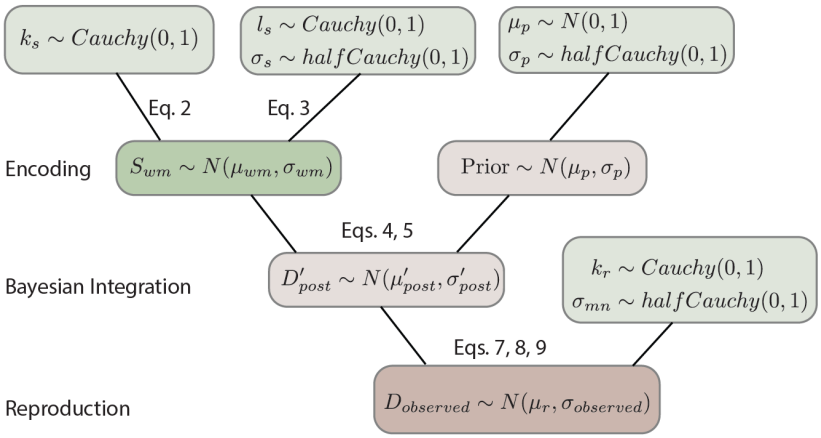

Figure 2 *Schematic illustration of the Bayesian hierarchical model through RStan in parameter estimation. Light green boxes show the standard distributions used for the hyperparameters. The green box represents the duration encoding stage, light brown boxes Bayesian integration stage, and the brown box reproduction stage.*

RESULTS

Memory task

The mean accuracy of individual participants for the working memory test was calculated by the proportion of correct responses, including the 'hit' responses (for trials with the same color of the probe and the memory item) and the 'correct rejection' responses (for the trial with different colors for the probe and memory item). Figure 3 shows the approximate linear relation between memory load and the mean accuracy in the memory test across the four experiments, with an overall decrease of the correct rates from Experiment 1 to 4.

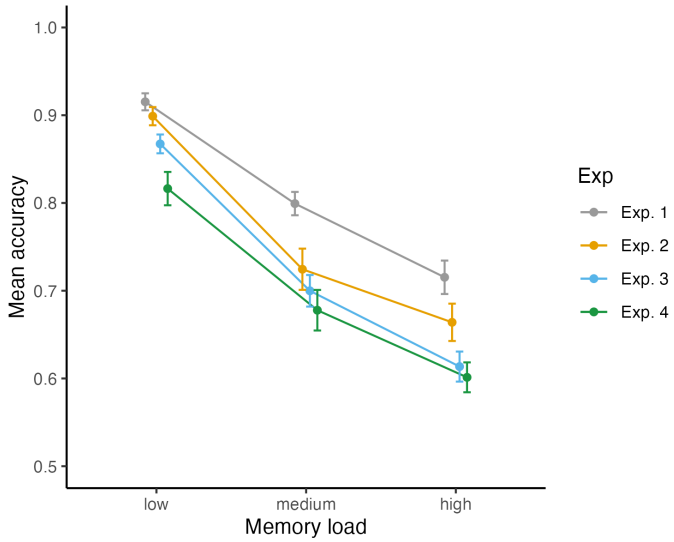

Figure 3 *Mean accuracy with associated standard errors (n=16) for the (secondary) working-memory task as a function of the memory load (low, medium, high), separated for Experiments 1-4.*

A mixed-design ANOVA of the mean correct rates with Experiment (1–4) as between-subject factor and Memory Load (small, medium and high) as within-subject factor revealed a significant main effects of Experiment, $F(3, 60) = 11.15$, $p < .001$, $\eta_g^2 = .36$, and of Memory Load, $F(2, 120) = 454.22$, $p < .001$, $\eta_g^2 = .88$. However, the Experiment × Memory-Load interaction did not reach significance, $F(6, 120) = 2.04$, $p = .066$, $\eta_g^2 = .093$. The mean accuracy in the working-memory task decreased linearly as the memory load increased from the 1 to 5 items (linear effect, $t(30) = -29.181$, $p < .001$, mean of .87, .73, .65 of low, medium and high load respectively), indicating that the manipulation of memory load was effective: larger set sizes engendered higher loads. Further LSD comparisons revealed memory accuracy to be significantly higher in Experiment 1 (mean of .81) than in Experiments 2–4 (means of .76, .73, and .70, respectively; all ps $< .025$), indicating that the working-memory performance declined significantly when the memory task was intermixed with the duration task (in Experiments 2–4). Interestingly, performance was also significantly better in Experiment 2 than in Experiment 4 ($p < .001$), that is: when the memory task was performed first (i.e., when it overlapped with the duration-production phase) vs. second (i.e., when it overlapped with the reproduction as well as the production phase).

Duration task

Figure 4 shows the reproduction biases and coefficients of variation (CVs) for all four experiments. The CV is a standardized measure of dispersion of reproduced durations (i.e.,

normalizing the standard deviation by the duration), which is a close approximation to the Weber fraction. As can be seen, there was a central-tendency effect in all experiments, with the short intervals being overestimated and long intervals underestimated, and the CVs decreased as the tested (i.e., to-be-reproduced) durations increased.

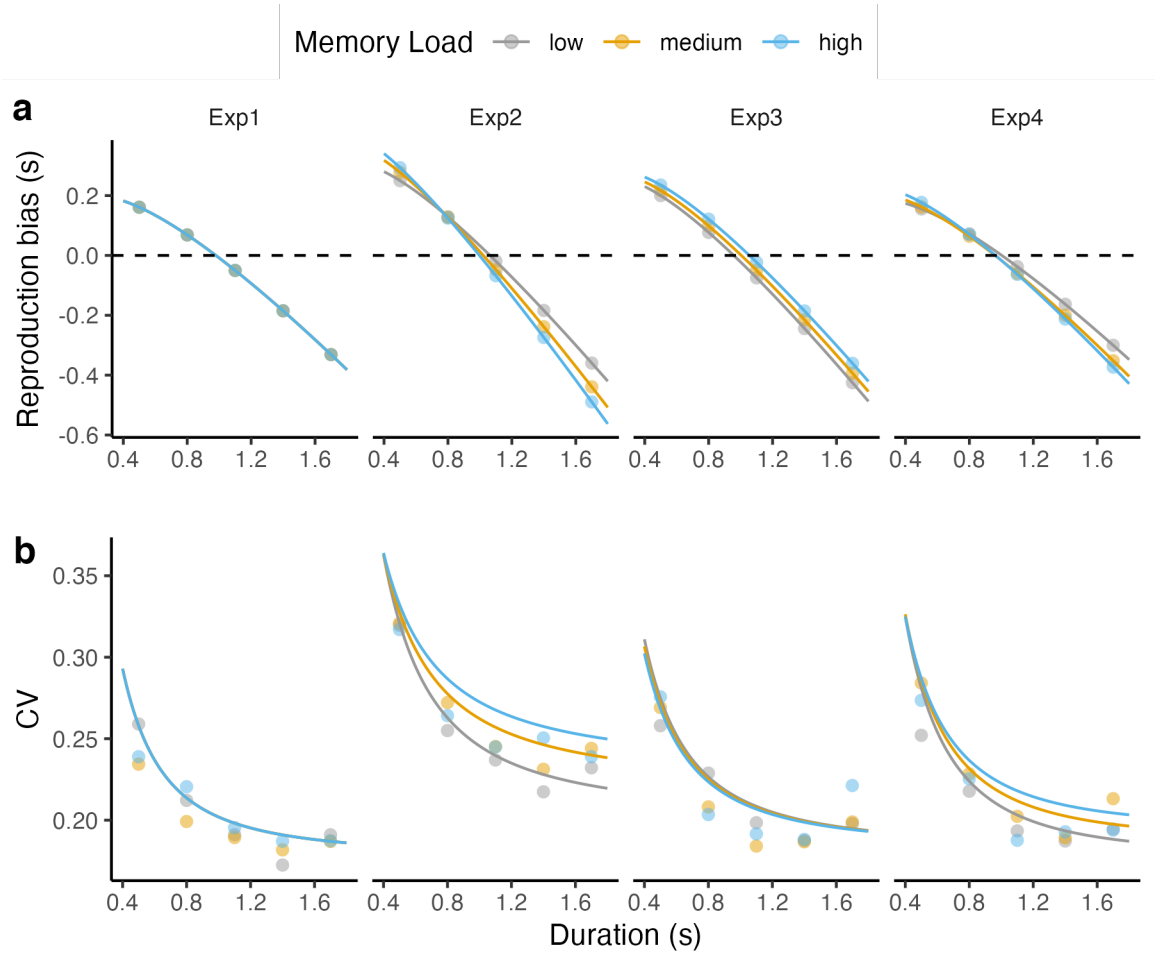

Figure 4. *(**a**) Mean reproduction biases and (**b**) coefficients of variation (CVs), for the four experiments. The dots represent the observed mean data, the curves the predictions from the Bayesian model outlined in the modeling section; the gray, orange, and cyan colors represent the low, medium, and high memory-load conditions, respectively.*

The central tendency effects

The central-tendency indices, calculated as the weight of the prior w_p in Bayesian modeling (Figure 4a), were significantly larger than 0 in all four experiments (all w_ps > 0.36, all *t*s > 8.14, all *p*s < 0.001), confirming robust central-tendency effects in our duration estimation tasks.

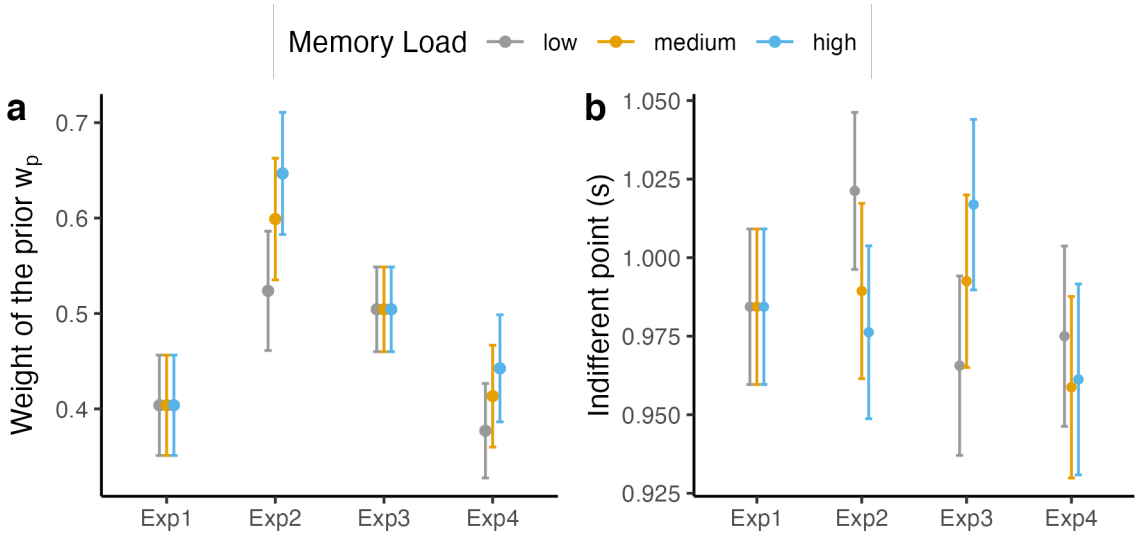

Figure 5 ***(a)*** *Mean central-tendency indices measured by the estimated weight of the prior (w_p) , separately for the three memory-load conditions in the four experiments.* ***(b)*** *Mean Indifferent Points, separately for the three memory-load conditions in the four experiments. The gray, orange, and cyan colors represent the low, medium, and high memory-load conditions, respectively. Note, w_p was set to the same value for Exps. 1 and 3.*

Because the working-memory task was introduced after the production phase in Experiments 1 and 3, the model assumed no influence of memory load on the central tendency (see Eq. 4 and Figure 3a). The mean weight of the prior (± standard error, SE) was 0.40 ± 0.05 in Experiment 1 and 0.50 ± 0.04 in Experiment 3. Interestingly, the model was able to predict the differential central-tendency effects between Experiments 2 and 4 (see the goodness of fit in Figure 3, and Appendix A). Repeated-measures ANOVAs and tests of the linear relation of the central-tendency indices (w_p) to the factor Memory Load (low, medium, high), conducted separately for Experiments 2 and 4, revealed a significant linear increase of the central tendency with increasing memory load in both Experiment 2 [mean central tendency for the low, median, and high memory loads: 0.52 ± 0.06, 0.6 ± 0.06 , 0.65 ± 0.06, respectively; ANOVA: $F(1.01, 15.2) = 55.09$, $p < .001$, $\eta_p^2 = .79$; linear effect, $t\ (30) = 10.41$ $p < .001$] and Experiment 4 [mean central tendency: 0.38 ± 0.05, 0.41 ± 0.05, and 0.44 ± 0.05, respectively; ANOVA: $F(1.01,15.2) = 33.98$, $p < .001$, $\eta_p^2 = .7$; linear effect: $t(30) = 8.23$, $p < .001$]. Recall that both Experiments 2 and 4 involved working-memory pressure in the duration-production phase, and this in turn caused stronger central-tendency effects for the high vs. the low memory loads.

When collapsing all levels of memory loads together for individual experiments and comparing across experiments, the mean central-tendency indices turned out to differ

significantly among the four experiments [one-way ANOVA: $F(3, 60) = 2.87$, $p = .044$]. Further LSD comparisons revealed significantly higher central-tendency indices in Experiment 2 (.590 ± .061) than in Experiment 1 (0.40 ± .05) and 4 (0.41 ± .04), all ps < .05; no other comparisons reached significance. The relatively stronger central-tendency effect in Experiment 2 may be caused by the increased sensory noise in having to perform the memory test task between the production and reproduction phases: this could increase sensory noise either due to the prolonged production-reproduction phase interval or the attentional sharing between the two tasks.

In a further, Wilcoxon-test analysis, we compared the model factor l_s (see Eq. 4), indicative of the extent to which the variance of the sensory measure is modulated by memory load, between Experiments 2 and 4: l_s turned out to be significantly higher in Experiment 2 [mean of 0.201 ± 0.139] vs. Experiment 4 [0.013 ± 0.004; W = 217, p <.001]. That is, sensory noise was larger when the working-memory task overlapped only with the production phase of the duration task (in Experiment 2) as compared to when it overlapped with both the production and reproduction phases (in Experiment 4). In turn, the larger sensory noise led to a higher central-tendency effect in Experiment 2 than in Experiment 4.

The mean indifferent points (IP) and general reproduction bias

The mean Indifferent Points (IPs; shown in Figure 4b), at which the duration reproduction is veridical, can be used to index participants' general reproduction bias: an IP larger than the mean tested duration is indicative of a general overestimate, while a smaller IP is indicative of a general underestimate. A mixed-design ANOVA with Memory Load (low, medium and high) as within-subject factor, Experiment (1–4) as between-subject factor revealed the Load × Experiment to be significant [$F(6, 120) = 6.110$, $p < .001$, $\eta_p^2 = .23$], while the main effects were non-significant [Memory Load: $F(1.01, 40.69) = .429$, $p = .652$, $\eta_p^2 = .007$; Experiment: $F(3, 60) = .285$, $p = .836$, $\eta_p^2 = .014$]. To understand the interaction effect, we examined the Load effect on the IPs separately for each experiment. In Experiment 1, the IPs were essentially the same for the three memory-load conditions (.984 ± .024; see also Figure 3a), given that the working-memory task was introduced after the completion of the duration task. In Experiment 2, the IPs *decreased* linearly with increasing memory load [one-way ANOVA: $F(1.09, 16.37) = 5.83$, $p = .007$, $\eta_p^2 = .28$, linear effect: $t(30) = -3.32$, p = .002, means of 1021 ± 26 ms, 989 ± 26 ms, and 976 ± 27 ms for the low, median, and high memory-load conditions, respectively], indicating that a memory load imposed on the production phase

caused a significant underestimation of the duration. In Experiment 3, the IPs increased linearly with increasing memory load [$F(1.03, 15.48) = 8.75$, $p = .007$, $\eta_p^2 = .37$, linear effect: $t(30) = 4.182$, $p < .001$; means of 966 ± 28 ms, 993 ± 27 ms, 1017 ± 26 ms, respectively]. The pattern was opposite to that in Experiment 2, showing that a memory load imposed on the reproduction phase led to an overestimation of the tested duration. In contrast, when the memory load spanned both the duration-production and reproduction phases in Experiment 4, the IPs were comparable among three different memory-load levels [one way ANOVA: $F(1.04, 15.62) = .74$, $p = .49$, $\eta_p^2 = .05$; means of 975 ± 28 ms, 959 ± 28 ms and 961 ± 29 ms, respectively]. The comparable IPs among three memory load conditions suggests that the opposite effects of Memory Load observed in Experiments 2 and 3 cancel each other out when the memory tasks spans the whole (production plus reproduction phases of the) duration task.

Recall that participants' general biases were modeled by two scaling factors, k_s and k_r: k_s represents the magnitude of the duration shortening per unit of memory load in the production phase, and k_r the magnitude of the duration lengthening per unit of memory load in the reproduction phase. When compared to zero, one sample t-tests revealed a significantly positive k_s-value in Experiment 2 ($k_s = .263 \pm .077$, $t = 3.409$, $p = .004$), a significantly positive k_r-value in Experiment 3 ($k_r = .025 \pm .003$, $t = 7.212$, $p < .001$), and both a significantly positive k_s-value and a significantly positive k_r-value in Experiment 4 ($k_s = .459 \pm .052$, $t = 8.824$, $p < .001$; $k_r = .229 \pm .047$, $t = 4.924$, $p < .001$). These results are confirmatory of the attentional-sharing hypothesis, that is: the concurrent memory and duration estimation tasks share the same attentional resource, which in turn leads to a duration underestimation with working memory pressure during the production phase, but duration overestimation with memory pressure during the reproduction phase. When the memory pressure occurs in both phases, the duration underestimation in production and overestimation in reproduction may cancel out each other, resulting in the diminishing of the general bias.

Coefficient of variations of the duration reproduction

The observed Coefficients of Variation (CVs) for each reproduced sample duration, calculated as $CV_i = \frac{\sigma_i}{R_i}$, where σ_i and R_i represent the standard deviation and the mean of the reproduction of a given interval D_i, are shown in Figure 4b. A mixed-design ANOVA of the CVs, with Sampled Duration (500, 800, 1100, 1400, 1700 ms) and Memory Load (low, medium, high) as within-subject factor and Experiment as between-subject factor revealed only the main effects of Duration and Experiment to be significant; no other effects (including the

main effect of memory load and all interactions) reached significance (all Fs < 1.47, all ps > .19). The main effect of Duration (F(2.429, 7.286) = 47.794, p < .001, η_p^2 = .44) was due to the CVs decreasing linearly with increasing sample duration (means of .274 ± .008, .229 ± .006, .205 ± .005, .195 ± .005, and .209 ± .006 from short to long durations, respectively; linear effect: t(60) = -11.423, p < .001]. And the main effect of Experiment [F(3, 60) = 6.535, p = .001, η_p^2 = .246; means of .203 ± .010, .259 ± .010, .213 ± .010, .and 216 ± .010 in Experiment 1–4, respectively] was owing to the significantly higher CV in Experiment 2 compared to the other experiments (LSD tests comparing Experiment 2 with Experiment 1, 3, and 4, respectively: all ps < .003; comparable CVs in Experiment 1, 3, and 4). The relatively larger response variation in Experiment 2 is likely due to the working-memory test phase being administered in-between the duration-production and reproduction phases, while in all the other three experiments the memory task was tested at the end of the reproduction phase.

Goodness of the modeling

The Bayesian model outlined above predicted the behavioral results strikingly well, in all four experiments. The model performance may be gauged by calculating the means of the relative prediction error on the estimated durations $\left(Mean\left(\frac{|D_{observed}-D_{predicted}|}{D_{observed}}\right)\right)$ and on the estimated variances ($Mean(|SD(D_{observed}) - sd(D_{predicted})|/sd(D_{observed}))$). The results revealed less than 4.17% error on the reproduced durations for all experiments (3.34%, 4.17%, 3.44%, and 3.53% for Exp. 1–4 respectively; see Appendix A for further details about the goodness of the model fitting).

GENERAL DISCUSSION

The present study examined cognitive-load interference in duration estimation using a Bayesian approach. Through computational modeling of the results from four experiments, we attempted to provide a generative model of load interference on the whole champion of processes in duration estimation, including duration encoding and reproduction. The Bayesian model we proposed predicted not only the mean but also the coefficient of variation (CV) of reproduction behavior.

We found a visual working-memory load to interfere with participants' duration reproduction both when the load was imposed during the duration-production and during the reproduction phase, albeit in a different way. In more detail, when the working-memory task

overlapped only with the production phase (in Exp. 2), participants on average underestimated the tested durations, and they exhibited a stronger central-tendency effect under high- vs. low memory-load conditions. In contrast, when the working-memory task overlapped only with the reproduction phase (in Exp. 3), the higher the memory loaded, the more duration participants over-reproduced, while the central-tendency effect was comparable across the different load conditions. Of note, when the working-memory task spanned both the production and reproduction phases (in Exp. 4), there was no longer a general over- or underestimation, but the central-tendency effect remained stronger with higher vs. lower memory loads. Finally, varying levels of memory load introduced between consecutive duration reproductions (in Exp. 1) had no discernible effects on either general reproduction biases (i.e., there was no general over-/underestimation) or the central-tendency bias, suggesting that the prior updating of the sampled durations was not influenced by the intermediate secondary task.

Importantly, this pattern of findings could be well fitted by our Bayesian model, which integrated the notion of attentional-resource (Fortin, 2003; Fortin & Rousseau, 1998; Macar et al., 1994) between two concurrent tasks. According to the Bayesian inference model (Jazayeri & Shadlen, 2010; Petzschner et al., 2015; Shi et al., 2013), the reproduced duration reflects an optimal integration (according to the Bayes rule) of the sensory estimate of a given duration with the prior distribution stored in the memory, where 'optimal' refers to achieving minimal variability in the final estimate. A by-product of this is that the duration estimates assimilated to the mean prior, as evidenced in the typical central-tendency bias. Standard Bayesian inference models make no assumptions about how memory load may influence Bayesian inference. To address this, here we combined the attentional-sharing account (Fortin, 2003; Fortin & Rousseau, 1998; Macar et al., 1994) with standard Bayesian inference, assuming that time units would be lost in the duration encoding and reproduction stages when attention is shared with a secondary task. Prior work (Fortin, 2003; Fortin & Rousseau, 1998) had shown that the loss of time units in the encoding and reproduction stages has a differential impacts: when attention is diverted away from the primary (temporal) task by another, concurrent non-temporal task during the duration-encoding phase, a certain amount of clock ticks would be lost, resulting in a shortened time estimation (underestimation); in contrast, when the secondary task is performed concurrently with the reproduction phase, the resulting loss of clock ticks (due to lapses in monitoring the elapsed time) would lead to a reproduced duration longer than the tested interval (overestimation). This descriptive explanation is quantitatively characterized by the linear scaling parameters k_s and k_r in our Bayesian model (Eqs. 2 and 7). Both the behavioral findings and the model confirm the dissociable influences of memory pressure at

different stages of time estimation. Specifically, the concurrent working-memory task gave rise to an underestimation when imposed during the production phase ($k_s > 0$), but an overestimation when imposed during the reproduction phase ($k_r > 0$).

While the attentional-sharing account (Fortin, 2003; Fortin & Rousseau, 1998; Macar et al., 1994) can well explain the general over- and, respectively, underestimation of the sample durations, it falls short in explaining the differential central-tendency effects. While there were central-tendency effects in all four experiments, the central-tendency bias was significantly modulated by memory load only in Experiments 2 and 4, that is, when the secondary memory task overlapped the production phase. We take this to indicate that the central-tendency bias was introduced mainly in the duration-encoding stage: the memory load imposed during this stage increased the uncertainty of the sensory measure (which we modeled with the scaling parameter l_s in Eq. 3), translating into a reduction of the sensory weight in Bayesian integration and, in turn, an increased central-tendency bias. In contrast, making the secondary task overlap the reproduction phase (Exp. 3) or introducing it after completion of the reproduction (Exp. 1) did not significantly change the central-tendency effect, which corroborates that the Bayesian integration occurred primarily at the encoding stage.

It should be noted that the influence of the memory load on the central tendency was more pronounced in Experiment 2 than in Experiment 4, where, in the latter, the secondary task extended across both the duration-encoding and reproduction phases of the primary task. In the model, this differential load effect was captured by the value of l_s – the scaling parameter of the sensory-measurement uncertainty – being significantly larger in Experiment 2 vs. Experiment 4. This finding raises a question: if the modulation of the central-tendency effect exclusively arises in the encoding stage, shouldn't the two experiments have engendered comparable effects of the memory load on the central-tendency bias? The fact that they didn't might be explained by the order of the primary and secondary tasks. In Experiment 2, the secondary memory task was the first task requiring a response (i.e., the memory test preceded the reproduction), whereas it was the second task in Experiment 4. As can be seen from Figure 2, accuracy in the memory test was significantly higher in Experiment 2 than in Experiment 4, that is, when the memory was probed first rather than second. Within the attentional-sharing framework, this pattern would indicate that more attentional resources were allocated to the first than to the second task. Consequently, allocating relatively less attention to the duration task in Experiment 2 relative to Experiment 4 would have led to an increase of the uncertainty of the duration estimates, rendering a stronger central-tendency effect. In addition, probing the

secondary memory task first also lengthened the gap between the encoding and reproduction stages, which might cause additional forgetting of the estimated duration. Such forgetting might assimilate the representation toward the distal (i.e., long-term) mean prior, as would be suggested by the adaptation-level theory (Helson, 1964). The influence of the prolonged gap between the encoding and reproduction stages was also numerically, though not significantly, evident in Experiment 3 as compared to the baseline Experiment 1 (see Fig. 4a). Unfortunately, because we did not record the completion time of the memory test in between the production and reproduction phases, we cannot quantitatively determine the impact of the prolonged gap on the central tendency bias. Thus, this conjecture deserves further investigation in future studies.

Interestingly, Allred et al., (2016) recently reported in a line-length judgment task that the central tendency is likewise influenced by the memory load. In their study, the memory items had to be held in working memory for the whole process of the primary line-judgment task, which is similar to our Experiment 4. The consistent influences of memory load on the central-tendency effect in non-temporal (Allred et al., 2016) as well as temporal tasks (the present study) suggest that the Bayesian model we propose here is generic, rather than being limited to the time domain. Given that Allred et al.'s study design did not separately manipulate the memory load in the encoding and reproduction stages, their finding does not tell at which stage the interference occurred. Here, with four experiments imposing the memory load in different stages, we found that the impact on the central-tendency effect was primarily attributable to cognitive-load interference during the encoding, rather than the reproduction (retrieval), stage – an attribution that informed the construction of our computational model.

As briefly discussed earlier, when the secondary memory task was imposed on the reproduction phase in Experiment 3, the central tendency was not influenced by the memory load. Interestingly, though, we saw a general shift (bias) in the reproduced duration (Figs. 3a): the higher the memory load, the larger a (general) shift we observed in the reproduced durations – as could also be seen in the shift of the indifference points (Fig. 4b). The dissociation between the central tendency and the general bias mainly came from the differential interference of the memory load in the duration-encoding and reproduction stages. The reproduction stage did not involve any Bayesian integration, just comparing the elapsed time to the estimated duration retained in memory. The primary impact of the memory load consists of the lapse of attention in monitoring the elapsed time (Fortin & Rousseau, 1998; Glasauer et al., 2007), which causes loss of some units of passage time and thus an over-reproduction of the (estimated) duration. This would explain the general shift in the reproduced duration that we observed (captured by

Eq. 7, also Eq. 10). Interestingly, the varying memory loads did not alter the memory representation of the reference (i.e., the estimated) duration, which was derived by Bayesian integration in the encoding stage. However, as considered above, across experiments the reference duration might have been influenced by the temporal gap between the encoding and reproduction stages. The fact that the memory load failed to change the reference duration suggests that the representation of the single reference duration is rather robust. However, this might change if the task requires two or more reference durations to be held in memory – a conjecture that would be interesting to examine in a future study.

It should be noted that the general shift in the reproduced duration was not limited to Experiment 3: we also observed a general shift in Experiment 2, though in the opposite direction (Fig. 4b). As shown in model Eq. 10, unlike the central-tendency effect (captured by w_p) which is only influenced by the memory load in the encoding stage, the indifference point is influenced by both stages, though in opposite directions. When a load was imposed only on the encoding stage, underestimated durations with higher vs. lower loads caused a general downshift in the indifference points, in addition to the influence on the central-tendency bias in the encoding stage. Given the opposite influences of memory load in the encoding and reproduction stages, we observed the opposite trends in Experiments 2 and 3. This could then also explain the absence of significant shifts in Experiment 4, as the net impacts of the general shifts roughly canceled each other out when the memory load was imposed on both stages (in Experiment 4).

In summary, imposing the memory load on the encoding and reproduction stages of duration estimation, we replicated the general underestimation and overestimation of a given duration when the memory load was increased in the encoding and reproduction stages, respectively, as suggested by the attentional-sharing account. In addition, we found the central-tendency bias was only influenced by the memory pressure in the encoding stage. Using a generative Bayesian model, we detailed when and how memory pressure affects time estimation and the concomitant effects on the central-tendency bias, and quantitatively predicted behavioral results from all four experiments. Last but not the least, the generative model we proposed here for the influence of the cognitive load on time perception might be generalizable to other forms of magnitude perception.

ACKNOWLEDGEMENTS

This study was supported by the Starting Research Fund from Hangzhou Normal University (RWSK20190315) to XZ, German Science Foundation (DFG) research grants SH 166/ 3-2 to Z.S and GL 342/3-2 to S.G.

REFERENCES

Acerbi, L., Wolpert, D. M., & Vijayakumar, S. (2012). Internal Representations of Temporal Statistics and Feedback Calibrate Motor-Sensory Interval Timing. In *PLoS Computational Biology* (Vol. 8, Issue 11, p. e1002771). https://doi.org/10.1371/journal.pcbi.1002771

Allred, S. R., Crawford, L. E., Duffy, S., & Smith, J. (2016). Working memory and spatial judgments: Cognitive load increases the central tendency bias. *Psychonomic Bulletin & Review*, *23*(6), 1825–1831.

Bays, P. M. (2015). Spikes not slots: Noise in neural populations limits working memory. *Trends in Cognitive Sciences*, *19*(8), 431–438.

Block, R. A., & Zakay, D. (1997). Prospective and retrospective duration judgments: A meta-analytic review. *Psychonomic Bulletin & Review*, *4*(2), 184–197.

Bürkner, P.-C. (2016). brms: Bayesian regression models using Stan. *R Package Version 0. 10. 0*.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software, Articles*, *76*(1), 1–32.

Cicchini, G. M., Arrighi, R., Cecchetti, L., Giusti, M., & Burr, D. C. (2012). Optimal encoding of interval timing in expert percussionists. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(3), 1056–1060.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Breitkopf u. Härtel.

Fortin, C. (2003). Attentional time-sharing in interval timing. In W. H. Meck (Ed.), *Functional and neural mechanisms of interval timing* (pp. 235–260). CRC Press.

Fortin, C., & Couture, E. (2002). Short-term memory and time estimation: Beyond the 2-second "critical" value. *Canadian Journal of Experimental Psychology*, *56*(2), 120–127.

Fortin, C., & Massé, N. (2000). Expecting a break in time estimation: attentional time-sharing without concurrent processing. *Journal of Experimental Psychology. Human Perception and Performance*, *26*(6), 1788–1796.

Fortin, C., & Rousseau, R. (1998). Interference from short-term memory processing on encoding and reproducing brief durations. *Psychological Research*, *61*(4), 269–276.

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

Glasauer, S., Schneider, E., Grasso, R., & Ivanenko, Y. P. (2007). Space–Time Relativity in Self-Motion Reproduction. *Journal of Neurophysiology*, *97*(1), 451–461.

Glasauer, S., & Shi, Z. (2021a). The origin of Vierordt's law: The experimental protocol matters. *PsyCh Journal*, *pchj.464*. https://doi.org/10.1002/pchj.464

Glasauer, S., & Shi, Z. (2021b). Individual believes about temporal continuity explain perceptual biases. In *bioRxiv* (p. 2021.07.13.452167). https://doi.org/10.1101/2021.07.13.452167

Gu, B.M., Jurkowski, A. J., Shi, Z., & Meck, W. H. (2016). Bayesian Optimization of Interval Timing and Biases in Temporal Memory as a Function of Temporal Context, Feedback, and Dopamine Levels in Young, Aged and Parkinson's Disease Patients. *Timing & Time Perception*, *4*(4), 315–342.

Gu, B. M., & Meck, W. H. (2011). New perspectives on Vierordt's law: memory-mixing in ordinal temporal comparison tasks. In A. Vatakis, A. Esposito, M. Giagkou, F. Cummins, & G. Papadelis (Eds.), *Multidisciplinary Aspects of Time and Time*

Perception (Vol. 6789, pp. 67–78).

Helson, H. (1964). Current trends and issues in adaptation-level theory. *The American Psychologist*, *19*(1), 26–38.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*, *13*(8), 1020–1026.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception ECVP Abstract Supplement*, *36*, 14.

Lejeune, H., & Wearden, J. H. (2009). Vierordt's The Experimental Study of the Time Sense (1868) and its legacy. *The European Journal of Cognitive Psychology*, *21*(6), 941–960.

Maaß, S. C., de Jong, J., van Maanen, L., & van Rijn, H. (2021). Conceptually plausible Bayesian inference in interval timing. *Royal Society Open Science*, *8*(8), 201844.

Macar, F., Grondin, S., & Casini, L. (1994). Controlled attention sharing influences time estimation. *Memory & Cognition*, *22*(6), 673–686.

Matthews, W. J., & Meck, W. H. (2016). Temporal Cognition: Connecting Subjective Time to Perception, Attention, and Memory. *Psychological Bulletin*, *2*. https://doi.org/10.1037/bul0000045

Penney, T. B., Gibbon, J., & Meck, W. H. (2000). Differential Effects of Auditory and Visual Signals on Clock Speed and Temporal Memory. *Journal of Experimental Psychology. Human Perception and Performance*, *26*(6), 1770.

Petzschner, F. H., & Glasauer, S. (2011). Iterative Bayesian estimation as an explanation for range and regression effects: a study on human path integration. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(47), 17220–17229.

Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A Bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences*, *19*(5), 285–293.

Petzschner, F. H., Maier, P., & Glasauer, S. (2012). Combining symbolic cues with sensory input and prior experience in an iterative bayesian framework. *Frontiers in Integrative Neuroscience*, *6*, 58.

Ren, Y., Allenmark, F., Müller, H. J., & Shi, Z. (2021). Variation in the "coefficient of variation": Rethinking the violation of the scalar property in time-duration judgments. *Acta Psychologica*, *214*, 103263.

Roach, N. W., McGraw, P. V., Whitaker, D. J., & Heron, J. (2017). Generalization of prior information for rapid Bayesian time estimation. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(2), 412–417.

Shi, Z., & Burr, D. (2016). Predictive coding of multisensory timing. *Current Opinion in Behavioral Sciences*, *8*, 200–206.

Shi, Z., Church, R. M., & Meck, W. H. (2013). Bayesian optimization of time perception. *Trends in Cognitive Sciences*, *17*(11), 556–564.

Stan Development Team. (2018). *RStan: the R Interface to Stan*.

Stan Development Team. (2020). *Stan Modeling Language Users Guide and Reference Manual. Version 2.28*. https://mc-stan.org

Teghtsoonian, R., & Teghtsoonian, M. (1978). Range and regression effects in magnitude scaling. *Perception & Psychophysics*, *24*(4), 305–314.

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(22), 8780–8785.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Vierordt, K. (1868). *Der Zeitsinn nach Versuchen*. H. Laupp.

Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and

widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research: JMLR*, *11*(12). https://www.jmlr.org/papers/volume11/watanabe10a/watanabe10a

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235.

APPENDIX

Appendix A. Model comparison

The three-stage Bayesian model introduced here assumes logarithmic encoding of subjective durations, based on Fenchner's logarithmic law (Fechner, 1860). Logarithmic encoding implicitly assumes that time percepts follow the scalar property (Gibbon et al., 1984; Shi et al., 2013), namely, a constant of Weber fraction of time estimation. It should be noted that it is not the only model in the field. An alternative assumption holds that the internal representation is linearly scaled, but with the noise increasing linearly with the absolute magnitude according to Weber's law. However, empirical justification of linear vs. logarithmic coding of the internal representation largely depends on the adopted experimental paradigms (Maaß et al., 2021; Matthews & Meck, 2016; Ren et al., 2021). Accordingly, we explored both logarithmic and linear encoding models and compared their predictions to determine which model performs better.

In the logarithmic-encoding model, the duration is first transformed to the logarithmic scale. Bayesian integration of the sensory input and memory prior, and the influences of the memory load operate on this scale. The duration thus estimated is then transformed back to the linear scale for the reproduction. The memory influence occurs at the reproduction stage on the linear scale (see main text for details of the model). In contrast, the linear-encoding model assumes that all processes operate at the linear scale. However, the model further assumes Weber scaling, that is: the sensory measure (S) of given sample duration (D) follows Weber' law: $S = D + \epsilon$, where ϵ indicates internal measurement noise. The standard deviation of sensory measurement (σ_s , estimated from the noise ϵ) is approximately proportional to the mean of the sensory measurement (μ_s), $\sigma_s = k * \mu_s$, where k is known as the Weber fraction of sensory measurements. Similar to the logarithmic-encoding model, both the mean estimate and its standard deviation are assumed to be linearly affected by the memory load. Let the memory representation without loss of clock ticks be normally distributed, $S_{wm} \sim N(\mu_{wm}, \sigma_{wm}^2)$. Accordingly, both the mean μ_{wm} and the variance σ_{wm}^2 are subject to the influence of the memory load:

$$\mu_{wm} = D \cdot (1 - k_s \cdot M) \quad \text{(Eq. 10)}$$

$$\sigma_{wm}^2 = \sigma_s^2 (1 + l_s \cdot M) \quad \text{(Eq. 11)}$$

At the Bayesian integration stage, the distribution of the internal prior is assumed as

$N(\mu_p, \sigma_p^2)$ with the mean μ_p and the variance σ_p^2, and the prior is integrated with the memory representation of duration S_{wm} according to the Bayesian rule, so that Eq. 4 and Eq. 5 are also applicable in the linear-scale model. During the reproduction process, time units could be lost due to attentional sharing and reproduction inherited additional motor noise. Let $D_r \sim N(\mu_r, \sigma_r^2)$ be the reproduction distribution with:

$$\mu_r = \mu'_{post}\,(1 + k_r \cdot M) \quad \text{(Eq. 12)}$$

$$\sigma_r{}^2 = \sigma^2{}_{post} + \sigma_{mn}^2 \quad \text{(Eq. 13)}$$

The notations of these key parameters in the linear model are the same as those in the logarithmic model.

Both the logarithmic- and the linear-scale model perform well in predicting the mean reproduction and the coefficient of variation (CV). However, while the predictions are comparable as regards the mean reproduction, the logarithmic model predicted the CV significantly better than the linear model. Figure A1 represents the mean absolute errors (MAE) of the predictions derived from the linear- and, respectively, logarithmic-scale models (reproduction and CV): each point represents the mean absolute prediction error in the reproduction means and their CVs in the various conditions across the four experiments. As can be seen, the logarithmic model produced generally smaller prediction errors than the linear model, with a particularly marked advantage in the CVs. The prediction errors of the logarithmic model never exceed 0.0155, indicated by the dashed line; but the linear model performed worse for more than half the conditions compared to the poorest condition from the logarithmic model.

To formally evaluate the models' performance, we calculated Watanabe–Akaike information criterion (WAIC) and leave-one-out cross-validation (LOO-CV) as predictive information criteria for Bayesian models, using the Loo package in R framework (Vehtari et al., 2017; Watanabe & Opper, 2010). Lower values of WAIC and LOO-CV imply higher prediction accuracy. Table A1 lists the averaged WAICs, LOO-CVs, and prediction accuracies (AUC) of the reproduction means and variances across all participants. As can be seen, the logarithmic-scale model was associated with lower WAIC and LOO-CV values and higher prediction accuracies across all experiments than the linear model.

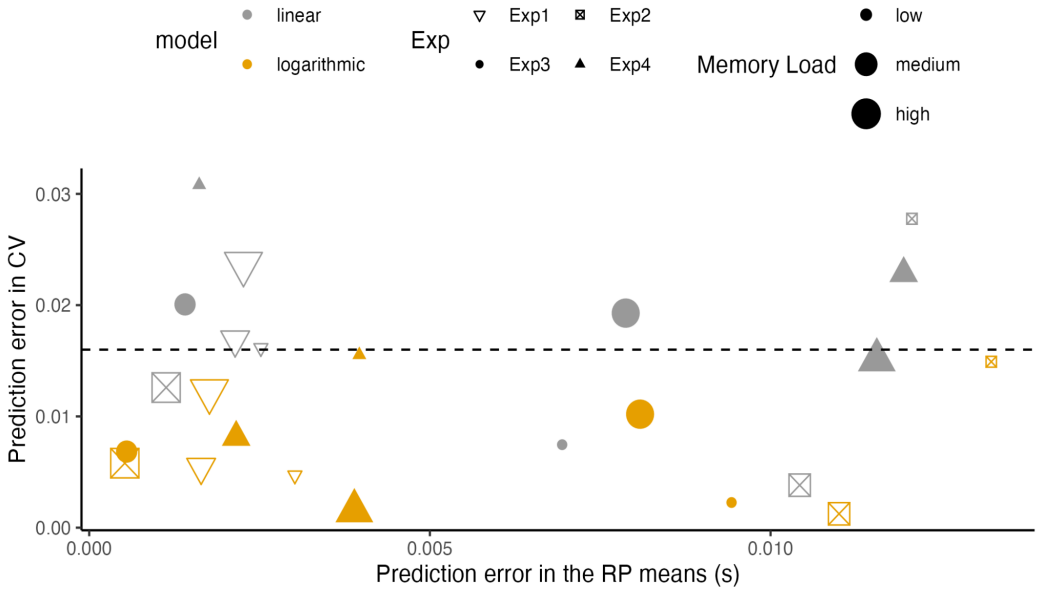

Figure A1 *Averaged absolute prediction errors of the reproduction and CV derived from the proposed logarithmic- and linear-scale models for individual observers in the four experiments.*

Table A1 *WAIC and WBIC as predictive information criteria for Bayesian models.*

Model		**Exp.1**	**Exp.2**	**Exp. 3**	**Exp. 4**
logarithmic model	**WAIC**	123.244	557.315	420.216	415.359
	LOO-CV	368.069	557.315	420.216	415.359
	Reproduction AUC	96.664%	95.828%	96.552%	96.470%
	CV AUC	82.385%	86.228%	84.151%	81.594%
linear model	**WAIC**	142.914	318.033	190.389	193.002
	LOO-CV	415.902	590.852	463.886	465.701
	Reproduction AUC	96.422%	95.632%	95.932%	95.932%
	CV AUC	74.588%	81.389%	77.578%	71.696%

2.3 Prior integration in Bayesian estimation under multi-prior context

CONTRIBUTIONS

Zhu and Zang share first authorship. ZS and Zhu conceived and designed the experiments. Zhu and Zang collected and analyzed the data, Zhu implemented the models. Zhu, Zang and FA discussed the results with ZS, SG and HJM. Zhu and Zang interpreted the results and wrote the paper. ZS, FA, SG and HRL commented and revised the manuscript.

Priors Integration in Bayesian Estimation for Time Interval Reproduction Task

XIUNA ZHU[1#], XUELIAN ZANG[2#], FREDRIK ALLENMARK[1], STEFAN GLASAUER[3], HERMANN J. MÜLLER[1,4], ZHUANGHUA SHI[1,*]

1. Experimental Psychology, Department of Psychology, LMU Munich, Germany
2. Center for Cognition and Brain Disorders, Institutes of Psychological Sciences, Hangzhou Normal University, China
3. Institut für Medizintechnologie, Brandenburgische Technische Universität Cottbus-Senftenberg, Germany
4. Department of Psychological Science, Birkbeck College (University of London), London, UK

# These authors contributed equally to the paper

ABSTRACT

To improve the reliability of sensory estimation, our humans use acquired contextual knowledge compensating uncertainty during perception. Ample previous studies suggested that multiple timing mechanisms exist in the brain across and within sensory modalities, but how sensory information integrates with multiple priors that are shaped by distinct context. To figure out how the hierarchical Bayesian estimation structure works in multiple prior temporal estimation, two duration reproduction experiments with two levels of variance of short and long range on logarithmic scale were carried out to uncover the underlying mechanism during multiple prior knowledge integration. The participants were firstly trained with the short and long intervals session-wise, and then tested with interleaved ranges randomly inter-mixed, to examine potential integration hierarchies. In the present study, three hierarchical structures of prior integration in Bayesian estimation were proposed to uncover the governing rules. The model results showed their performance were equally well performance in reproduction mean and variance prediction despite their different structural assumptions. In the temporal context with smaller variability, time estimates counted more on the global prior knowledge and less on sensory measurement. Besides, session order impacts on the evaluation of global prior, since context farther in the past is less weighted than more recent context. When the short session was taken as the first training session, observers acquired a higher mean of global prior.

Keywords: time perception, Bayesian modeling, contextual effect, temporal reproduction, prior integration

INTRODUCTION

Timing and time perception plays a vital role in our daily life, from motor planning to social interactions. However, temporal estimates in everyday activities are not always accurate. A variety of biases in time perception (i.e., the perceived time deviates from the physical time) comes from not only inaccurate sensory timing but also from surrounding contexts. A typical contextual bias is that the perceived time for the same duration would be different when the duration is embedded in different temporal ranges. This phenomenon has been referred to as the *range or regression effect (Petzschner & Glasauer, 2011; Teghtsoonian & Teghtsoonian, 1978)*.

In recent decades, with the widespread usage of Bayesian reference, various Bayesian observer models have been proposed to probabilistically interpret various types of contextual biases in magnitude estimates (Berniker et al., 2010; Knill & Richards, 1996; Ryan, 2011; Wei & Stocker, 2015; Zimmermann & Cicchini, 2020). The key assumption of those Bayesian observer models is that human perception is an optimal integration process, combining the sensory measurement (referred to as *likelihood*) and the contextual knowledge (e.g., the mean and variations of the previously encountered stimuli, referred to as *prior*) together (Fritsche et al., 2020; Jazayeri & Shadlen, 2010; Roach et al., 2017a; Shi, Church, et al., 2013; Westheimer, 2008). One key benefit of this integration is to improve the reliability of the final estimate to cope with uncertainty (Knill & Richards, 1996; von Helmholtz, 1867; Westheimer, 2008). For example, when a truck in front of your car blocks the traffic light at a crossroad, you may still be able to estimate the timing of the green light by your past experience. The learned contextual knowledge about the past stimuli encountered may influence further timing process as they have registered in the memory (Glasauer et al., 2020; Jazayeri & Shadlen, 2010; Nagai et al., 2012; Roach et al., 2017a). One natural outcome of the integration process is that the final estimate is often assimilates to the prior knowledge, which is a typical central tendency effect[4] (Hollingworth, 1910).

One classical paradigm to observe a typical central tendency effect is the duration production-reproduction paradigm, which can be traced back to early Vierordt's effect (Vierordt, 1868). Vierordt asked his assistant to arbitrarily produce a duration and he then

[4] Central tendency effects were confirmed by later researchers as a common phenomenon in time perception (Hollingworth, 1910), i.e., the reproductions of short durations are overestimated, meanwhile the reproductions of long durations are underestimated.

reproduced it. Nowadays we use a computer to automatically generate a given interval and participants to reproduce it. It has been repeatedly demonstrated that when intervals are randomly sampled from a single range, observers are relatively accurate to estimate the sampled range and distribution with a single prior approximation (Acerbi et al., 2012). When intervals from different ranges are presented blockwise and demarcated by distinct cue or spatial location, observers are able to generalize temporal context experiences and formed separate priors for individual ranges, hence a central tendency effect was observed in each range (Cicchini et al., 2012; Gu et al., 2016; Jazayeri & Shadlen, 2010; Petzschner et al., 2012b). It should be noted that most previous studies were interested in the central tendency effect, assuming simple Baysian integration with one single prior.

Distributed timing system and multiple priors of temporal contexts

Regarding the timing system, it remains controversial whether we pose a single central dedicated clock or distributed multi-clock systems (Buhusi & Meck, 2009). There is evidence that different modalities may use different timing systems (Buhusi & Meck, 2009; Ivry & Richardson, 2002; Paton & Buonomano, 2018). Ample evidence suggests that multiple timing mechanisms exist in the brain across and within sensory modalities (Gau & Noppeney, 2016; Heron et al., 2012). However, up to date it is not clear if the prior representation of multimodal timing is also distributed for different sensory modalities. The question is important given that the prior representation, whether modality-specific or modality-independent, could help us understand how temporal context is represented in our brain.

Recent studies have shown that the prior representation might not be like the distributed clock systems, each having their own priors. For instance, a recent study by Roach et al. (2017a) showed that spatial separation or modality of two different ranges of durations for the duration reproduction task did not yield separate priors for the separate locations or modalities. Separate priors were only formed when two tasks were distinct (e.g., reproduction vs. bisection) for two separate durations ranges. More recently, Zang et al. (2022) illustrated that multiple separated priors could only be developed when two ranges (short and long) were clearly separable in the range. When two ranges were overlapped, maintaining and updating two similar priors could be costly, even though two ranges of durations were modality-specific. Their finding is partially consistent with the results of Roach et al.'s study (2017b) that observers formed a unified prior (global prior) by generalizing across the two interleaved stimulus sets instead of separate priors for each stimulus set in multi-prior context, and they did not distinguish the different stimulus

sets in the random interleaving condition even if the stimuli used for the two interleaved duration distributions were clearly discriminable. In addition to the context of sampled distribution of durations, other contexts may also influence duration judgments. For example, auditory duration is often perceived longer than the visual duration (Shi, Ganzenmüller, et al., 2013; Wearden et al., 1998). Binary categorical cue (Petzschner et al., 2012a) or categorical judgments (Luu & Stocker, 2018) may also influence final decision making. For instance, Petzschner et al.(2012b) used overlapping short and long distance in a production-reproduction task with three different experimental conditions ("blocked-ranges, no cue", "interleaved range, no cue", and "interleaved range, cue") and demonstrated two separate priors in the "interleaved range, cue" condition due to the presence of the categorical cue ('Short' vs. 'Long').

Bayesian model in multi-prior context

It remains an open question how those different contexts (priors) integrate together with the sensory measurement. Most current Bayesian approaches (Cicchini et al., 2012; Petzschner et al., 2015; Roach et al., 2017a) have been mainly aimed to explain subjects' performance by incorporating prior experience into the time estimation process with an implicit assumption of uni-prior temporal context without considering the interaction among different contexts. To obtain Bayesian optimality in a multi-prior context, we need to investigate the internal structure of prior integration in the multi-prior context. Would it be a flat integration (i.e., combining priors and sensory measurement linearly) or combined priors and sensory measurement in a hierarchical order? If the latter is true, what kind of structure order would be? Suppose the internal prior consists of a multi-level structured prior, constituting something like a local-global hybrid prior or a generalization of the global prior and local prior information, the question of how Bayesian optimization of time perception for the temporal context with multiple priors to explain subjects' performance by the incorporation of multiple prior experience into the time estimation process remains unknown. The internal prior may consist of a multi-level structured prior, constituting something like a local-global hybrid prior. On this ground, we investigated the structure of the prior in detail to attain a better understanding of the hierarchical structure of the prior. In any case, fundamental questions remain regarding how these acquired knowledge (priors) are integrated with global priors and generalize to sensory evidence and behavioral contexts. Inspired by this, we intend to further investigate the underlying structure of the context-specific prior in time estimation of prior information in the

duration reproduction task. Furthermore, the interactive influence between priors in the multi-prior context can be posed rigorously.

This study

In this study, we employed two with two levels of variance of short and long range on logarithmic scale by rescaling short and long ranges and expanding the separation between the short and long intervals to uncover the underlying mechanism during prior knowledge integration. Note that the geometric means of the interleaved range in two experiments were kept to be the equal. As no single Bayesian observer model yet has been proposed to be able to explain prior integration in multi-prior contexts, we aimed to extend our basic Bayesian observer model to multi-prior temporal contexts, and further uncover the rules governing how learned prior knowledges are grouped together. We proposed three possible hierarchical Bayesian models to explore hierarchical structure of the priors integration under multi-prior context. All proposed hierarchical Bayesian models were able to predict the central tendency effects observed from two experiments. The model results showed their performance were equally well performance in reproduction mean and variance prediction despite their different structural assumptions. These proposed models were strong evidence showing the existence of an additional global prior which represents the context of interleaved range for multi-prior context. Additionally, we found the global prior was not only influenced by the geometric mean of the interleaved range, but also influenced by more recent context (prior) learned in the session before a randomly interleaved session. In addition, the global prior was weighted more in time estimates with the less variability of the interleaved range. When the global range and variability were reduced, the influence of the global prior increased accordingly.

METHODS

Participants

Thirty-two volunteers recruited from Ludwig-Maximilians-Universität Munich took part in two experiments (16 each; Exp1. : 13 females; mean age: 25.5 yrs; SD of age: 4.32 yrs; Exp.2: 5 females; mean age: 24.8 yrs; SD of age: 2.28 yrs). All had normal or corrected-to-normal visual acuity. All participants were given written informed consent before the experiments and were paid 9 Euro per hour for their participation. The study was approved by the ethics committee of the Psychology Department of LMU Munich. All participants were

naive to the purpose of the research. The number of participants was determined according to a power analysis conducted with G*Power (Erdfelder et al., 1996) with an alpha of .05, power of 95% and a relative large size effect (f(U) = 0.426) in a repeated-measures analysis of variance (ANOVA; η_p^2= 0.4), the effect size was determined according to reported data in previous studies (Cicchini et al., 2012; Petzschner et al., 2012b; e.g., Roach et al., 2017a).

Stimuli and apparatus

The experiments were carried out in a sound-attenuated and dimly lit experimental cabin. Visual stimuli (a gray disk, 36.5 cd/m^2, 5°of visual angle in diameter) and experiment instructions / feedback, were generated with Matlab software (The MathWorks Inc, 2016 version) and Psychophysics Toolbox-3 (Kleiner et al., 2007), and were presented on a 21-inch LACIE CRT monitor with a monitor resolution of 1024×768 pixels and a refresh rate of 100 Hz. The viewing distance was fixed at around 57 cm. The display background was dark gray (18.9 cd/m^2).

Procedure and Design

We adopted the temporal production and reproduction task in two experiments. The task remained the same for all experiments. Participants were asked to reproduce the duration of a given visual stimulus as accurately as possible by pressing and holding the left button of the mouse (see Figure 1 for an example). Each trial started with a visual fixation cross for a random interval between 500 ms to 1s, prompting participants to get ready for the new trial. A production phase started immediately after the offset of the fixation display, with the visual stimulus (i.e., gray disk, 5°) being presented for a randomly selected interval either on the left (-5°) or the right (5°) side of the fixation. The test durations were sampled from two ranges: the short-range (< 1s) and the long-range (> 1s, details see Figure 1 and the text of the next paragraph). The ranges were assigned to the left or the right, fixed for a given participant, but counter-balanced among participants. After the presentation of the probe duration, there was a blank screen with 1s. Then, a white down arrow was shown, prompting participants to reproduce the probe duration by pressing and holding the left button of the mouse as accurately as possible. After the mouse was pressed, a gray circle was immediately presented in the middle of the screen and it disappeared when the mouse was leased, to help participants perceive their reproduced duration. A feedback display was then shown to indicate the relative reproduction accuracy [(reproduced duration - target duration)/ target duration]. The feedback display

consisted of five circles, corresponding to the relative deviation: below -30%, [-30%,-5%], (-5%, 5%), [5%, 30%], and above 30% from the left to the right respectively. The feedback for the 'accurate' trials falling in between [-30%, 30%] was presented in green color for 0.5s, while the feedback for the 'inaccurate' trials (i.e., outside [-30%, 30%]) was shown in red (two outer circles) for 1.5s as a warning. After an inter-trial interval of 1s, the next trial began.

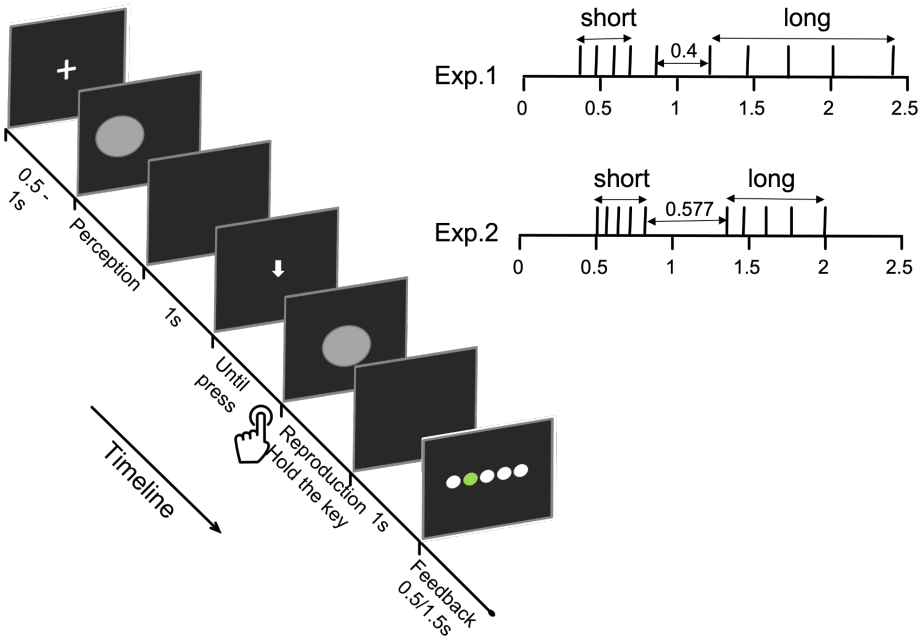

Figure 1 *An illustration of trial procedure and stimuli in Experiments 1 and 2. The target intervals were shown in the left/right side of the screen according to the range that the interval came from (either short or long range). Participants were required to reproduce the time duration of the target stimulus by pressing the mouse button as accurately as possible for an equivalent interval of target intervals.*

To fully examine the context effects due to the blockwise ranges, we adopted two levels of variance on log scale by expanding or rescaling the range of short and long intervals and varied the separation between the short and long ranges: higher log variance (0.634 in Exp. 1) and lower log variance (0.538 in Exp. 2), as depicted in Figure 1. The intervals of the short range and long range were not overlapped and the geometric means of interleaved range for Exp. 1 and Exp. 2 were the same. In Exps. 1 and 2, participants performed three experimental sessions: the short session (five interval samples with equal logarithmic spacing were chosen from short-range, 400, 476, 566, 673, 800 ms for Exp. 1, 490, 543, 600, 663, 730 ms for Exp. 2), the long session (five interval samples with equal logarithmic spacing were chosen from long-range, 1200, 1427, 1697, 2018, 2400 ms for Exp. 1, 1310, 1448, 1600, 1768, 1954 ms for

Exp. 2), and the mixed session (10 intervals from both short and long ranges were randomly mixed together). Each duration was repeated 32 times in the short and long session and 16 times in the mixed session. In Exp. 1, the intervals from the short-range were equally spaced in the log-scale between 400 and 800 ms, with a geometric mean of 566 ms, a standard deviation of 158.459ms; the intervals from long-range were also equally spaced in the log-scale between 1200 and 2400 ms, with a geometric mean of 1697ms and a standard deviation of 475.407ms; distribution of intervals for the mixed session had a geometric mean of 980ms and a standard deviation of 699.307ms. In Exp. 2, the intervals from the short-range were equally spaced in the log-scale between 490 and 733 ms, with a geometric mean of 600 ms and a standard deviation of 95.571ms; the intervals from long-range were equally spaced in the log-scale between 1310 and 1954 ms, with a geometric mean of 1600 ms and a standard deviation of 254.864 ms; distribution of intervals for the mixed session had a geometric mean of 980 ms and a standard deviation on of 562.446ms. Note that the geometric means of interleaved range for Exps. 1 and 2 were kept the same (980 ms).

In order to form separate priors for blocked ranges, we adopted distinct spatial location for stimulus sets from short and long ranges, similar as in Roach et al.'s study (2017b)[5]. Henceforth the short session and long session were referred to as the blocked range (BR) conditions, whereas the mixed session was referred to as the interleaved range (IR) condition. Both experiments start with BR sessions first (the sequence of the short and long sessions were randomly and counterbalanced designed). The IR session in which test intervals are sampled from short and long range was always the last session. Each session has 8 blocks and each block has 20 trials (in total 160-trial per session). Prior to the 24-block formal experiment, participants got familiarized with the task in a practice block of 20 trials, which involved the same procedure as the formal experiment.

Data Analysis

To exclude duration reproduction likely reflecting lapses of attention or accidental responses, outlier criterion was adopted based on the interquartile range (IQR): production bias outside of the 0.025 and 0.975 quantiles for each participant in each experimental condition were omitted from the further data analyses. Statistical differences were assessed by repeated-

[5] The visual stimuli were presented at different spatial locations (left or right of fixation cross) in short and long sessions to build a stable prior based on spatial and temporal information, and alternated from trial to trial in the mixed session to remove any spatial and temporal uncertainty.

measures Bayesian analyses of variance (ANOVA), and we applied one-tailed paired t-tests when two levels were compared. We additionally report the estimated Bayes factors(BF_{10}), which gives the evidence required for acceptance of the null hypothesis, obtained from comparable Bayesian statistics using JASP v.0.13.1 (*http://www.jasp-stats.org*) with default settings (Love et al., 2019). Further repeated-measures ANOVA analysis with conditions (IR vs. BR) and ranges (short vs. long) as within-subject factors to assess participants' response difference between condition and range.

BEHAVIOR RESULTS

Mean Reproduction and Variability

The mean reproduction responses and reproduction bias in Exps. 1 and 2 are depicted in Figure 2 (Exp. 1: a and b, Exp. 2: d and e). By visual inspection, we observed remarkable central tendency effects in all conditions. In addition, comparing two types of sessions: the IR session had a general shift as compared to the BR session. The short range durations were in general overestimated and the long range duration underestimated in IR relative to BR session.

A 2×2 repeated-measures ANOVA on the reproduction bias in Exp. 1 revealed a significant main effect of Range (short vs. long) [$F(1, 15) = 12.863$, $p < 0.01$, $\eta_p^2 = .462$, $BF_{10} > 1000$], together with a significant interaction of Range (short vs. long) × Prior Context (BR vs. IR) [$F(1, 15) = 1.09$, $p < .01$, $\eta_p^2 = .424$, $BF_{10} = 5.50$], main effect of Prior Context was not significant [$F(1, 15) < .019$, $p = .892$, $\eta_p^2 = .0001$, $BF_{10} = .358$]. A post-hoc one tailed paired t-test comparing the reproduction difference between the BR and IR sessions for the short and long range separately revealed significant difference between the BR and IR condition [for the short range: reproduction bias for intervals in the BR condition was less than IR condition, $t(15) = -4.943$, $p < .0001$, $BF_{10} = 345.098$; for the long range: reproduction bias in the BR condition was greater than the IR condition, $t(15) = 2.073$, $p = 0.028$, $BF_{10} = 2.763$], suggesting a stronger overestimation for the short intervals and underestimation for the long intervals in the IR relative to the BR session.

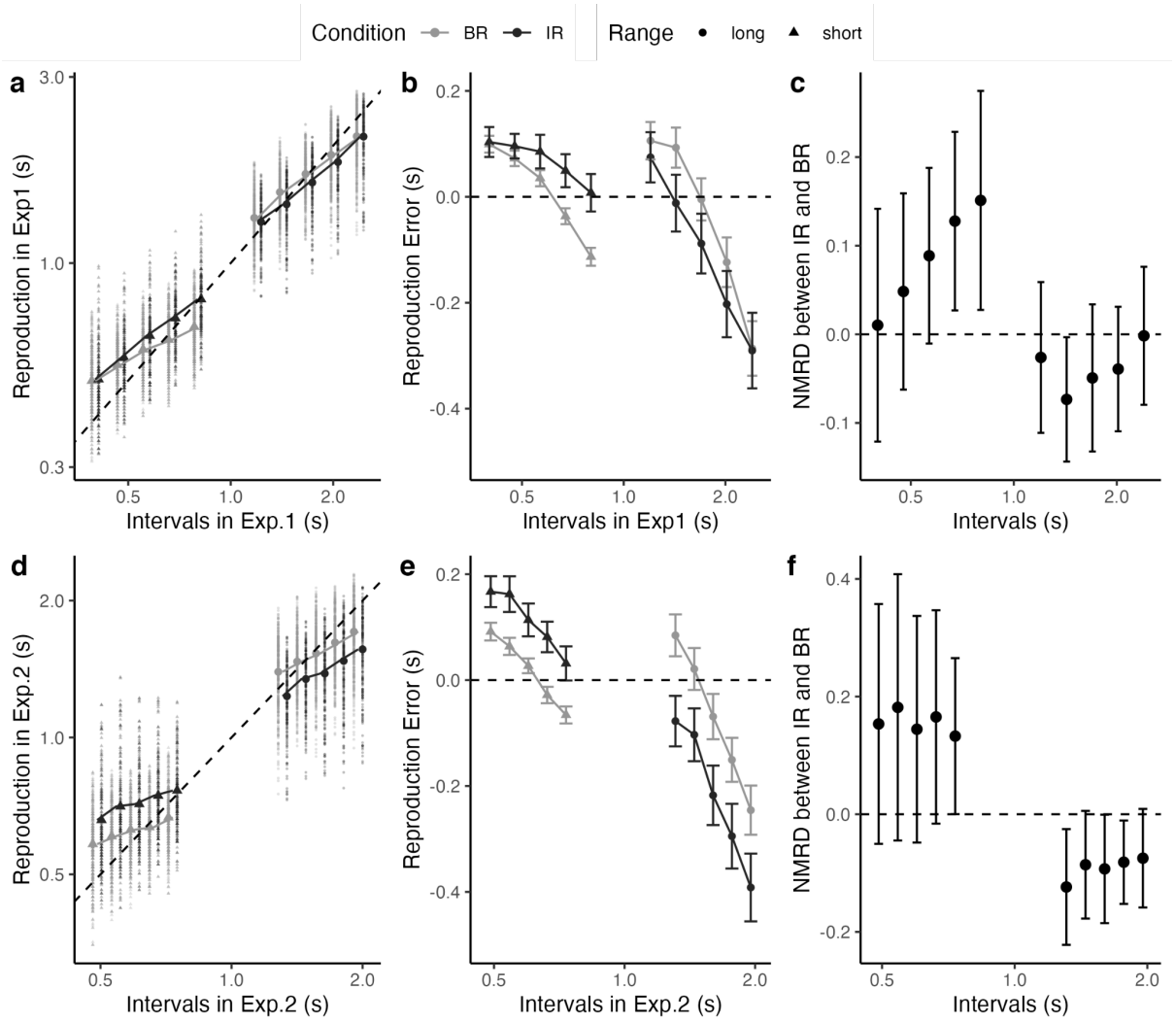

Figure 2 *Production-reproduction performance in Exp. 1 (small separate 400ms) and Exp. 2 (large separate 577ms). (1) Mean reproduction of estimation behavior across subjects was plotted against test intervals, separate for the BR vs. IR conditions in Exp.1(a) and Exp.2(d). The dashed diagonal lines indicate veridical performance (the responses and sample stimulus would be equal). The scatterplots indicate the reproduced duration of subjects for target intervals. (2). Mean of reproduction bias in Exp.1(b) and Exp.2(e). The bias increased for increasing sample durations because of scalar property. Error bars denote standard errors of the mean across participants. (3). Normalized mean reproduction difference between IR condition and BR condition in Exp. 1(c) and Exp.2(f). Note that the x axis in subplots was transformed to log10 scale.*

Moreover, a 2×2 repeated-measures ANOVA on the reproduction bias in Exp. 2 yielded significant main effects of Range (short vs. long), $F(1, 15) = 42.183$, $p < .0001$, $\eta_p^2 = .738$, $BF_{10} > 1000$, together with a highly significant interaction of Prior Context × Range, $F(1, 15)$

= 26.494, p <.001, η_p^2 = .638, BF_{10} >1000, but not significant for the change Prior Context [*F(1, 15) = 2.452, p = 0.15, η_p^2 = .138, BF_{10} = .358*]. A post-hoc ANOVA on the reproduction bias revealed a significant effect of the Prior Context (BR vs. IR) in short range [*F(1, 15) =12.58, p < .01, η_p^2 = .289*] and no significant effect of Prior Context in long range [*F(1, 15) = 22.88, p < .001, η_p^2 = .33*]. Similarly, one tailed paired t-test comparing the reproduction difference between the BR and IR conditions for the short range and long range separately was conducted on Exp. 2 and the results showed that the reproduction bias for intervals in short range in the BR condition was less than IR condition [*t(15) = -3.55, p <.001, BF_{10} = 30.55*]; reproduction bias for intervals in long range in BR condition was greater than IR condition [*t(15) = 4.748, p = .0001, BF_{10} = 262.83*], implying a stronger overestimation for short intervals and underestimation for long intervals in IR condition than BR condition in Exp. 2.

Mean reproduction difference between prior context condition

The mean observed reproduction difference between the IR and BR conditions in Exp. 1 and Exp. 2 were illustrated in Figure 2 **(c** and **f)**. Normalized Mean Reproduction Difference (NMRD) was used to scale the results of the mean reproduction bias difference between IR condition and BR condition, making Exps. 1 and 2 comparable. NMRBD was calculated as $NMRD_i = \frac{R_{i,IR} - R_{i,BR}}{D_i}$, where $R_{i,IR}$ and $R_{i,BR}$ represent the mean of the reproduction of a given interval D_i in the IR and BR conditions. A mixed-design ANOVA with Rang (short and and) as the within-subject factor, Experiment (1 and 2) as the between-subject factor revealed the the main effect Rang to be significant [*F(1, 30) = 38.047, p < .001, η_p^2 = .445*], while the main effect of Experiment was non-significant [*F(2, 30) = .596, p = .5575, η_p^2 = .0014*] and the Exp × Rang interaction was non-significant [*F(2, 30) = 2.0813, p = .142, η_p^2 = .0.08*]. In addition, we conducted one-tailed paired t-test comparing the reproduction difference between the BR and IR conditions $\left(R_{i,IR} - R_{i,BR}\right)$ for the short and long range separately, and these analysis revealed that NMRD in Exp. 1 was less than in Exp. 2 [*t(30) = -1.449, p =.079, BF_{10} = 1.487*] for short range, NMRD for long range in Exp. 1 was significant greater than in Exp. 2 [*t(30) = 2.140, p < .05, BF_{10} = 3.639*], suggesting less reproduction difference in IR and BR condition in Exp. 1, namely that in Exp.1 there was less biased towards the grand mean compared to Exp. 2 for both the short and long intervals.

Mean reproduction difference between session order

Exp. 1 and 2 start with the short session or long session, and end with IR condition. The order of the short and long sessions was randomly and counterbalanced. To check the session order effect on the behavior, we plotted the mean observed reproduction as a function of the test interval, separate for the session, and Session order [Short-Long-Mixed (S-L-M) vs. Long-Short-Mixed (L-S-M)] in Exp. 1 and Exp. 2 (Figure 3). By the visual inspection, the Session order did not influence much for the long range duration reproduction, but showed visible separation for the short range duration reproduction. Specifically, durations in the short IR session were shifted higher when it was preceded by the long session as compared to the short session.

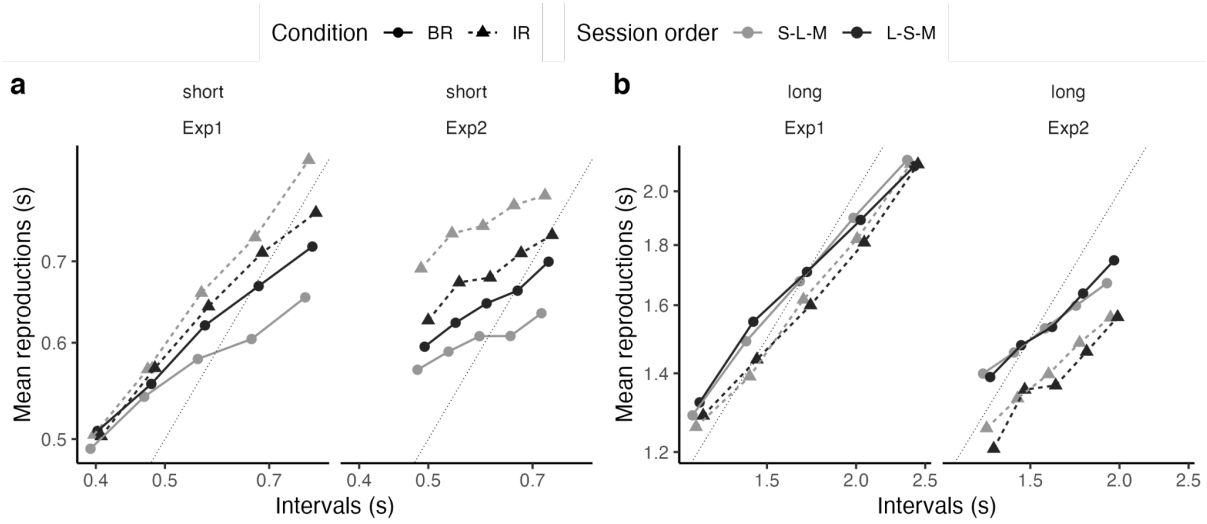

Figure 3 *Mean observed reproduction across subjects was plotted against test intervals, separate for Range (short (**a**) vs. long (**b**)), Prior Context (IR vs. BR) and Session order (S-L-M vs. L-S-M) in Exp.1 and Exp.2. The dashed diagonal lines indicate veridical performance (the responses and sample stimulus would be equal).*

A mixed-design ANOVA of the mean reproduction bias with Experiment (1–2) and Session Order as between-subject factors, Prior Context (BR vs. IR) and Range (short vs. long) as within-subject factors revealed a significant main effects of Range [$F(1, 28) = 47.89$, $p < .001$, $\eta_g^2 = .631$], Range × Prior Context interaction [$F(1, 28) = 36.329$, $p < .001$, $\eta_g^2 = .565$], and Session Order × Prior Context interaction [$F(1, 28) = 8.85$, $p < .01$, $\eta g^2 = .011$]. The rests were not significant ($Fs < 4.2$, $p > .052$). The interaction between the Session order and Prior context mainly originated from the short range (see Figure 3a). A long session prior to the mixed session inflated the reproduction in the short range, but not in the long range. Of note,

in BR condition, for Session Order S-L-M (short session as the first session of the experiment), there is no session prior to short session, so the reproduction of the short session has no general shift. In contrast, the inflated reproduction for Session Order L-S-M was observed compared to S-L-M (see Figure 3a), suggesting the reproduction of the short range in BR condition was influenced by long session prior to short session. The effect of Session Order on the IR condition was opposite to the effect on the BR condition. In the IR condition, the reproductions of the short and long-range were both influenced by the long session prior to the mixed session for session order S-L-M, and both influenced by the short session prior to the mixed session for session order L-S-M. Hence, we observed more inflated reproduction of S-L-M than L-S-M for short range (see Figure 3a), and more deflated reproduction of L-S-M than M-S-L for long range in IR condition (see Figure 3b). Combining the results of the BR and IR conditions, we concluded that the reproduction of later sessions is influenced by the previous session, that is, when the session prior to the current session is a long session, the reproduction would be increased, at least for the short range, and when the session prior to the current session is a short session, the reproduction would be reduced.

Regression indices of central tendency

To better understand the central tendency effect, we use the indicator - Centrality index(CI), calculated as CI = 1-slope (slope denoting the slope of linear regression function between reproduced intervals and produced intervals), which has been used in previous studies (Glasauer & Shi, 2019). In Experiment 1, a 2×2 repeated-measures ANOVA on mean CI with the factors Prior Context (BR vs. IR) and *Range* (short vs. long) revealed a significant main effect of Prior Context, $F(1, 15) = 24.18$, $p < .001$, $\eta_p^2 = .617$, $BF_{10} = 80.25$, together with a significant interaction between Prior Context and Range, $F(1, 15) = 10.973$, $p < .005$, $\eta_p^2 = .422$, $BF_{10} = 20.287$. Moreover, the reproduction of IR condition in Exp. 1 had a lower CI than it in the BR condition (see Figure 2c), suggesting the IR condition showed lower central tendency effects than the BR condition. More specifically, there was significant effect between IR and BR condition for the short range [mean: 0.244 (IR) vs. 0.541 (BR), $F(1,15) = 18.5$, $\eta_p^2 = .552$, $BF_{10} = 220.238$], implying there were significant less central tendency effects in the IR condition; but there was no significant difference between IR and BR for long range[mean: 0.306 (IR) vs. 0.340 (BR), $F(1,15) = 1.709$, $p = 0.211$, $\eta_p^2 = .102$, $BF_{10} = 0.617$].

For Exp. 2, with a 2×2 repeated-measures ANOVA on mean CI with the factors Prior Context (BR vs. IR) and Range (short vs. long) we found no significant difference of the factors

Prior Context [*F(1, 15) =1.177, p =0.295,* η_p^2 *= .073,* BF_{10}*=.325*] and no significant Prior Context ×Range interaction [*F(1, 15) = .870, p =.366,* η_p^2 *= .055,* BF_{10}*= .368*]. The CI in the Exp. 2 (as depicted in Figure 4) were analyzed by one tailed paired t-test with factor Prior Context, which failed to find any significant differences: the short range [mean: 0.468 (IR) vs. 0.426 (BR), *t(15) = 1.323, p = .103,* BF_{10}*= 1.07*]; and the long range [mean: 0.608(IR) vs. 0.7625(BR) , *t(15) = .115, p = .455,* BF_{10}*= 0.514*].

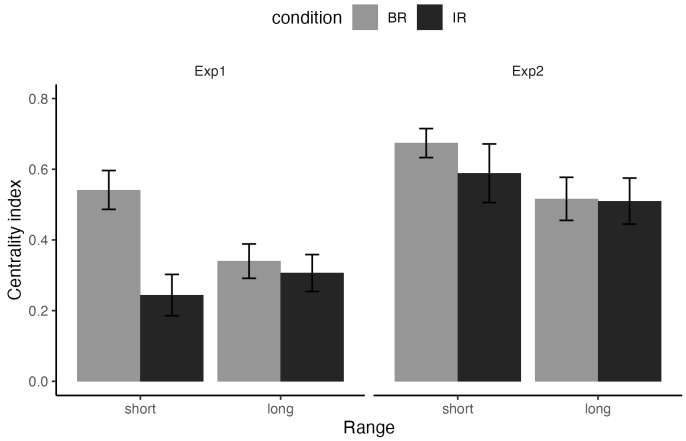

Figure 4 *Mean regression index of central tendency (CI) as a function of target duration of the short and long interval conditions in Exp. 1 and Exp.2.*

Figure 4 illustrates the centrality index (CI) of Exp. 1 and Exp. 2 among different prior-condition-associated and range-associated conditions. The CI of the least-squares regression in Exp. 2 were significantly greater than that in Exp. 1 (see Figure 4), suggesting that the subjects participating in the experiment with narrower local ranges and a wider separation (Exp. 2) between short and long intervals showed higher central tendency effects.

Indifference Points (IPs)

Indifference point denotes the mean of the prior that production times were biased towards. To investigate the effect of the migration toward the center varying in Prior Context and Range, we analyzed indifference points to denote the center of compressive central tendency bias, as shown in Figure 5. In the BR condition, the mean IPs were located close to the means of the stimulus sets in the corresponding range, indicating that the local prior in the BR session worked independently and indifference points induced by the local prior were located around the center of the sampled distribution.

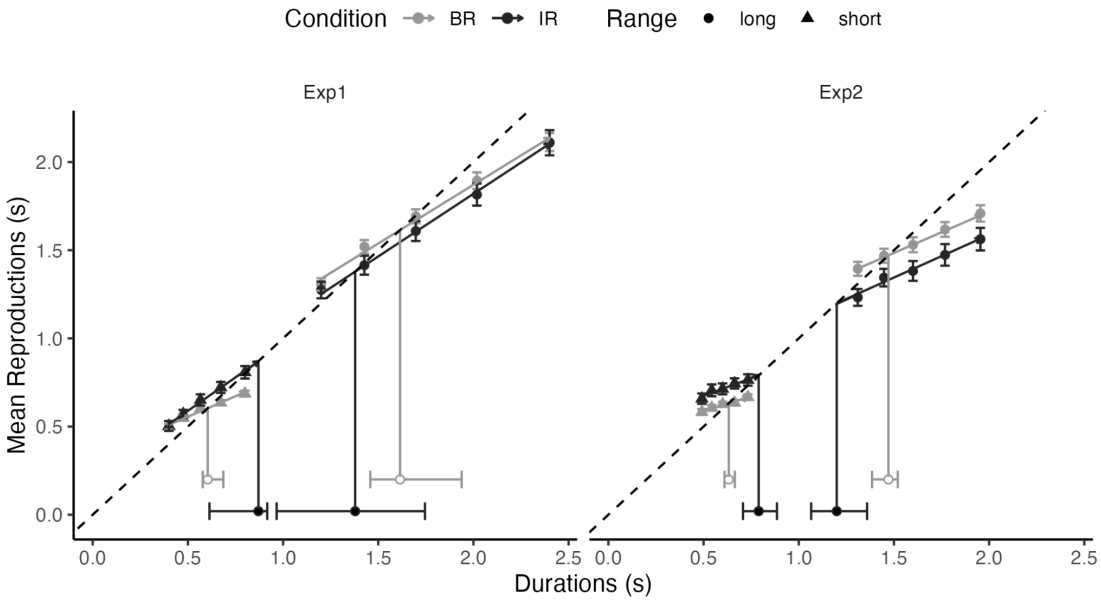

Figure 5 *Mean reproduced durations as a function of stimulus duration (filled symbols) presented in separate prior context and range condition. Solid continuous lines through triangles and solid points represent linear fits of the reproduced durations from the short and long interval condition, respectively, whereas the dotted diagonal lines denote veridical (unbiased) performance. Horizontal error bars denote standard errors of the mean of indifferent points across participants. Open points represent the estimated indifference points along with bootstrapped 95% confidence intervals.*

A 2×2 repeated-measures ANOVA on indifference points in Exp. 1 with the factors Prior Context (IR vs. BR) and Range (short vs. long) revealed a significant main effect of Range [$F(1, 15) = 34.39$, $p < .001$, $\eta_p^2 = .696$, $BF_{10} > 10000$], a significant Prior Context x Range interaction [$F(1, 15) = 7.75$, $p < .05$, $\eta_p^2 = .341$, $BF_{10} = 2.14$], but failed to find any significant effect of Prior Context [$F(1, 15) = 1.538$, $p = .234$, $\eta_p^2 = .093$]. A post-hoc analysis on the estimated indifferent points showed a significant effect of Prior Context (BR vs. IR) for the short and long range separately [short range, mean: 0.632(±0.10) vs. 0.765 (±0.28), $t(15) = -1.940$, $p = .035$, $BF_{10} = 2.286$; long range, mean: 1.70 (±0.45) vs.1.36 (±0.73), $t(15) = 2.203$, $p = .0218$, $BF_{10} = 3.342$], suggesting two individual local priors for the range short and long for the BR session were assimilated toward the grand mean of all intervals in the IR condition. This finding can be observed in Figure 5a, where the IPs induced by IR condition (solid circles) were migrated toward the mean of the grand mean of stimuli (1.16s).

Similarly, a repeated-measures ANOVA on the IPs in Exp. 2 revealed a significant main effect of Range [$F(1, 15) = 108.142$, $p < 0.001$, $\eta_p^2 = .878$, $BF_{10} > 10000$], a significant interaction [$F(1, 15) = 24.394$, $p < .001$, $\eta_p^2 = .619$, $BF_{10} = 986.124$], but again failed to show

any significant effect of Prior Context [$F(1, 15) = 1.412, p = .253, \eta_p^2 = .086$]. Furthermore, we conducted a one-tailed paired t-test comparing the estimated IPs between the BR and IR conditions, separately for the short and long range. These revealed IPs for the short intervals in the BR condition were significantly less than IR condition [mean: 0.637($\pm$0.05) vs. 0.8 ($\pm$0.17), $t(15) = -3.61, p < .005, BF_{10} = 34.096$], meanwhile the IPs for the long intervals in BR condition were significantly greater than the IR condition in Exp. 2 [mean: 1.45($\pm$0.127) vs. 1.21($\pm$0.276), $t(15) = 3.96954, p < .001, BF_{10} = 63.889$], indicating in IR condition two local priors were assimilated to each other, that is, migrated toward the grand mean of stimuli (1.12s).

BAYESIAN MODELING

In the present study, both experiments exhibited reliable central tendency effects in two prior context conditions, with a more "centered" reproduction for the IR condition. This indicated a tendency toward the mean of the underlying stimulus distribution in both IR and BR conditions, and the two local centers (indifference points) were migrated toward the grand center for the IR condition. In the BR condition, the indifference points were pulled toward the mean of duration set in the associated testing conditions, therefore we reasoned that observer learned two separate priors, a short prior and a long prior for the short and long intervals, respectively. In the IR session, the mean reproduction shows their own central tendency with the migration toward the grand mean of testing intervals. Of note, regression indices of the central tendency in the IR condition was significantly less than the BR condition. For the IR condition, a stronger central tendency with narrower local ranges and a larger separation (in Exp. 2) was observed, compared to Exp. 1. Inspired by those results, we hypothesized that in addition to the short and long priors formed in the BR condition, there was a general global prior, which played a role in the estimation process in the last IR session. The subjective estimation in the IR condition was produced as the integration of the global prior with the local priors.

Uni-prior Bayesian observer model

The central tendency effect of subjective estimates in the BR condition can be well explained by the uni-prior Bayesian observer model (Shi, Church, et al., 2013; Shi & Burr, 2016), where the distributions of the priors, likelihoods and posteriors are assumed to be Gaussian and the optimal estimate is calculated as the weighting of the sensory measurement of the current interval with the previous learned prior. Figure 6 illustrates how a Bayesian

observer learns the statistics of the history of stimuli and forms a prior around the mean of stimuli. The likelihood function $D_s(x)$, denoting the distribution of sensory measurement when stimulus is x, was assumed to be Gaussian distribution with mean x. As illustrated in Figure 6, two separate sessions in the present study were assumed to have their own priors $P_{Short}$and P_{Long} around the mean of the test interval set. The reproduced duration (represented as D_r) was considered as posterior by combing the prior (either P_{Long} or P_{Short}) with the likelihood of noisy sensory measurement D_s according to Bayes' rule. Taking the test interval 1.2s as an example, $D_s(1.2)$ denoting the likelihood function of 1.2s was understood as the distribution of sensory measurement of 1.2s. The distribution of reproduced duration was assumed as a joint probability distribution $(D_{L(1.2)})$, which was modeled as a posterior distribution by integrating the sensory measurement $(D_{s(1.2)})$ with the prior of long rang P_{Long}. The most likely estimate is represented as the maximum of posterior D_r, which has the highest accuracy. Coupled with measurement noise, it leads to the consequence that the measured interval in BR condition may differ from the test duration.

Multi-prior Bayesian observer model

Based on the results of indifference points, the centers of observed reproduction were migrated toward the grand centers of stimuli with regard to the IR condition. This migration in the IR condition cannot be easily explained by the uni-prior Bayesian model as the influence caused by two independent local priors. Otherwise, indifference points induced by the local priors would be predicted to be the mean of the distribution. However, the results of this study (see Figure 2 and Figure 5) showed indifferent points for the short and long intervals in the IR condition were biased towards to the grand mean.

Although the uni-prior Bayesian model explained how a prior around the mean of the test duration bias posterior estimates towards the center of the range in the BR experimental conditions, it cannot explain the migration from the reproductions of the observers for the BR-long (red line) and BR-short (green line) condition to the IR-long (blue line) and IR-short (purple line) condition, respectively, as shown in Figure 6. As mentioned above, we proposed that this migration was caused by a general global prior influencing the perceptual estimates in multiple prior context conditions. Our hypothesis is that the brain combines current sensory measurement, local priors, and a global prior for the final reproduction. Therefore, a multi-prior Bayesian observer model with a hierarchical structure describing how sensory

measurement and prior knowledge including local priors and global prior are integrated together may better predict the behavioral findings.

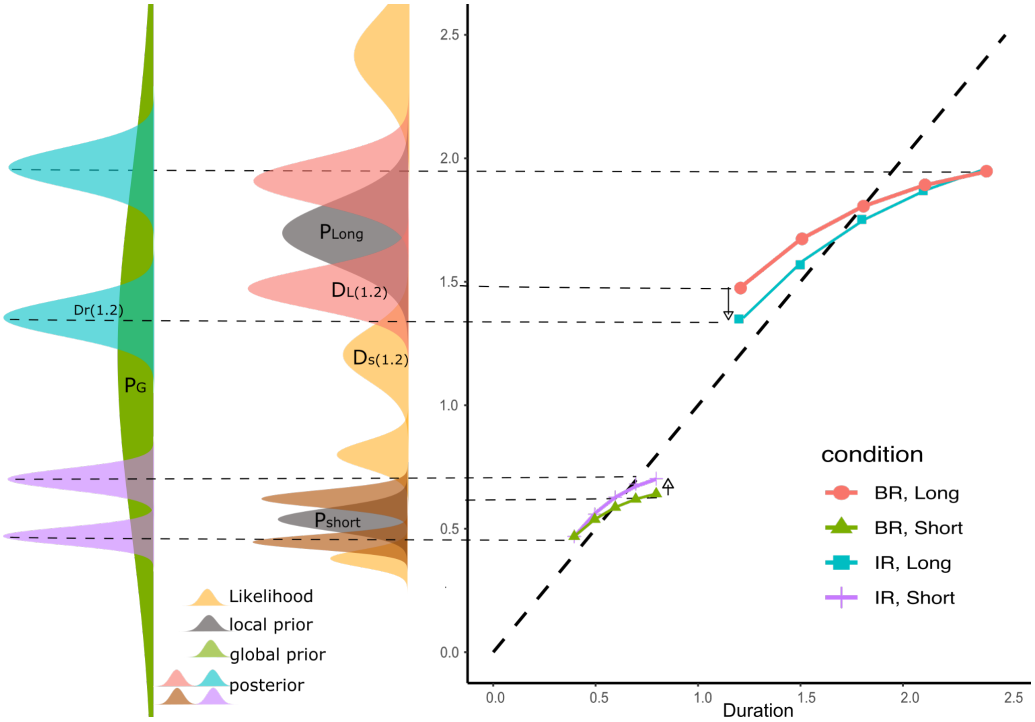

Figure 6 *Schematic illustration of a hierarchical Bayesian estimator model (taking local-global model as an example). In the IR condition, in addition to two local priors for the separate ranges, a global prior based on the whole range may also be developed and maintained in Bayesian updating. Both local and global priors integrate with the sensory measurement to form the time estimation. The sensory measurement* (D_s) *first integrates with the local prior* (P_L) *to obtain a posterior* (D_L)*, which further integrates with the global prior* (P_G) *to generate a final posterior for reproduction* (D_r)*. Taking testing duration 1.2s as an example,* $D_s(1.2)$ *denotes the distribution of sensory measurement. Two test ranges, range short and range long, have their own priors* P_{Short} and P_{Long}*. The distribution of reproduced duration is assumed as a joint probability distribution* $(D_{L(1.2)})$ *which can be modeled as posterior by integrating the sensory measurement* $(D_{s(1.2)})$ *with the local prior of rang long* P_{Long}*. The formed posterior* $(D_{L(1.2)})$ *further integrates with the global prior* (P_G) *to generate a final posterior for reproduction* $(D_{r(1.2)})$*.*

To better explain the perceptual estimates in the mixed session, we extended our previous uni-prior Bayesian model to the multi-prior temporal condition by adding a global prior P_G. To uncover the rules governing local priors and global prior, three alternative interpretations about the prior integration in the estimation process were proposed. The first

interpretation, referred to as the *hierarchical local-global* model, assumes that the sensory measurement is first integrated with the local prior, and then integrated with the global prior (see Figure 7a and Figure 6). The sensory measurement (D_s) first integrates with the local prior $(P_L = P_{Short} | P_{Long})$ to form a posterior (D_L), then the formed posterior (D_L) further integrates with the global prior (P_G) to generate a final posterior for reproduction (D_r). Alternatively, we proposed a *dual-integration* model (DIM) hypothesizing that the local prior for each range integrates with the sensory measurement separately firstly, and then the outputs from these integrations could be combined together to form the estimate (Figure 7b). The third model we proposed, referred to as the prior integration model (Figure 7c), assumes that global priorP_G and local prior P_L are firstly integrated to form a integrated prior P_I, and integrated prior P_I is then integrated with sensory measurement D_S into the duration estimation. The mathematical descriptions for above proposed models are available in Appendix A.

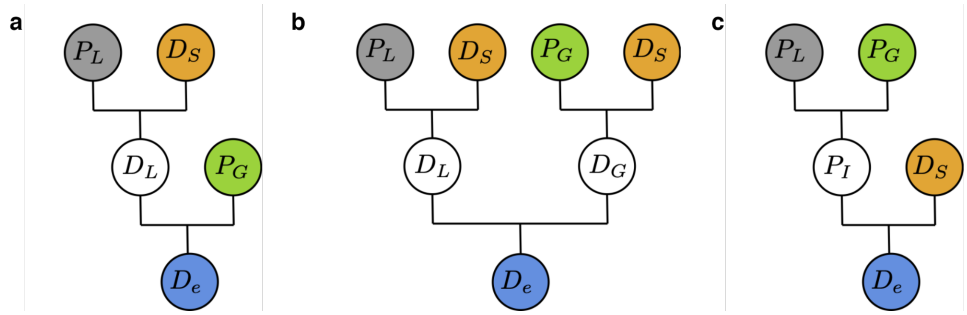

Figure 7 *Dendrogram of three proposed hierarchical structures of global and local priors integration in Bayesian estimation: local-global model(LGM), dual integration model(DIM), prior integration model(PIM)* ***a****.The local-global model(LGM) assumes that the sensory measurement* (D_S) *first combines with the local prior* (P_L) *to form posterior* (D_L)*, and the formed posterior* (D_L) *further integrates with the global prior* (P_G) to generate *reproduction* (D_r)*.* ***b****. The dual integration model (DIM) assumes that each range has its own prior and integrates with the sensory measurement separately firstly, and then the outputs from these integrations could be combined into the posteriors. Both local and global priors independently integrate with the sensory measurement to generate two posteriors* (D_L) *and* (D_G)*, the latter two are combined together for estimate*(D_e)*.* ***c****. The prior integration model (PIM) assumes that global prior*P_G *and local prior* P_L *are firstly integrated to form a integrated prior* P_I*, and this integrated prior* P_I *is then integrated with sensory measurement* D_S *to form the estimated duration.*

Implementation of models and parameter estimation

There are two alternative assumptions on the encoding of subjective durations, logarithmic scaled and linear scaled. The logarithmic encoding of subjective durations assumes a constant of Weber fraction of time estimation based on scalar property (Gibbon et al., 1984a; Shi, Church, et al., 2013), while the linear encoding of subjective durations supports a linearly increasing noise with the absolute magnitude according to Weber's law(Cliff et al., 2019; Fechner, 1966). Since empirical justification of logarithmic or linear scaled on internal duration mainly depends on the adopted experimental paradigms (Matthews & Meck, 2016; Ren et al., 2021a), hence whether adopting linear or logarithmic coding of the internal representation was not ascertained, therefore we conducted both logarithmic and linear encoding models and compared their predictions to explore which model can achieve better performance (see Appendix B).

In our Bayesian framework, we applied the uni-prior model for the BR session data to yield the distribution of the short prior (Gaussian distribution with mean $\mu_{X_S}$and variance $\sigma_{X_S}^2$), the long prior (Gaussian distribution with mean $\mu_{X_L}$and variance $\sigma_{X_L}^2$), and the motor noise ($N(0, \sigma_m^2)$). In addition, we have the parameter σ_s^2 for the variance of the sensory measurement in the log scale (which we assume the Weber scaling in duration perception). After we obtained the parameters from the BR session, we used those six parameters for the IR session, assuming the local priors (either the short or the long) and the motor noise did not change. The additional parameters we set free were the distribution of the global prior (μ_{X_G} and $\sigma_{X_G}^2$ denoting the mean and variability of global prior).

All proposed models (see Figure 6) were implemented by Stan in Rstudio (Carpenter et al., 2017; Team & Others, 2018), a platform for statistical modeling and Bayesian statistical inference. After implementation of the notational specification of probability models, the model specification was compiled from Stan's probabilistic programming language into C++ program by the Stan platform. Sequentially, Markov Chain Monte Carlo (MCMC) algorithms for Bayesian Inference were used to sample the parameters to maximize the joint posterior distribution for the parameters of interest. In this study, 12000 iterations per chain (8 chains in total) with 2000 warmup iterations were applied to achieve the optimization of parameters and our Stan models were diagnosed with their assess convergence to check if parameters were drawn from the actual posterior distribution of interest.

The proposed models were described in detail in Appendix A, and all experimental

dataset, R-analysis and RStan modeling codes are available at https://github.com/msenselab/PriorsIntegration.

MODEL RESULTS

Prediction of the mean of reproduction

To visualize the goodness of the fit of each model, the prediction error (i.e., the error between predicted reproduction and observation) in the IR condition across all observers were plotted in Figure 8a. As depicted in Figure 8**b**, the mean absolute errors on predicted reproduction mean was plotted versus errors on predicted reproduction variance. Each point represents the errors in prediction error estimation derived from the proposed model per individual participant. The bigger points indicate the averaged means and variances of reproductions predicted by the model across all subjects. To further evaluate the prediction performance, we applied the mean absolute percentage errors (MAPE) to measure the accuracy of the prediction by averaging the absolute percentage error of prediction. As depicted in Figure 8b, the dual integration model (DIM) had better prediction performance on mean relative accuracy of the reproduced durations prediction(DIM 98.08%, LGM 97.42%, PIM 97.42% for Experiment 1, and DIM 97.80%, LGM 97.49%, PIM 97.49% for Experiment 2), but worst relative accuracy of the reproduction variance prediction(DIM 51.92%, LGM 56.27%, PIM 97.42% for Experiment 1, and DIM 68.46%, LGM 72.70%, PIM 72.70% for Experiment 2). Note that LGM and PIM models yielded equally well prediction performance in reproduction mean and variance prediction for varying experimental conditions.

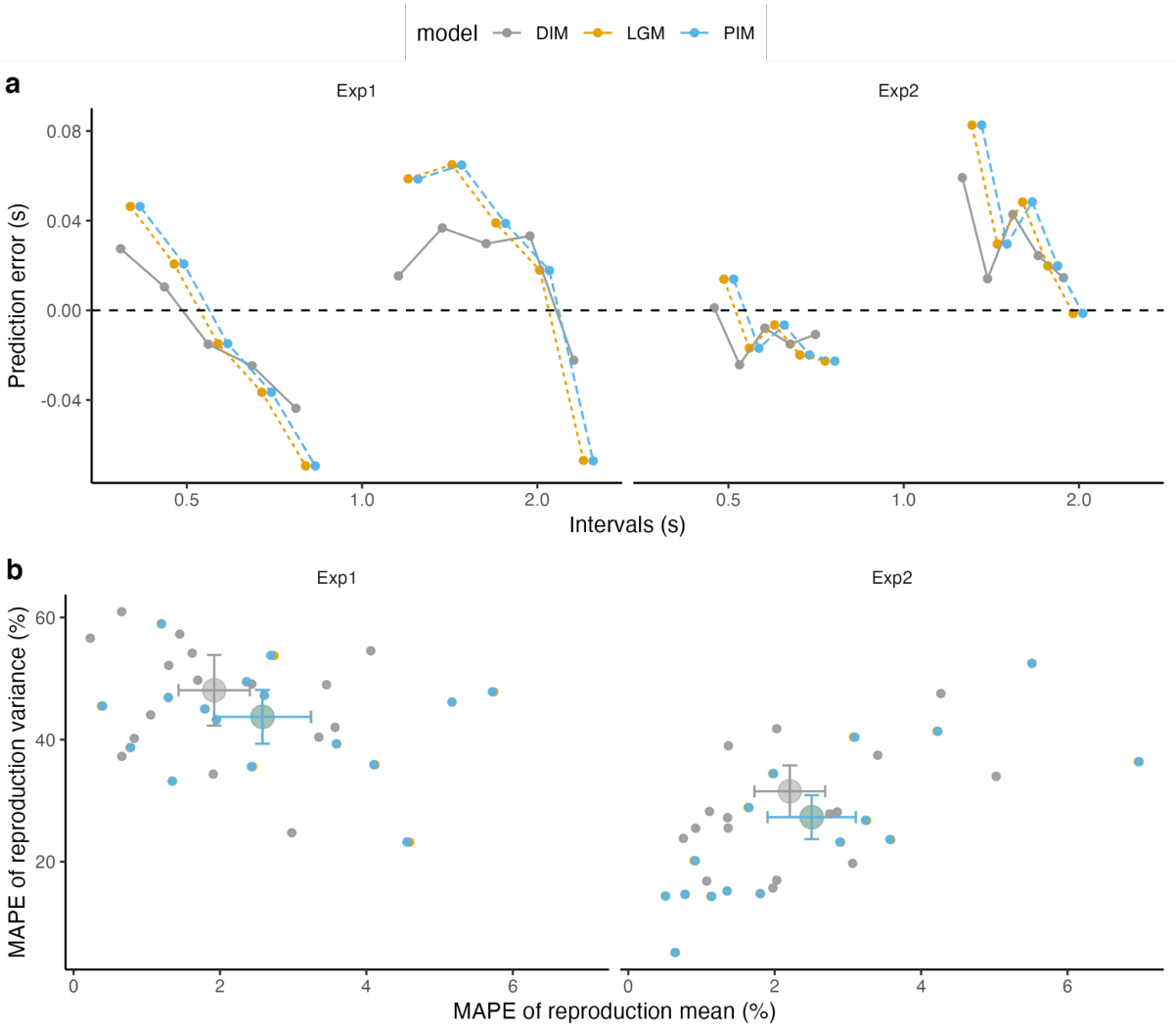

Figure 8 a. *Mean prediction error from proposed models (mean predicted reproduction - mean of observed reproduction) across observers for all test intervals from all conditions. The dotted lines, dashed lines and solid lines represent mean prediction error derived from the hierarchical local-global model (LGM), the prior integration model (PIM) and the dual integration model (DIM), respectively. Horizontal dotted line indicates mean observed reproduction.* ***b****. The scatterplot of mean absolute percentage errors (MAPEs) of the means and variances of reproduction predicted by models LGM, PIM, and DIM for individual observers in Experiments 1 and 2. Note that the results of LGM were overlapped by the results from PIM, since they yielded almost the same predictions.*

Furthermore, Watanabe-Akaike information criterion (WAIC) and leave-one-out information criterion (LOOIC) were used as fit indices for each of three Bayesian models. The WAIC and LOOIC are widely applied in practical Bayesian model evaluation and model selection. These measures not only take into account the goodness-of-fit, but also penalize models with more free parameters (Vehtari et al., 2017). Lower WAIC values indicate better

model performance. As shown in Table A1, the DIM had the best performance in both logarithmic and linear scale models for Experiments 1 and 2, given that it yielded the lowest WAIC and the lowest LOOIC.

Estimated global priors

In this study, we used two parameters (μ_{X_G} and $\sigma^2_{X_G}$) for the mean and variability of the global prior distribution (for more details see Appendix A). A mixed-design ANOVA of the estimated mean global prior (μ_{X_G}) between Exp. 1 and 2 as the between-subject factor and Model(DIM, LGM, and PIM) as the within-subject factor revealed a significant main effect of Experiment [*F(1, 30) = 6.391, p = .0170,* η_p^2 *= .339,* BF_{10}*>1000*], but failed to reveal any significant of Model [*F(2, 60) = .00045, p=.999,* η_p^2*<.0001,* BF_{10}*>1000*], or the interaction between two factors [*F(2, 60) = .0263, p =.974,* η_p^2 *<.0001,* BF_{10}*= 4.362*].

Similarly, a mixed-design ANOVA of the estimated variance of global prior ($\sigma^2_{X_G}$) between Exp. 1 and 2 as the between-subject factor and Model (DIM, LGM, and PIM) as the within-subject factor revealed a significant main effect of Model [*F(2, 60) =14.545, p = <.0001,* η_p^2 *= .0578,* BF_{10}*>1000*], but failed to reveal any significant of Experiment [*F(1, 30) = .842, p=.366,* η_p^2*<.045,* BF_{10}*=.952*], or the interaction between two factors [*F(2, 60) = 2.148, p =.125,* η_p^2 *<.001,* BF_{10}*= 2.172*]. The short range and long range in Exp.2 were shrunken and separation between the short and long-range were expanded compared to Exp.1, a more sharply peaked local prior (smaller variance) was expected in Exp.2. The results showed a smaller variance of global prior in Exp.2 was observed, consistent with our expectation. Interestingly, though, Exp. 1 and Exp. 2 had the same geometric mean for the IR session, we found a significant difference in the mean of global prior between Exp. 1 and Exp. 2.

Effect of session order

To examine the effect of session order on the performance during the mixed session, the estimated global priors (±SE) were plotted against Models (DIM, LGM, and PIM), for session order Long-Short-Mixed (L-S-M) and Short-Long-Mixed (S-L-M) in Exp.1 and 2 (see Figure 9). A mixed-design ANOVA of the estimated mean global prior (μ_{X_G}) with Experiment (Exp. 1 and 2) and Session Order(L-S-M vs. S-L-M) as the between-subject factor and Model (DIM, LGM, and PIM) as the within-subject factor revealed a significant main effect of Experiment [*F(1, 28) = 8.624, p < .01,* η_p^2 *= .235,* $BF_{10} = 2.546$], Session Order [*F(2, 60) = 12.48, p<.001,* η_p^2*=.308,* BF_{10}*=7.422*] and Model [*F(2, 60) = 14.544, p<.001,* η_p^2*=0.045,*

BF_{10}=*.977*]. Furthermore, pairwise T-tests between session order revealed significantly higher mean global priors for session order S-L-M than L-S-M in both Exp.1 and Exp.2 (see *Table 1*), suggesting the mean of global prior was higher when the long session as the previous session prior to the mixed session.

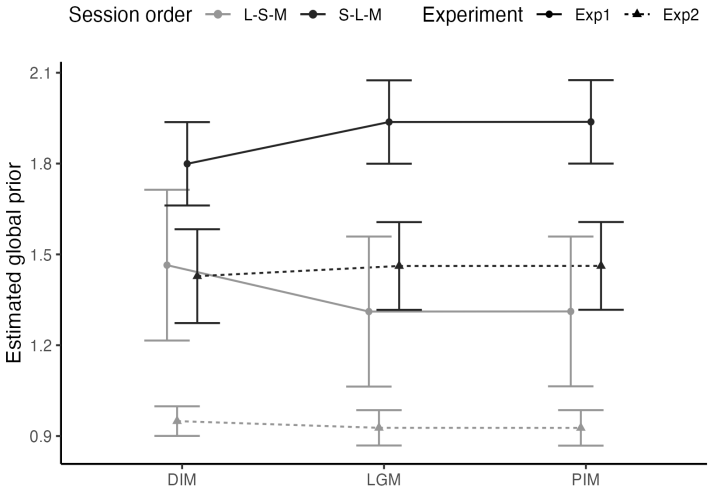

Figure 9 *Mean estimated global priors (and associated standard errors) derived from DIM, LGM, and PIM models plotted against session order Long-Short-Mixed (L-S-M) and Short-Long-Mixed (S-L-M) in Exp.1 and Exp.2.*

Table 1 Statistical results of one-way pairwise t-tests between session order on the mean of estimated global priors derived from DIM, LGM, PIM for Exp.1 and Exp.2.

	model	t-value	p-value	BF_{10}
Exp. 1	DIM	-1.258	.115 (ns)	1.438
	LGM	-2.361	.017 (*)	4.569
	PIM	-2.364	.017 (*)	4.585
Exp. 2	DIM	-3.150	.004 (**)	13.34
	LGM	-3.658	<.001 (***)	28.188
	PIM	-3.658	<.001 (***)	28.208

Weights of the priors in timing estimation

The estimated weight of global prior, local prior and sensory measurement in Bayesian time estimation derived from the proposed models are illustrated in Figure 10.

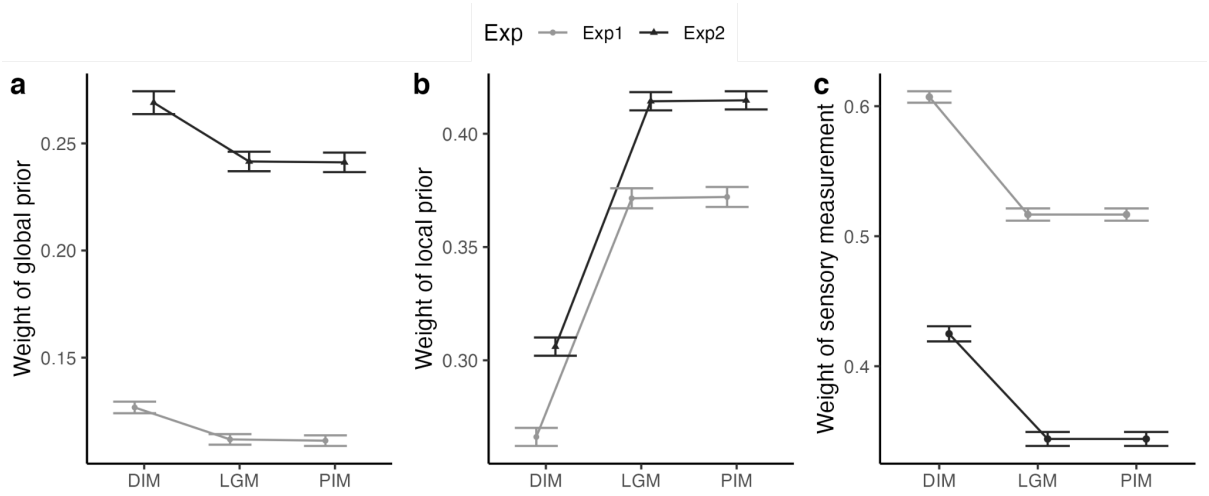

Figure 10 *Weight of global prior (a), local prior (b) and sensory measurement D_s(c) in timing estimation derived from the Bayesian models.*

A mixed-design ANOVA on the estimated weight of global prior (Figure 10a) with Experiment (1-2) as between-subject factor and Model (DIM, LGM, and PIM) as within-subject factor revealed a significant main effect of Experiment [$F(1, 30) = 5.024$, $p = .003$, $\eta_p^2 = .275$, $BF_{10} > 1000$], and of Model [$F(2, 60) = 17.971$, $p < .0001$, $\eta_p^2 = .495$, $BF_{10} = .119$], but no significant interaction between two factors [$F(2, 60) = 1.535$, $p = .224$, $\eta_p^2 = .116$, $BF_{10} = .161$]. As mentioned above, we used the same uni-prior model to obtain the same distribution of local priors, therefore the variability of the mixed session is the main independent variable causing the difference in weight of global prior. We found that the weight of the global prior was larger in Exp. 2 than in Exp. 1 [paired t-test LGM: $t(15) = -2.392$, $p < .05$, DIM: $t(15) = -2.360$, $p < .05$, PIM: $t(15) = -2.391$, $p < .05$], suggesting the weight of global prior is influenced by standard deviation of the mixed session, that is, when the standard deviation of the mixed session is smaller (Exp. 2), the estimation relies more on the global prior, compared to standard deviation is greater (Exp. 1)

A 2×3 mixed-design repeated-measures ANOVA on the weight of local prior (Figure 10b) with the factors Model between Experiments (1-2) revealed a significant main effect of Model [$F(1, 30) = 440.559$, $p = < .00001$, $\eta_p^2 = .973$, $BF_{10} = 2579.45$]. However, neither Experiment [$F(2, 60) = .858$, $p = .36$, $\eta_p^2 = .0577$, $BF_{10} = 1.603$], nor the Experiment

×Model [*F(2, 60) = .072, p =9.309,* η_p^2*=.004,* BF_{10}*=.145*], were significant. Another mixed-design ANOVA on the estimated weight of sensory measurement (D_s) revealed a significant main effect of Experiment [*F(1, 30) = 5.398, p = .027,* η_p^2 *=.263 ,* BF_{10}*>1000*], and Model [*F(2, 60) = 153.297, p <.0001,* η_p^2*= .877,* BF_{10}*=.838*], but no significant interaction between of the factors [*F(2, 60) = .535, p =.588,* η_p^2*=.0598 ,* BF_{10}*=.155*]. The weight of sensory measurement (D_s) was larger in Exp.1 than in Exp. 2 [LGM: *t(15) =2.286, p < .05,* DIM: *t(15)=2.351, p < .05, PIM: t(15)=2.285, p<.05*] (see Figure 10c). As the variability of the mixed session in Exp. 2 was smaller than Exp.1, the estimated weight of sensory measurement (Ds) was larger in Exp. 1 relative to Exp. 2, and this in turn caused higher central tendency effects in Exp. 2.

DISCUSSION

The present study aimed to investigate the research question of how multiple contexts are integrated with the sensory measurement in time estimation. The difference of the central tendency effects between the IR and BR conditions and the migration of the indifference points in the IR relative to the BR condition suggest that time estimation in the final random intermixed session (IR) was not merely influenced by the local context of the sampled duration, rather also influenced by the previous acquired context and the global range of the sample intervals (we refer this as to the global prior). To reveal the structure of prior integration, we proposed three Bayesian multi-prior integration models assuming estimation to be an optimal integration of the noisy sensory measurement with multiple priors (the local short or long prior, and the global prior) in a hierarchical order. Afterwards, our proposed multi-prior integration models were validated by fitting behavior data and the model prediction revealed our proposed models are in good agreement with the behavioral data.

The structure of the prior integration

As an extension of the basic iterative model (Petzschner & Glasauer, 2011), Petzschner et al. (2012b) proposed two symbolic cue integration models (categorical model and cue-combination model) to account integration process of multiple contexts (categorical cue, and the sample distribution) in a distance reproduction task for the “interleaved-ranges, no cue” and the “interleaved-ranges, cue” sessions. Particularly, based on an assumption that distinct categories from which the stimuli are drawn (Feldman et al., 2009), the categorical model combines prior and symbolic cue as combined prior in Gaussian mixture distribution, then

ingrates combined prior with sensory measurement to yield the posterior. Additionally, taking the symbolic cue as additional modality, a cue-combination model combines symbolic cue and sensory measurement as fused signal firstly, then combines fused signal with prior to yield the posterior. Both models showed similar performance in the prediction of observed behavior, and the model results revealed that the cue context was incorporated into the distance reproduction.

Similar to the approach of combination between the categorical context and prior in the previous study (Petzschner et al., 2012b), our proposed prior integration model combines global prior and local prior as integrated prior firstly, then the integrated prior combines with sensory measurement to yield posterior. Differ from the discrete symbolic cue in Petzschner's symbolic cue models, acquired priors in our study are assumed to be continuous given the multiple contexts are the sampled durations (such as within the session, and the previous session). In Petzschner's study, the combined prior was a mixture Gaussian distribution which is achieved by weighting each prior with the conditional probability of the respective category, whereas the integrated prior in our PIM model were two normal distributions of local prior and global prior. In addition to the PIM model, we proposed two other alternative prior integration structures, which showed equally good in prediction: the local-global model integrated the sensory measurement with the local prior, then further integrates with the global prior, which adopts a different integration order with PIM (see Figure 7). Interestingly, the model results showed LGM and PIM models with two different hierarchical orders yielded the same prediction results in reproduction mean and variance prediction for varying experimental conditions, implying the different integration order of priors with sensory measurement did not cause difference in prediction. The overlapping of two models is partly owing to the linear combination in nature with the Gaussian distribution.

Different from the sequential combination of LGM and PIM, the last proposed dual integration model (DIM) assumed that the sensory measurement is integrated with the local prior and the global prior separately, and the final decision is a combination of the two, which is more complex than the LGM and PIM. The idea of DIM is that the brain might keep two decisions first separately (one for the local and one for the global) for different decision scenarios (e.g., based on the relevance of the local and global contexts). As depicted in Figure 7, in the DIM the sensory measurement integrates twice with the global and local priors, thus two posteriors are partially correlated. In order to combine these two posteriors together, we applied the correlated distribution theory (Oruç et al., 2003). The model results of DIM showed better performance in prediction of reproduction mean but worse performance in prediction of reproduction variance, as compared to LGM and PIM.

Context further in the past is less weighted than more recent context

In the present study, the same geometric mean of the interleaved range was applied to both experiments. Thus, if the global prior was learnt independent of the session order, we should observe comparable global priors. However, the behavioral results showed the session order effect. The session prior to the IR session influenced the global prior more than the first session of the experiment, rather than equal impact on that. And all models also confirmed that the mean of the global prior is correlated to the mean prior of the previous session. That is, the mean global prior was longer for the previous long-range session as compared to the short-range session. Recent studies have also shown a similar history dependent effect, such as the statistics of past interval samples adaptively incorporated in perceptual estimates and samples farther in the past being less taken into account during estimation because of their less reliability (Zimmermann & Cicchini, 2020). That means, the short prior learned in the first session (farther in the past) was less reliable and the long prior learned in the second session was higher reliable, implying the long prior was weighted in the estimation of global prior which caused higher global prior. This is similar to the sequential dependence in trial-by-trial updating (Glasauer & Shi, 2022), remote history trials had less impact on the current trial. It can be explained by the dynamic Kalman filter (Petzschner and Glasauer 2011), a special moving average process, which weights recent trials more than the earlier trials.

Rely more on prior information with expanding separation and narrowing local ranges

According to the Bayesian modeling framework, when the sensory measurement has high precision, there would be less influence of prior knowledge, resulting in a lesser central-tendency bias (Cicchini et al., 2012; Shi, Church, et al., 2013; Shi & Burr, 2016). The prior knowledge is weighted more when sensory estimates are imprecise, suggesting subjects rely more on prior knowledge to reduce noise (Karaminis et al., 2016). Initially, we expected a trade-off between the weights of local and global prior by adjusting the ranges of short and long sessions, while the sensory measurement would be similar. When the short and long ranges were shrunk and the separation between short-range and long-range was spanded in Exp. 2 as compared to Exp. 1, responses of all subjects were expected to be attracted more toward the mean of local priors for short-range and long-range in Exp. 2 relative to Exp. 1. Surprisingly, the behavior and model results showed that the weight of prior information (including global prior and local prior) increases with the shortened mixed range and the widened separation in Exp. 2. We observed the centrality index (CI) was higher in Exp. 2 than Exp. 1 (see Figure 4),

suggesting the central tendency in duration reproduction tasks in Exp. 2 was stronger than in Exp.1. It is likely owing to the fact that the shortened range reduced the variability (uncertainty) as well, which subsequently led to higher reliable local and global priors.

To check the effect of separation between the short and long range, we had to consider the changes of indifference points and reproduction errors between the IR and BR condition. The results of indifference points from two ranges in the IR session were assimilated to each other and a stronger migration towards the grand mean of stimuli in Exp. 2 relative to Exp. 1 (see Figure 5). The mean of the reproduction bias showed a larger difference between the BR and IR session in Exp. 2 relative Exp. 1 (see Figure 2**c** and **f**). Taking the above results together, the experiment with a wider separation showed a stronger driving force toward the local range center (stronger central tendency for the local range) and more migration to the grand center (reflected as more changes in indifference points).

Based on the results generated from the models, we found that local and global priors showed the same trend on their weights in estimation. In addition, both local priors and global prior weighted more in Exp. 2 (with lower variability) relative to Exp. 1 (with higher variability) (see Figure 10). From another perspective, sensory measurement was expected to be weighted less in Exp. 1, since the variability of the IR condition in Exp. 1 was larger than Exp.2 (0.634 s in Exp.1 vs. 0.538 s in Exp.2). The model results suggested the weight of global prior was mainly influenced by the variability of the interleaved range. We can conclude that subjects rely more on prior knowledge (including local priors and global prior) but less on sensory measurement in the condition of wider separation between short-range and long-range.

CONCLUSION

The present study concentrated on the underlying structure of the prior knowledge integration in multi-prior contexts and the question of how multiple priors influence subjective duration timing in duration reproduction tasks. Two duration reproduction experiments with the same geometric mean of interleaved range but with two levels of variance of short and long range by rescaling short and long ranges and expanding the separation between the short and long intervals were carried out. Three possibilities of integration models were proposed to explore the integration structure of combining multiple prior information with sensory measurement, and results showed their performance were equally well performance in reproduction mean and variance prediction despite their different structural assumptions. We found the order of the

training session influences the formation of global prior, since context farther in the past is less weighted than more recent context.

ACKNOWLEDGEMENTS

This study was supported by German Science Foundation (DFG) research grants SH 166/ 3-2, awarded to ZS, and GL 342/3-2, awarded to SG, as well as the Starting Research Fund from Hangzhou Normal University (RWSK20190315) awarded to XZ.

REFERENCES

Acerbi, L., Wolpert, D. M., & Vijayakumar, S. (2012). Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. *PLoS Computational Biology*, *8*(11), e1002771. https://doi.org/10.1371/journal.pcbi.1002771

Berniker, M., Voss, M., & Kording, K. (2010). Learning priors for Bayesian computations in the nervous system. *PloS One*, *5*(9). https://doi.org/10.1371/journal.pone.0012686

Buhusi, C. V., & Meck, W. H. (2009). Relativity theory and time perception: single or multiple clocks? *PloS One*, *4*(7), e6268. https://doi.org/10.1371/journal.pone.0006268

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software, Articles*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Cicchini, G. M., Arrighi, R., Cecchetti, L., Giusti, M., & Burr, D. C. (2012). Optimal encoding of interval timing in expert percussionists. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(3), 1056–1060. https://doi.org/10.1523/JNEUROSCI.3411-11.2012

Cliff, J., Jackson, S., McEwan, J., & Bizo, L. (2019). Weber's Law and the Scalar Property of Timing: A Test of Canine Timing. *Animals*, *9*(10), 801. https://doi.org/10.3390/ani9100801

Dyjas, O., Bausenhart, K. M., & Ulrich, R. (2012). Trial-by-trial updating of an internal reference in discrimination tasks: evidence from effects of stimulus order and trial sequence. *Attention, Perception & Psychophysics*, *74*(8), 1819–1841. https://doi.org/10.3758/s13414-012-0362-4

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, *28*(1), 1–11. https://doi.org/10.3758/BF03203630

Fechner, G. (1860). Elements of psychophysics. In *W. Dennis (Ed.), Readings in the history of psychology* (pp. 206–213). https://doi.org/10.1037/11304-026

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–782. https://doi.org/10.1037/a0017196

Fritsche, M., Spaak, E., & de Lange, F. P. (2020). A Bayesian and efficient observer model explains concurrent attractive and repulsive history biases in visual perception. *eLife*, *9*. https://doi.org/10.7554/eLife.55389

Gau, R., & Noppeney, U. (2016). How prior expectations shape multisensory perception. In *NeuroImage* (Vol. 124, pp. 876–886). https://doi.org/10.1016/j.neuroimage.2015.09.045

Gibbon, J., Church, R. M., & Meck, W. H. (1984a). Scalar timing in memory. *Annals of the New York Academy of Sciences*, *423*, 52–77. https://www.ncbi.nlm.nih.gov/pubmed/6588812

Gibbon, J., Church, R. M., & Meck, W. H. (1984b). Scalar timing in memory. *Annals of the New York Academy of Sciences*, *423*, 52–77. https://www.ncbi.nlm.nih.gov/pubmed/6588812

Glasauer, S., & Shi, Z. (2019). *Central Tendency as Consequence of Experimental Protocol*. Conference on Cognitive Computational Neuroscience, Berlin, Germany.

Glasauer, S., & Shi, Z. (2022). Individual beliefs about temporal continuity explain variation of perceptual biases. *Scientific Reports*, *12*(1), 10746. https://doi.org/10.1038/s41598-022-14939-8

Glasauer, S., Shi, Z., Maier, P., & Petzschner, F. H. (2020). Short-term adaptation to stimulus statistics depends on behavioral relevance. *(Manuscript)*.

Gu, B.-M., Jurkowski, A. J., Shi, Z., & Meck, W. H. (2016). Bayesian optimization of interval timing and biases in temporal memory as a function of temporal context, feedback, and dopamine levels in young, aged and Parkinson's disease patients. *Timing & Time Perception*, *4*(4), 315–342. https://doi.org/10.1163/22134468-00002072

Heron, J., Aaen-Stockdale, C., Hotchkiss, J., Roach, N. W., McGraw, P. V., & Whitaker, D. (2012).

Duration channels mediate human time perception. *Proceedings. Biological Sciences*, *279*(1729), 690–698. https://doi.org/10.1098/rspb.2011.1131

Hollingworth, H. L. (1910). The Central Tendency of Judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, *7*(17), 461–469. https://doi.org/10.2307/2012819

Ivry, R. B., & Richardson, T. C. (2002). Temporal control and coordination: the multiple timer model. *Brain and Cognition*, *48*(1), 117–132. https://doi.org/10.1006/brcg.2001.1308

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*, *13*(8), 1020–1026. https://doi.org/10.1038/nn.2590

Karaminis, T., Cicchini, G. M., Neil, L., Cappagli, G., Aagten-Murphy, D., Burr, D., & Pellicano, E. (2016). Central tendency effects in time interval reproduction in autism. *Scientific Reports*, *6*, 28570. https://doi.org/10.1038/srep28570

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception ECVP Abstract Supplement*, *36*, 14.

Knill, D. C., & Richards, W. (Eds.). (1996). Perception as Bayesian inference. *Cambridge University Press*. https://doi.org/10.1017/CBO9780511984037

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., Ly, A., Gronau, Q. F., Smira, M., Epskamp, S., & Others. (2019). JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software*, *88*(2), 1–17. https://doi.org/10.18637/jss.v088.i02

Luu, L., & Stocker, A. A. (2018). Post-decision biases reveal a self-consistency principle in perceptual inference. *eLife*, *7*. https://doi.org/10.7554/eLife.33334

Matthews, W. J., & Meck, W. H. (2016). Temporal Cognition: Connecting Subjective Time to Perception, Attention, and Memory. *Psychological Bulletin*, *2*. https://doi.org/10.1037/bul0000045

Miyazaki, M., Nozaki, D., & Nakajima, Y. (2005). Testing Bayesian models of human coincidence timing. *Journal of Neurophysiology*, *94*(1), 395–399. https://doi.org/10.1152/jn.01168.2004

Nagai, Y., Suzuki, M., Miyazaki, M., & Kitazawa, S. (2012). Acquisition of multiple prior distributions in tactile temporal order judgment. *Frontiers in Psychology*, *3*(August), 276.

https://doi.org/10.3389/fpsyg.2012.00276

Oruç, I., Maloney, L. T., & Landy, M. S. (2003). Weighted linear cue combination with possibly correlated error. *Vision Research*, *43*(23), 2451–2468. https://doi.org/10.1016/s0042-6989(03)00435-8

Paton, J. J., & Buonomano, D. V. (2018). The Neural Basis of Timing: Distributed Mechanisms for Diverse Functions. *Neuron*, *98*(4), 687–705. https://doi.org/10.1016/j.neuron.2018.03.045

Petzschner, F. H., & Glasauer, S. (2011). Iterative Bayesian estimation as an explanation for range and regression effects: a study on human path integration. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(47), 17220–17229. https://doi.org/10.1523/JNEUROSCI.2028-11.2011

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

Petzschner, F. H., Maier, P., & Glasauer, S. (2012). Combining symbolic cues with sensory input and prior experience in an iterative bayesian framework. *Frontiers in Integrative Neuroscience*, *6*, 58. https://doi.org/10.3389/fnint.2012.00058

Ren, Y., Allenmark, F., Müller, H. J., & Shi, Z. (2021a). Variation in the "coefficient of variation": Rethinking the violation of the scalar property in time-duration judgments. *Acta Psychologica*, *214*, 103263. https://doi.org/10.1016/j.actpsy.2021.103263

Ren, Y., Allenmark, F., Müller, H. J., & Shi, Z. (2021b). Variation in the "coefficient of variation": Rethinking the violation of the scalar property in time-duration judgments. *Acta Psychologica*, *214*, 103263. https://doi.org/10.1016/j.actpsy.2021.103263

Roach, N. W., McGraw, P. V., Whitaker, D. J., & Heron, J. (2017a). Generalization of prior information for rapid Bayesian time estimation. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(2), 412–417. https://doi.org/10.1073/pnas.1610706114

Roach, N. W., McGraw, P. V., Whitaker, D. J., & Heron, J. (2017b). Generalization of prior information for rapid Bayesian time estimation. *Proceedings of the National Academy of*

Sciences of the United States of America, *114*(2), 412–417. https://doi.org/10.1073/pnas.1610706114

Ryan, L. J. (2011). Temporal context affects duration reproduction. *Journal of Cognitive Psychology* , *23*(1), 157–170. https://doi.org/10.1080/20445911.2011.477812

Shi, Z., & Burr, D. (2016). Predictive coding of multisensory timing. *Current Opinion in Behavioral Sciences*, *8*, 200–206. https://doi.org/10.1016/j.cobeha.2016.02.014

Shi, Z., Church, R. M., & Meck, W. H. (2013). Bayesian optimization of time perception. *Trends in Cognitive Sciences*, *17*(11), 556–564. https://doi.org/10.1016/j.tics.2013.09.009

Shi, Z., Ganzenmüller, S., & Müller, H. J. (2013). Reducing bias in auditory duration reproduction by integrating the reproduced signal. *PloS One*, *8*(4), e62065. https://doi.org/10.1371/journal.pone.0062065

Team, S. D., & Others. (2018). *RStan: the R Interface to Stan. R package version 2.17. 3*.

Teghtsoonian, R., & Teghtsoonian, M. (1978). Range and regression effects in magnitude scaling. *Perception & Psychophysics*, *24*(4), 305–314. https://doi.org/10.3758/BF03204247

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vierordt, K. von. (1868). *Der Zeitsinn nach Versuchen [The sense of time according to research]*. H. Laupp.

von Helmholtz, H. (1867). *Treatise on Physiological Optics Vol. III*. Dover Publications.

Wearden, J. H., Edwards, H., Fakhri, M., & Percival, A. (1998). Why “sounds are judged longer than lights”: application of a model of the internal clock in humans. *The Quarterly Journal of Experimental Psychology. B, Comparative and Physiological Psychology*, *51*(2), 97–120. https://doi.org/10.1080/713932672

Wearden, J. H., & Lejeune, H. (2008). Scalar properties in human timing: Conformity and violations. *The Quarterly Journal of Experimental Psychology*, *61*(4), 569–587. https://doi.org/10.1080/17470210701282576

Wei, X. X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can

explain’anti-Bayesian'percepts. *Nature Neuroscience*. https://doi.org/10.1038/nn.4105

Westheimer, G. (2008). Was Helmholtz a Bayesian? *Perception*, *37*(5), 642–650. https://doi.org/10.1068/p5973

Wiener, M., Thompson, J. C., & Coslett, H. B. (2014). Continuous carryover of temporal context dissociates response bias from perceptual influence for duration. *PloS One*, *9*(6), e100803. https://doi.org/10.1371/journal.pone.0100803

Zang, X., Glasauer, S., Assumpcao, L., Wu, J., Xiu, C., & Shi, Z. (2022). What makes it so hard to acquire multiple priors in duration estimation? *In Submission*.

Zimmermann, E., & Cicchini, G. M. (2020). Temporal Context affects interval timing at the perceptual level. *Scientific Reports*, *10*(1), 8767. https://doi.org/10.1038/s41598-020-65609-6

APPENDIX

Appendix A. Mathematical expression of prior integration models

In the Bayesian approach, the sensory measurement is noisy and uncertain. The probability distribution of the sensory measurement is presented by a likelihood function D_s, $D_s \sim N(\mu_s, \sigma_s^2)$. Local prior of short range P_{Short} is assumed to be Gaussian distribution with mean μ_{X_S} and variance $\sigma_{X_S}^2$, local prior of long range P_{Long} is assumed to be Gaussian distribution with mean μ_{X_L} and variance $\sigma_{X_L}^2$, and global prior P_G is assumed to be distributed normally with mean μ_{X_G} and variance $\sigma_{X_G}^2$ [i.e., $P_{Short} \sim N(\mu_{X_S}, \sigma_{X_S}^2)$, $P_{Long} \sim N(\mu_{X_L}, \sigma_{X_L}^2)$, $P_G \sim N(\mu_{X_G}, \sigma_{X_G}^2)$]. When stimulus x is sampled from range short X_s, the local priors P_L (in Figure 7a and b) would be P_{Short}. Otherwise, when the stimulus x is sampled from range long X_L, then local prior P_L for current stimulus x would be P_{Long}.

1. Local-global model (LGM)

In the sensory measurement phase, hierarchical local-global model assumes that sensory measurement could first integrate with the local priors, and then integrate with the global prior to obtain the time estimation (see Figure 7a). The uni-prior model was adopted in the process of integration local prior with sensory measurement. When stimulus x is chosen from range short X_s, the local priors P_L (in Figure 7a) would be P_{Short}. In this case, the weight of local prior w_S indicating the dependency of prior in Bayesian time estimate $w_S = \frac{\sigma_s^2}{\sigma_{X_S}^2 + \sigma_s^2}$. We assume local prior P_L is the prior of the long range P_{Long} when the stimulus is sampled from range long X_L, therefore weight of local prior for the stimulus from range long w_L is proportional to the inverse of the variability of the local prior $\sigma_{X_L}^2$, $w_L = \frac{\sigma_s^2}{\sigma_{X_L}^2 + \sigma_s^2}$. Meanwhile, when stimulus is sampled from the range short X_S, the weight of local prior for the stimulus from range short w_S is proportional to the inverse of the variability of the local prior $\sigma_{X_S}^2$.

The posterior of integrating local priors P_L with sensory measurement [$D_s \sim N(\mu_s, \sigma_s^2)$] is denoted by D_L, $D_L \sim N(\mu_L, \sigma_L^2)$,

$$\mu_L(x) = \begin{cases} w_S \times \mu_{X_S} + (1 - w_S) \times \mu_s & x \in X_S \\ w_L \times \mu_{X_L} + (1 - w_L) \times \mu_s & x \in X_L \end{cases} \quad \text{(Eq.1)}$$

$$\sigma_L^2(x) = \begin{cases} \frac{\sigma_{X_S}^2 \sigma_s^2}{\sigma_{X_S}^2 + \sigma_s^2} & x_i \in X_S \\ \frac{\sigma_{X_L}^2 \sigma_s^2}{\sigma_{X_L}^2 + \sigma_s^2} & x_i \in X_L \end{cases}$$

and (Eq.2)

where $w_S = \frac{\sigma_s^2}{\sigma_{X_S}^2 + \sigma_s^2}$ and $w_L = \frac{\sigma_s^2}{\sigma_{X_L}^2 + \sigma_s^2}$indicates the weight of local prior in time estimation during the first stage of integration with sensory measurement.

Afterwards, the global prior P_{X_G} integrates with the output of integrating sensory measurement with local prior. The posterior of integration of global prior $P_{X_G} \sim N(\mu_{X_G}, \sigma_{X_G}^2)$ with posterior D_L, could be presented as$D_e \sim N(\mu_e, \sigma_e^2)$,which is the outcome of sensory measurement phase without consideration of motor noise,

$$\mu_e = w_G \mu_{X_G} + (1 - w_G)\mu_L \quad \text{(Eq.3)}$$

$$\sigma_e^2 = \frac{\sigma_{X_G}^2 \times \sigma_L^2}{\sigma_{X_G}^2 + \sigma_L^2} \quad \text{(Eq.4)}$$

where w_G indicates the weight of global prior in Bayesian time estimation $w_G = \frac{\sigma_L^2}{\sigma_{X_G}^2 + \sigma_L^2}$ and D'_r is the estimate of the reproduction without consideration of motor noise.

In motor reproduction phase, according to the previous studies (Shi, Church, et al., 2013; Shi & Burr, 2016), the distribution of estimates of the reproduction with the consideration of related motor noise [$D_m \sim N(0, \sigma_m^2)$] was assumed to be modeled as $D_r \sim N(\mu_r, \sigma_r^2)$ where

$$\mu_r = \mu_e \quad \text{(Eq.5)}$$

and $$\sigma_r^2 = \sigma_e^2 + \sigma_m^2 \quad \text{(Eq.6)}$$

2. Dual integration model (DIM)

The dual integration model (Figure 7b) assumes the sensory measurement could integrate with the local and global priors separately, then outputs from these integrations could be combined into the final estimate. In the sensory measurement phase, the posterior of integration local priors P_L with sensory measurement D_s is denoted by $D_L \sim N(\mu_L, \sigma_L^2)$, which has the same definition in the hierarchy global-local model, refer to Eq.1 and Eq.2. Meanwhile, the global prior P_G is known to be distributed normally with mean μ_{X_G} and variance $\sigma_{X_G}^2$, the

posterior of integration global prior P_G with sensory measurement D_s is denoted by D_G, where D_G is assumed to be Gaussian distribution $N(\mu_G, \sigma_G^2)$with mean $\mu_G$and variance σ_G^2,

$$\mu_G = w_g \mu_{X_G} + \left(1 - w_g\right)\mu_{X_G} \quad \text{(Eq.7)}$$

and

$$\sigma_G^2 = \frac{\sigma_{X_G}^2 * \sigma_s^2}{\sigma_{X_G}^2 + \sigma_s^2} \quad \text{(Eq.8)}$$

$w_g = \frac{1/\sigma_{X_G}^2}{1/\sigma_{X_G}^2 + 1/\sigma_s^2}$ indicates the dependency of global prior P_G in Bayesian estimation of D_G.

D_G and D_L are correlated because they both integrate with sensory measurement D_s, hence we adopt an parameter $\varphi(-1 \leq \varphi \leq 1)$[6] denoting the correlation between $D_G$and D_L, referring to combining correlated distribution theory (Oruç et al., 2003). The estimation in the dual integration model is denoted by D_e, which integrates $D_L \sim N(\mu_L, \sigma_L^2)$ and $D_G \sim N(\mu_G, \sigma_G^2)$. The outputs of integrating the sensory measurement with the local prior is denoted by $D_L \sim N(\mu_L, \sigma_L^2)$, meanwhile the posterior of integrating the sensory measurement with global priors is denoted by $D_G \sim N(\mu_G, \sigma_G^2)$. Furthermore, the observed reproduction is assumed to as a lognormal distribution $D_r \sim logN(\mu_e, \sigma_e^2 + \sigma_m^2)$,

$$\mu_e = w_e \mu_G + (1 - w_e)\mu_L \quad \text{(Eq.9)}$$

$$w_e = \frac{\sigma_L^2 - \varphi\,(\sigma_G \sigma_G)}{\sigma_G^2 + \sigma_L^2 - 2\varphi(\sigma_G \sigma_G)} \quad \text{(Eq.10)}$$

The reliability of D_G in time estimation is indicated by

$$\sigma_e^2 = \frac{(1 - \varphi^2)\sigma_G^2 \sigma_L^2}{\sigma_G^2 + \sigma_L^2 - 2\varphi(\sigma_G \sigma_G)} \quad \text{(Eq.11)}$$

In the motor reproduction phase, the sensory measured estimate D_e integrates motor noisy $\log(D_m) \sim N(0, \sigma_m^2)$ in the same way defined in the hierarchical global-local model, $D_r \sim logN(\mu_r, \sigma_r^2)$. The definition of μ_r and σ_r^2 refers to Eq.5 and Eq.6, respectively.

3. Prior integration model (PIM)

The prior integration model (Figure 7c) assumes that global priorP_G and local prior P_L are firstly integrated to conduct an integrated priorP_I, $P_I \sim N(\mu_P, \sigma_P^2)$, and the generated

---

[6] The correlation φ ranges from -1 to +1, where ±1 means completely positive correlation or negative correlation, and 0 means irrelevant.

integrated priorP_I is then integrated with sensory measurement D_s to form the estimation$D_e \sim N(\mu_e, \sigma_e^2)$. In more detail, in the sensory measurement phase, the posterior of integration local priors P_L with global priorP_G is denoted by P_I. The integrated prior P_I,$P_I \sim N(\mu_P, \sigma_P^2)$, is assumed to be distributed normally with mean μ_P and variance σ_P^2,

$$\mu_P(x) = \begin{cases} w_G \times \mu_{X_G} + (1 - w_G) \times \mu_{X_s} & x \in X_S \\ w_G \times \mu_{X_G} + (1 - w_G) \times \mu_{X_L} & x \in X_L \end{cases} \quad \text{(Eq.12)}$$

$$\sigma_P^2(x) = \begin{cases} \frac{\sigma_{X_G}^2 \sigma_{X_s}^2}{\sigma_{X_G}^2 + \sigma_{X_L}^2} & x_i \in X_S \\ \frac{\sigma_{X_G}^2 \sigma_{X_S}^2}{\sigma_{X_G}^2 + \sigma_{X_L}^2} & x_i \in X_L \end{cases} \quad \text{(Eq.13)}$$

The posterior of integrate the integrated prior P_I with sensory measurement D_s is denoted by D_e, where D_e is assumed to be Gaussian distribution $N(\mu_e, \sigma_e^2)$with mean $\mu_e$and variance σ_e^2,

$$\mu_e = w_P \mu_P + (1 - w_P)\mu_s \quad \text{(Eq.14)}$$

and

$$\sigma_e^2 = \frac{\sigma_P^2 * \sigma_s^2}{\sigma_P^2 + \sigma_s^2} \quad \text{(Eq.15)}$$

$w_P = \frac{1/\sigma_P^2}{1/\sigma_P^2 + 1/\sigma_s^2}$ indicates the dependency of integrated prior P_I in Bayesian estimation.

Appendix B. Model comparison

One basic assumption of the prior integration introduced in this paper is Weber scaling, which follows Weber-Fechner's law. According to Fechner's law, internal representation is logarithmically related to the external stimulus intensity. An alternative assumption of the internal representation is linear internal coding, according to Weber's law that the uncertainty of the inner perception linearly increases with the absolute intensity. Numerous previous studies revealed that distinguishing between the linear and logarithmic internal representation is largely constrained by the adopted experimental paradigms (Matthews & Meck, 2016; Ren et al., 2021b), however, there is no existing method to determine which internal representation (either in logarithmic or linear scale) is better for current experimental paradigms. Therefore, we conducted both logarithmic and linear coding models and compared their prediction results to find the best model for time reproduction prediction.

According to scalar variability property that the standard deviations of sensory measurement are linearly increasing with the mean of stimuli (Acerbi et al., 2012; Gibbon et al., 1984b), the standard deviation of the priors could be denoted by$\sigma_{X_S} = k \times \mu_{X_S}$, $\sigma_{X_L} = wf \times \mu_{X_L}$, $\sigma_{X_G} = k \times \mu_{X_G}$, here k is known as *Weber's fraction* of sensory measurements.

Table A1 Average results of WAIC, LOOIC and prediction accuracy (AUC) of mean and standard variance the reproductions by fitting measured behavioral data to LGM, DIM, PIM models.

		Exp1			Exp2		
		LGM	DIM	PIM	LGM	DIM	PIM
logarithmic model	reproduction AUC	97.416%	98.079%	97.422%	97.497%	97.794%	97.493%
	CV AUC	56.272%	51.923%	56.266%	72.698%	68.463%	72.701%
	WAIC	-67.251	-73.728	-67.222	-91.803	-92.864	-91.852
	LOOIC	51.45	45.160	52.007	27.023	26.335	26.378
linear model	reproduction AUC	96.855%	96.747%	96.860%	97.305%	97.031%	97.308%
	CV AUC	54.779%	48.048%	54.778%	69.031%	64.295%	69.023%
	WAIC	-74.918	-107.318	-74.979	-107.722	-128.202	-107.703
	LOOIC	44.569	11.172	43.744	10.937	-9.920	11.0

To find the best model for the observed time reproduction prediction, Table A1 presents predicted error of mean and standard variance in the models and the average WAIC and LOOIC across all subjects. As shown in Table A1, the results from DIM in the logarithmic scale had the lowest WAICs and LOO-CV, suggesting the proposed logarithmic model can better predict duration reproduction biases than other models.

3 General Discussion

The present dissertation investigated mechanisms underlying subjective interval timing from Bayesian inference perspective, focusing on Bayesian modeling of contextual bias in the temporal estimation process. The main research objective of this accumulative work was to examine the contextual bias under three types of temporal contexts in subjective interval timing: temporal ensemble context in a bisection task (where ensemble set properties were manipulated by varying in stimulus spacing, set mean, and set variance), duration reproduction under memory pressure (visual working-memory cognitive load), and duration reproduction under multi-prior temporal context.

The first study (Chapter 2.1) focused on the temporal ensemble context in temporal bisection and we developed a unified ensemble-distribution account that assumes that the mean and variance of the duration set serve as the main reference, rather than the short and long standards, in duration comparison. We measured the influence of set properties on PSEs and JNDs by varying in stimulus spacing, set mean, and set variance in three experiments , and examined whether the mean of the stimuli set accounts for the shift of bisection points and whether the variance of the stimuli set accounts for change of the sensitivity of temporal judgments. Previous studies on spacing account have shown the logarithmic versus linear spacing of the probe durations determined the shift of bisection point (Brown et al., 2005; Penney et al., 2014). Our study revealed the ensemble means of the stimulus set is a critical factor accounting for shifts of the bisection. Our findings go beyond this by indicating that variance information also plays an important role in temporal judgments, because the variance of the sample distribution includes useful information for discerning the location of a probe duration relative to the ensemble mean, thus enhancing temporal sensitivity. To investigate whether temporal-bisection judgments would be best explained by ensemble-distribution account, we proposed and implemented a hierarchy model that explains in which way subjective judgments of time intervals vary according to the distribution summary statistics of the set mean and variance values.

The second study (Chapter 2.2) addressed the question of whether the central tendency would be differentially influenced by the cognitive load on the encoding and reproduction phases in duration judgments. Four dual, attention-sharing tasks were carried out to investigate the impact of cognitive load on duration estimation in the duration encoding and reproduction

stages. By integrating an attentional-sharing account into a hierarchical Bayesian model, we proposed a generative model that simulates cognitive load influence on the perceived and reproduced duration in both the mean and the variability. The proposed model was able to predict both the general over- and underestimation and the central-tendency effects observed in all four experiments.

In the third study (Chapter 2.3), we concentrated on the question of how multiple prior temporal contexts influence subjective duration timing in duration reproduction tasks. We found the existence of a global prior and investigated how global prior plays its role in multiple prior contexts. Aiming to explore a hierarchical structure of the priors integration which may better predict the behavioral results than the previous uni-prior observer model, we suggested three possible hierarchical structures of global and local priors integration in Bayesian estimation. Using the Bayesian inference framework, the question of how our brain learns the multiple modality-specific priors(local priors) through training in block-wise ranges before interleaved range, and uses local priors together with a global prior in multi-prior temporal contexts can be simulated and investigated.

I will briefly summarize the results of the three studies, discuss how they contribute to the current framework and point out potential future directions.

3.1 Summary of results

3.1.1 Ensemble perception for temporal sequence

The first study (Chapter 2.1) focused on influences of the distribution of sampled intervals in the temporal-bisection tasks and developed a unified ensemble-distribution account that assumes that the mean and variance of the duration set serve as the main reference, rather than the short and long standards, in duration comparison. It demonstrated that the bisection points of sample intervals were biased towards the mean duration of the whole set of auditory signals in the memory traces (i.e., prior in long-term memory). We tested set properties by varying in stimulus spacing, set mean, and set variance in three experiments, to investigate whether temporal-bisection judgments depend on the shape of the testing interval distribution.

In the first study, three experiments were carried out by manipulating the distribution of auditory duration sets to determine factors that influence temporal-bisection performance.

Adopting the similar spacing samples as Penney et al. (2014), Experiment 1 replicated their spacing effect, as expected, to examine for the shift of the bisection point in sets with positively skewed (PS) versus negatively skewed (NS) spacing. To further examine the temporal-bisection task with equally spaced durations but different sampled frequency (thus different ensemble means), Experiment 2 applied two skewed frequency-distribution sets: ascending frequency (AF) and descending frequency (DF). The results of Experiment 2 revealed differential bisection points, suggesting that merely spacing cannot explain the findings. Rather, using ensemble mean as the comparison interval can explain both findings from Experiments 1 and 2. To further corroborate the ensemble summary plays a critical role in the temporal bisection, we further tested if the second moment of the distribution also matters in Experiment 3. Specifically, Experiment 3 manipulated the variability of the sample distributions while keeping the mean of the distributions the same, by introducing a U-shaped and an inverted U-shaped sets, with the former having a greater variance than the latter, but both sampled distributions had the same ensemble means. As predicted by the ensemble distribution account (EDA), we found the bisection points in two conditions were comparable, but the discrimination sensitity of the bisection task was worse in the U-shape condition as compared to the inverted U-shape condition. Additionally, we applied hierarchical Bayesian modeling to the behavioral data according to various assumptions of how temporal bisection may be performed. The aim of the model fitting and comparison was to look at the data patterns obtained in the three experiments with respect to the manipulations of stimulus spacing, distribution means, and variances, so as to identify the best possible account of how the ensemble context modulates performance of the bisection task.

The results of the first study revealed the mean and variance of the stimulus set to be critical factors in the bisection task. Specifically, during the bisection task participants were not comparing the sampled interval to the standard Short or Long, but rather comparing to the mean of experienced intervals (i.e., the ensemble mean). In addition, the uncertainty of the comparison came from the representation of the ensemble statistics. When the sampled intervals had a large variability, the performance of the bisection task became more uncertain as compared to the sampled intervals with a low variability. These findings demonstrate that we automatically use ensemble statistics in time perception which works similarly to other perceptual properties in the visual and auditory domains. Thus, the EDA framework can explain sampled contextual biases in temporal judgments of time.

3.1.2 Cognitive load in central tendency bias

The second study (Chapter 2.2) addressed the question of whether the central tendency would be differentially influenced by the cognitive load on the encoding and reproduction phases in duration judgments. We used the duration production-reproduction task as a primary task with a secondary color memory task in four experiments to investigate the impact of cognitive load on duration estimation. A typical duration production-reproduction task has two phases: duration encoding phase and duration reproduction phase. In the duration encoding phase, an interval is automatically presented via a presentation of a stimulus, whereas in the duration reproduction phase, participants are asked to reproduce that duration by pressing a key to generate a stimulus as long as the perceived one. Experiment 1 served as a baseline that the duration task and the memory task were sequentially separated, such that the memory task had a minimal impact on the duration task. The potential impact of the secondary task in Experiment 1 was prior updating, given that the secondary task was 'inserted' in between the duration tasks. The memory task was imposed on the duration encoding stage in Experiment 2, whereas the memory task overlapped with the duration reproduction stage in Experiment 3. We aimed to examine the differential roles of memory in the two stages. In the final Experiment 4, the memory task was spanned over the whole duration production-reproduction task to test how memory pressure on both stages influences the duration reproduction.

According to an attentional-sharing account (Fortin, 2003; Fortin & Rousseau, 1998; Macar et al., 1994) and Bayesian inference in time estimation (Jazayeri & Shadlen, 2010; Shi, Church, et al., 2013; Shi & Burr, 2016), we expected that cognitive load would influence the perceived and reproduced duration in both the mean and the variability. In Experiment 1, we found that varying levels of memory load between consecutive duration reproductions had no discernible effects on either general reproduction biases or the central-tendency bias, which suggests that the prior updating of the sampled duration was not influenced by the intermediate secondary task. With Experiments 2 and 3, we found the impact of cognitive load in duration reproduction works in a different way when the working-memory load was imposed during the duration-encoding (production) and the reproduction phase. To be more specific, when the working-memory task overlapped only with the production phase (Experiment 2), participants underestimated the target durations, and a stronger central-tendency effect was shown under higher memory-load conditions. By contrast, when the working-memory task spanned only the reproduction phase (Experiment 3), participants over-reproduced more under the higher the

memory load condition. These findings can be qualitatively explained by the attentional-sharing account (Fortin, 2003; Macar et al., 1994). When attention is shared in the duration encoding stage, some 'ticks' may be lost in the accumulating process, such that the reproduced duration is shortened. In contrast, when the attention is shared in the duration reproduction phase, 'ticks' lost in the monitoring of the passage time would 'lengthen' the reproduction as it requires additional time to compensate those lost 'ticks'. Interestingly, we found the absence of significant shifts in the reproduced duration because the impacts of the general shifts roughly canceled each other out when the working-memory task spanned both the production and reproduction phases (in Experiment 4), but the central-tendency effect remained stronger with higher vs. lower memory loads.

To achieve a generative model for the results from four experiments, we proposed a hierarchical Bayesian model taking memory load interference on both duration encoding and reproduction phases into consideration. More specifically, following the classical internal clock models, loss of pacemaker generated ticks caused by memory pressure during the duration encoding phase leads to underestimation and increases the sensory noise. In addition, monitoring the elapsed time can be disrupted by the memory load during the duration reproduction phase. Moreover, uncertainty of reproduction is further influenced by motor noise. The model prediction not only showed a good agreement with the reproduction behavior for individual participants, but also predicted the mean and the coefficient of variation of reproduction behavior. Of note, the pattern of findings was well fitted by our proposed Bayesian inference model which combined the attentional-sharing account with the standard Bayesian inference model. The dissociable influences of memory pressure in the encoding and reproduction stages was explained as the loss of time units reported in previous studies (Fortin, 2003; Fortin & Rousseau, 1998). More specifically, when attention is diverted away from the primary (temporal) task by a concurrent non-temporal task during the duration-encoding phase, a certain amount of clock ticks would be lost, resulting in a shortened time estimation (underestimation). In contrast, when the secondary task is performed concurrently with the reproduction phase, due to lapses in monitoring the elapsed time there would lead to a loss of clock ticks during the reproduction phase which results in a longer reproduction than the tested interval (overestimation). This descriptive explanation is quantitatively characterized by the linear scaling parameters in our Bayesian model. Both the behavioral findings and the model confirm differential impacts of the loss of time units of memory pressure at different stages of time estimation.

3.1.3 Prior integration in multi-prior temporal context

In the third study (Chapter 2.3), we concentrated on the underlying structure of the prior integration in the multi-prior context and the question of how multiple priors influence subjective duration timing in duration reproduction tasks. Two duration reproduction experiments with two levels of variance of short and long range on logarithmic scale and the same geometric mean of interleaved range were carried out to uncover the underlying mechanism during prior knowledge integration. We examined whether the subjective estimation of duration is influenced by priors learned from previous training block ranges (intervals from different ranges are presented blockwise) and illustrated the existence of a global prior which takes place in time estimation in multiple prior contexts.

To investigate the questions of how multiple prior knowledge integrates together with the sensory inputs to form the estimation, we extended our previous uni-prior Bayesian model(Shi, Church, et al., 2013; Shi & Burr, 2016) by taking a global prior generalizing across the interleaved range into consideration. Based on an assumption that prior representation is modality-specific, priors based on the distinct stimulus set were treated as distinct modality to investigate the interaction between priors. To uncover potential prior integration structure, we proposed three alternative structures for integration of local prior, global prior and sensory inputs, including the local-global model (LGM, which integrates the sensory measurement with the local prior, then further integrates with the global prior to yield the posterior), prior integration model (PIM, which global prior and local prior are integrated as a integrated prior, then the integrated prior integrates with sensory measurement to form the estimated duration) and dual integration model (DIM, which global and local priors integrate with the sensory measurement separately firstly, then outputs from these integrations combine together to yield the posterior). These three models mimicked three possibilities of integration structures, and the model results showed they performed equally well in reproduction mean and variance prediction, despite their different structural assumptions. According to the parameter estimation results derived from models, we found that in the temporal context with smaller variability of the interleaved range the time estimates would count more on the prior knowledge and less on sensory inputs. Besides, the order of the training session influences the formation of global prior, since context farther in the past is less weighted than more recent context. When the short session was taken as the first training session, observers acquired a higher mean of global prior, becaused the influence of the long session in the near past is weighted more.

3.2 Future directions

Using Bayesian inference framework and the hierarchical modeling approach, the current dissertation aimed to understand three specific temporal contexts in duration judgment: ensemble sampled distributions, memory pressure, and global ranges. However, there are still some inherent limitations that current Bayesian computational frameworks can not achieve. Firstly, our current Bayesian models can well predict an optimal fusion of an experience-dependent prior expectation with the current noisy sensory measurement, and estimate the parameters for individual participants by maximizing the joint posterior distribution of parameters of interest. Those models are steady-state models. How our brain reaches such a steady-state level remains open. It requires trial-to-trial simulation of the decision processing of timing. Therefore, further studies are required to develop a new methodology for sequential state and parameter estimation combining with the recursive Bayesian estimation (Kalman filter) in Bayesian estimation to obtain the trial-wise simulation of timing processing. The behavior of the observers could be better simulated by an iterative Bayesian filtering framework (Doucet et al., 2000; Särkkä, 2013) where estimates are derived from the current noisy measurement merged with updating prior on the trial by trial basis. Furthermore, Simple Bayesian observer models, like we implemented in the studies, often assume Gaussian distribution (unimodal distribution), which is lacking its ability to account for bimodality prior distributions (Acerbi et al., 2014; Sanborn & Beierholm, 2016). It remains challenging to select priors in Bayesian inference. When the assumption of the prior is too flexible, there is a risk that Bayesian inference becomes merely fitting the empirical data (M. Jones & Love, 2011). Therefore, given that our models were constrained to use normal distribution for prior and sensory measurement during the Bayesian estimate process, it remains to improve for future studies to select proper priors, which are flexible yet have biological implications.

Moreover, Studies 2 and 3 of the current dissertation focus on Bayesian estimation of the central tendency bias in reproduction tasks. Although we mentioned above that the perceptual process is well-explained as a process in that humans learn statistical regularities and exploit this knowledge to improve perceptual decisions, the nature of the precise timing mechanism itself is still an open question. Despite decades of research, the process of how people perceive and process timing remains relatively unclear. Taking the second study as an example, we observed the influence of the memory load on the central tendency was more pronounced in Experiment 2 (working memory task overlapped the production phase but not the reproduction phase) than in Experiment 4 (working memory task was extended across both

the duration-encoding and reproduction phases), which was captured as well by a scaling parameter that denotes the sensory-measurement uncertainty. The estimated result of this parameter from our proposed model was significantly larger in Experiment 2 than in Experiment 4. This finding raised a question if order of the primary and secondary tasks was related to the proportion of allocating attention, and we hypothesized that more attentional resources were likely allocated to the first than to the second task. In Experiment 2, the secondary memory task was the first task requiring a response, whereas duration reproduction was the first task and the secondary memory task was the second task in Experiment 4. Given that we did not record the completion time of the memory test in between the production and reproduction phases, we cannot quantitatively determine the impact of the prolonged gap on the central tendency bias. Thus, further investigation is required to address these remaining questions.

Finally, Study 1 of the dissertation illustrated that set properties play a major decisive role in influence of bisection points and sensitivity during temporal bisection judgments. However, it remains unknown whether only these set properties as a comparison interval or two distinct time traces of the current and comparison interval simultaneously contribute to the decision process. Thus, it is worth exploring the neural evidence of the timing mechanism using event-related electroencephalogram (EEG) techniques, specifically the contingent negative variation (CNV) which manifests itself in the supplementary motor area (SMA) and often is taken as an index of decision making and temporal memory (Kononowicz & van Rijn, 2014; Macar & Vidal, 2003). If we could find the differences on the CNV amplitudes for the same target intervals but from different temporal contexts, that could be strong evidence to show how temporal judgment was influenced by two temporal contexts in neural underpinnings and uncover relations between the CNV changes with ensemble means of interval set during a temporal bisection task. Also, the findings in Study 3 of the dissertation suggests that a general global prior in addition to local priors affects participants' behavior in the multi-prior temporal context in duration reproduction tasks, it remains lack of neural evidence supporting of Bayesian integration in timing (Ma et al., 2008; Singletary et al., 2021). A further study of the neural substrates of ensemble processing under multi-prior temporal context by exploring the CNV activity is required.

3.3 Conclusions

Taken above three studies together, the current dissertation investigated contextual biases in subjective interval timing with the Bayesian inference framework and hierarchical Bayesian modeling approach.

Previous studies had shown that duration itself and stimulus spacing affects bisection points in temporal-bisection judgments, but in which way subjective judgments of time intervals vary according to spacing and ensemble statistics need to be explained. The first study (Chapter 2.1) reported three experiments that varied the probed duration spacing (Experiment 1), the frequency of the probed duration (Experiment 2), and the variability of the probed duration (Experiment 3) to determine factors that influence the temporal-bisection performance, and proposed a unified ensemble-distribution account (EDA) to explain the contextual effect in temporal bisection. Our findings suggested that not only the set properties (i.e., set mean and set variance) play a major decisive role in the influence not only on bisection points but also sensitivity during temporal-bisection judgments. Comparing EDA to the extant prior accounts showed that ensemble statistics can easily explain how the effect of various stimulus set-related properties (e.g., spacing, frequency, variance, skewness) are actually combined in temporal judgments. These findings demonstrated that the human timing mechanism involves an ensemble averaging process, similarly to other ensemble perceptual properties in the visual and auditory domains.

The second study (Chapter 2.2) examined the impact of cognitive (visual working-memory) load on duration estimation in the duration encoding and reproduction phases through four dual experiments (with duration reproduction task as the primary task and working memory task as the secondary task). Our findings suggested more memory load press (i.e., set size) during the duration-encoding phase increased the central-tendency bias and shortened the reproduced duration; in contrast, more memory load press during the reproduction phase increased the reproduced duration but without influencing the central tendency. Additionally, by integrating an attentional-sharing account into a hierarchical Bayesian model, the shifts of reproduction observed in all four experiments were well predicted. The model results further suggested that the central-tendency bias was only influenced by the memory pressure in the encoding stage. Memory pressure during the encoding stage increases the sensory noise, which elevates the central-tendency effect. In contrast, memory pressure during the reproduction stage only influences the monitoring of elapsed time, leading to a general shift of reproduction

without impacting the central tendency. Last but not the least, the generative model we proposed here for the influence of the cognitive load on time perception might be generalizable to other forms of magnitude perception.

The third study (Chapter 2.3) focused on investigating how the hierarchical Bayesian estimation structure works in multiple prior temporal estimation. Two duration reproduction experiments with two levels of variance of blocked range on logarithmic scale and the same geometric mean of interleaved range were carried out to uncover the underlying mechanism during multiple prior integration. The participants were firstly trained with the short and long intervals session-wise, and then tested with interleaved ranges randomly inter-mixed, to examine potential integration hierarchies. Three hierarchical prior integration models, assuming estimation to be an optimal integration of the noisy sensory inputs with multiple priors (short prior, long prior and global prior) in a hierarchical order, were proposed to uncover the governing rules. The model results showed their performance were equally well performance in reproduction mean and variance prediction despite their different structural assumptions. We found that the order of the short session and long session influences the formation of global prior. Taking the short session as the last training session before the mix session increased the estimation of global prior, in contrast, taking short session prior to the mix session decreased the estimation of global prior, which was illustrated by the parameter estimation results derived from models. Additionally, for the experiment with smaller variance on log scale by narrowing short and long range and expanding separation between short and long-range, the estimates would rely more on the global prior information and less on sensory inputs.

References (General Introduction and General Discussion)

Acerbi, L., Vijayakumar, S., & Wolpert, D. M. (2014). On the origins of suboptimality in human probabilistic inference. *PLoS Computational Biology*, *10*(6), e1003661.

Allan, L. G., & Gibbon, J. (1991). Human bisection at the geometric mean. *Learning and Motivation*, *22*(1), 39–58.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131.

Angrilli, A., Cherubini, P., Pavese, A., & Mantredini, S. (1997). The influence of affective factors on time perception. *Perception & Psychophysics*, *59*(6), 972–982.

Ariely, D. (2001). Seeing sets: Representation by Statistical Properties. *Psychological Science*, *12*(2), 157–162.

Bausenhart, K. M., Dyjas, O., & Ulrich, R. (2014). Temporal reproductions are influenced by an internal reference: explaining the Vierordt effect. *Acta Psychologica*, *147*, 60–67.

Berniker, M., Voss, M., & Kording, K. (2010). Learning priors for Bayesian computations in the nervous system. *PloS One*, *5*(9). https://doi.org/10.1371/journal.pone.0012686

Block, R. A., Hancock, P. A., & Zakay, D. (2010). How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychologica*, *134*(3), 330–343.

Brown, G. D. A., McCormack, T., Smith, M., & Stewart, N. (2005). Identification and bisection of temporal durations and tone frequencies: common models for temporal and nontemporal stimuli. *Journal of Experimental Psychology. Human Perception and Performance*, *31*(5), 919–938.

Buhusi, C. V., & Meck, W. H. (2009). Relativity theory and time perception: single or multiple clocks? *PloS One*, *4*(7), e6268.

Chen, L., Zhou, X., Müller, H. J., & Shi, Z. (2018). What you see depends on what you hear: Temporal averaging and crossmodal integration. *Journal of Experimental Psychology. General*, *147*(12), 1851–1864.

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, *45*(7), 891–900.

Cicchini, G. M., Arrighi, R., Cecchetti, L., Giusti, M., & Burr, D. C. (2012). Optimal encoding of interval timing in expert percussionists. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(3), 1056–1060.

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the Bandwidth of Perceptual Experience? *Trends in Cognitive Sciences*, *20*(5), 324–335.

Coull, J. T., Davranche, K., Nazarian, B., & Vidal, F. (2013). Functional anatomy of timing differs for production versus prediction of time intervals. *Neuropsychologia*, *51*(2), 309–319.

Curtis, D. W., & Mullin, L. C. (1975). Judgments of average magnitude: Analyses in terms of the functional measurement and two-stage models. *Perception & Psychophysics*, *18*(4), 299–308.

Doucet, A., Godsill, S., & Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, *10*(3), 197–208.

Dyjas, O., Bausenhart, K. M., & Ulrich, R. (2013). Effects of Stimulus Order on Duration Discrimination Sensitivity are Under Attentional Control. *Journal of Experimental Psychology. Human Perception and Performance*. https://doi.org/10.1037/a0033611

Eagleman, D. M. (2008). Human time perception and its illusions. *Current Opinion in Neurobiology*, *18*(2), 131–136.

Erickson, M. H., & Erickson, E. M. (1958). Further Considerations of Time Distortion: Subjective Time Condensation as Distinct from Time Expansion. *The American Journal*

of Clinical Hypnosis, *1*(2), 83–88.

Fechner, G. T. (1966). Elements of psychophysics [elemente der Psychophysik.(New York, Holt, Rinehart and Winston:).[First published. 1860. *Editors DH Howes and EG Boring*, *1*.

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience*, *17*(5), 738–743.

Fortin, C. (2003). Attentional time-sharing in interval timing. In W. H. Meck (Ed.), *Functional and neural mechanisms of interval timing* (pp. 235–260). CRC Press.

Fortin, C., & Massé, N. (2000). Expecting a break in time estimation: attentional time-sharing without concurrent processing. *Journal of Experimental Psychology. Human Perception and Performance*, *26*(6), 1788–1796.

Fortin, C., & Rousseau, R. (1998). Interference from short-term memory processing on encoding and reproducing brief durations. *Psychological Research*, *61*(4), 269–276.

Fritsche, M., Spaak, E., & de Lange, F. P. (2020). A Bayesian and efficient observer model explains concurrent attractive and repulsive history biases in visual perception. *eLife*, *9*. https://doi.org/10.7554/eLife.55389

Gau, R., & Noppeney, U. (2016). How prior expectations shape multisensory perception. In *NeuroImage* (Vol. 124, pp. 876–886). https://doi.org/10.1016/j.neuroimage.2015.09.045

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psychological Review*, *84*(3), 279–325.

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

Glasauer, S., Schneider, E., Grasso, R., & Ivanenko, Y. P. (2007). Space-time relativity in self-motion reproduction. *Journal of Neurophysiology*, *97*(1), 451–461.

Glasauer, S., & Shi, Z. (2021a). The origin of Vierordt's law: The experimental protocol

matters. *PsyCh Journal*, *pchj.464*. https://doi.org/10.1002/pchj.464

Glasauer, S., & Shi, Z. (2021b). Individual believes about temporal continuity explain perceptual biases. In *bioRxiv* (p. 2021.07.13.452167). https://doi.org/10.1101/2021.07.13.452167

Gu, B. M., Jurkowski, A. J., Shi, Z., & Meck, W. H. (2016). Bayesian optimization of interval timing and biases in temporal memory as a function of temporal context, feedback, and dopamine levels in young, aged and …. *Timing & Time Perception*. https://brill.com/view/journals/time/4/4/article-p315_1.xml

Gu, B. M., & Meck, W. H. (2011). New perspectives on Vierordt's law: memory-mixing in ordinal temporal comparison tasks. In A. Vatakis, A. Esposito, M. Giagkou, F. Cummins, & G. Papadelis (Eds.), *Multidisciplinary Aspects of Time and Time Perception* (Vol. 6789, pp. 67–78).

Haberman, J., & Whitney, D. (2009). Seeing the mean: ensemble coding for sets of faces. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(3), 718–734.

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. *From Perception to Consciousness*. https://books.google.com/books?hl=en&lr=&id=Kw9pAgAAQBAJ&oi=fnd&pg=PA339&dq=Ensemble+Perception+summarizing+scene+broadening+limits+visual+processing+Haberman+Whitney&ots=a9drUmc0wz&sig=eV7LYUmDHdaxk9eshuEMyNzN9mw

Helson, H. (1964). Current trends and issues in adaptation-level theory. *The American Psychologist*, *19*(1), 26–38.

Heron, J., Aaen-Stockdale, C., Hotchkiss, J., Roach, N. W., McGraw, P. V., & Whitaker, D. (2012). Duration channels mediate human time perception. *Proceedings. Biological Sciences / The Royal Society*, *279*(1729), 690–698.

Hoffman, D. D., Singh, M., & Prakash, C. (2015). The Interface Theory of Perception. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-015-0890-8

Holland, M. K., & Lockhead, G. R. (1968). Sequential effects in absolute judgments of loudness. *Perception & Psychophysics*, *3*(6), 409–414.

Hollingworth, H. L. (1910). The Central Tendency of Judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, *7*(17), 461–469.

Ivry, R. B., & Richardson, T. C. (2002). Temporal control and coordination: the multiple timer model. *Brain and Cognition*, *48*(1), 117–132.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*, *13*(8), 1020–1026.

Jones, L. A., & Wearden, J. H. (2004). Double standards: Memory loading in temporal reference memory. In *The Quarterly Journal of Experimental Psychology Section B* (Vol. 57, Issue 1b, pp. 55–77). https://doi.org/10.1080/02724990344000088

Jones, M., & Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *The Behavioral and Brain Sciences*, *34*(4), 169–188; disuccsion 188–231.

Kanai, R., & Watanabe, M. (2006). Visual onset expands subjective time. *Perception & Psychophysics*, *68*(7), 1113.

Kiyonaga, A., Scimeca, J. M., Bliss, D. P., & Whitney, D. (2017). Serial Dependence across Perception, Attention, and Memory. *Trends in Cognitive Sciences*, *21*(7), 493–497.

Kononowicz, T. W., & van Rijn, H. (2014). Decoupling interval timing and climbing neural activity: a dissociation between CNV and N1P2 amplitudes. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *34*(8), 2931–2939.

Lee, G. (2017). Making sense of subjective time. In *The Routledge handbook of philosophy of temporal experience* (pp. 157–168). Routledge.

Lejeune, H., & Wearden, J. H. (2009). Vierordt's The Experimental Study of the Time Sense (1868) and its legacy. *The European Journal of Cognitive Psychology*, *21*(6), 941–960.

Luu, L., & Stocker, A. A. (2018). Post-decision biases reveal a self-consistency principle in perceptual inference. *eLife*, *7*. https://doi.org/10.7554/eLife.33334

Macar, F., Grondin, S., & Casini, L. (1994). Controlled attention sharing influences time estimation. *Memory & Cognition*, *22*(6), 673–686.

Macar, F., & Vidal, F. (2003). The CNV peak: an index of decision making and temporal memory. *Psychophysiology*, *40*(6), 950–954.

Marchant, A. P., Simons, D. J., & de Fockert, J. W. (2013). Ensemble representations: effects of set size and item heterogeneity on average size perception. *Acta Psychologica*, *142*(2), 245–250.

Ma, W. J., Beck, J. M., & Pouget, A. (2008). Spiking networks for Bayesian inference and choice. *Current Opinion in Neurobiology*, *18*(2), 217–222.

Parducci, A. (1963). Range-frequency compromise in judgment. *Psychological Monographs: General and Applied*, *77*(2), 1–50.

Paton, J. J., & Buonomano, D. V. (2018). The Neural Basis of Timing: Distributed Mechanisms for Diverse Functions. *Neuron*, *98*(4), 687–705.

Penney, T. B., Brown, G. D. A., & Wong, J. K. L. (2014). Stimulus spacing effects in duration perception are larger for auditory stimuli: data and a model. *Acta Psychologica*, *147*, 97–104.

Penney, T. B., Gibbon, J., & Meck, W. H. (2000). Differential Effects of Auditory and Visual Signals on Clock Speed and Temporal Memory. *Journal of Experimental Psychology. Human Perception and Performance*, *26*(6), 1770.

Petzschner, F. H., & Glasauer, S. (2011). Iterative Bayesian estimation as an explanation for range and regression effects: a study on human path integration. *The Journal of*

Neuroscience: The Official Journal of the Society for Neuroscience, *31*(47), 17220–17229.

Petzschner, F. H., Maier, P., & Glasauer, S. (2012a). Combining symbolic cues with sensory input and prior experience in an iterative Bayesian framework. *Frontiers in Integrative Neuroscience*. https://www.frontiersin.org/articles/10.3389/fnint.2012.00058/full

Petzschner, F. H., Maier, P., & Glasauer, S. (2012b). Combining symbolic cues with sensory input and prior experience in an iterative bayesian framework. *Frontiers in Integrative Neuroscience*, *6*, 58.

Piazza, E. A., Sweeny, T. D., Wessel, D., Silver, M. A., & Whitney, D. (2013). Humans use summary statistics to perceive auditory sequences. *Psychological Science*, *24*(8), 1389–1397.

Polti, I., Martin, B., & van Wassenhove, V. (2018). The effect of attention and working memory on the estimation of elapsed time. *Scientific Reports*, *8*(1), 6690.

Ren, Y. (2020). *Context Effects in Interval Timing*. München, Ludwig-Maximilians-Universität.

Ren, Y., Allenmark, F., Müller, H. J., & Shi, Z. (2021). Variation in the "coefficient of variation": Rethinking the violation of the scalar property in time-duration judgments. *Acta Psychologica*, *214*, 103263.

Ren, Y., Müller, H. J., & Shi, Z. (2020). Ensemble perception in the time domain: evidence in favor of logarithmic encoding of time intervals. In *Scientific Report* (p. 2020.01.25.919407). https://doi.org/10.1101/2020.01.25.919407

Roach, N. W., McGraw, P. V., Whitaker, D. J., & Heron, J. (2017). Generalization of prior information for rapid Bayesian time estimation. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(2), 412–417.

Sackett, A. M., Meyvis, T., Nelson, L. D., Converse, B. A., & Sackett, A. L. (2010). You're

having fun when time flies: the hedonic consequences of subjective time progression. *Psychological Science*, *21*(1), 111–117.

Sanborn, A. N., & Beierholm, U. R. (2016). Fast and Accurate Learning When Making Discrete Numerical Estimates. *PLoS Computational Biology*, *12*(4), e1004859.

Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press.

Schweickert, R., Han, H. J., Yamaguchi, M., & Fortin, C. (2014). Estimating averages from distributions of tone durations. *Attention, Perception & Psychophysics*, *76*(2), 605–620.

Shi, Z., & Burr, D. (2016). Predictive coding of multisensory timing. *Current Opinion in Behavioral Sciences*, *8*, 200–206.

Shi, Z., Church, R. M., & Meck, W. H. (2013). Bayesian optimization of time perception. *Trends in Cognitive Sciences*, *17*(11), 556–564.

Shi, Z., Ganzenmüller, S., & Müller, H. J. (2013). Reducing bias in auditory duration reproduction by integrating the reproduced signal. *PloS One*, *8*(4), e62065.

Singletary, N. M., Gottlieb, J., & Horga, G. (2021). A neural substrate for Bayesian integration in human parietal cortex. In *bioRxiv* (p. 2021.10.30.466508). https://doi.org/10.1101/2021.10.30.466508

Stanfield-Wiswell, C., & Wiener, M. (2020). The effect of an unexpected modality on time reproduction: Clock speed or memory mixing? In *PsyArXiv*. https://doi.org/10.31234/osf.io/vbqrj

Stetson, C., Fiesta, M. P., & Eagleman, D. M. (2007). Does time really slow down during a frightening event? *PloS One*, *2*(12), e1295.

Teghtsoonian, R., & Teghtsoonian, M. (1978). Range and regression effects in magnitude scaling. *Perception & Psychophysics*, *24*(4), 305–314.

Teixeira, S., Machado, S., Paes, F., Velasques, B., Silva, J. G., Sanfim, A. L., Minc, D., Anghinah, R., Menegaldo, L. L., Salama, M., Cagy, M., Nardi, A. E., Pöppel, E., Bao,

Y., Szelag, E., Ribeiro, P., & Arias-Carrión, O. (2013). Time perception distortion in neuropsychiatric and neurological disorders. *CNS & Neurological Disorders Drug Targets*, *12*(5), 567–582.

Treisman, M. (1963). Temporal discrimination and the indifference interval: Implications for a model of the“ internal clock.” *Psychological Monographs: General and Applied*, *77*(13), 1.

Treisman, M., Faulkner, A., Naish, P. L., & Brogan, D. (1990). The internal clock: evidence for a temporal oscillator underlying time perception with some estimates of its characteristic frequency. *Perception*, *19*(6), 705–743.

van Wassenhove, V., Buonomano, D. V., Shimojo, S., & Shams, L. (2008). Distortions of subjective time perception within and across senses. *PloS One*, *3*(1), e1437.

Vierordt, K. (1868). *Der Zeitsinn nach Versuchen*. H. Laupp.

Watamaniuk, S. N., & Duchon, A. (1992). The human visual system averages speed information. *Vision Research*, *32*(5), 931–941.

Wearden, J. H. (1991). Human performance on an analogue of an interval bisection task. *The Quarterly Journal of Experimental Psychology. B, Comparative and Physiological Psychology*, *43*(1), 59–81.

Wearden, J. H., Edwards, H., Fakhri, M., & Percival, A. (1998). Why “sounds are judged longer than lights”: application of a model of the internal clock in humans. *The Quarterly Journal of Experimental Psychology. B, Comparative and Physiological Psychology*, *51*(2), 97–120.

Wearden, J. H., & Ferrara, A. (1995). Stimulus spacing effects in temporal bisection by humans. *The Quarterly Journal of Experimental Psychology. B, Comparative and Physiological Psychology*, *48*(4), 289–310.

Webster, F. (2014). *Theories of the information society*. Routledge.

Westheimer, G. (2008). Was Helmholtz a Bayesian? *Perception*, *37*(5), 642–650.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble Perception. *Annual Review of Psychology*, *69*, 105–129.

Zakay, D. (1989). Subjective Time and Attentional Resource Allocation: An Integrated Model of Time Estimation. In *Time and Human Cognition: A Life-Span Perspective* (Vol. 59, pp. 365–397). Elsevier.

Zakay, D., & Block, R. A. (1996a). The role of attention in time estimation processes. In M. A. Pastor & J. Artieda (Eds.), *Time, internal clocks and movement.* (pp. 143–164). North-Holland/Elsevier Science Publishers.

Zakay, D., & Block, R. A. (1996b). Time, Internal Clocks and Movement. *Advances in Psychology*, *115*, 143–164.

Zang, X., Glasauer, S., Assumpcao, L., Wu, J., Xiu, C., & Shi, Z. (2022). What makes it so hard to acquire multiple priors in duration estimation? *In Submission*.

Acknowledgments

I would like to first thank my two supervisors: Prof. Zhuanghua Shi and Prof. Hermann Müller, who accepted me as their student and supervised my studies throughout the whole process of this dissertation. Great thanks to Prof. Zhuanghua Shi for his constant guidance on my scientific road, teaching me how to design experiments well and rigorously compose a scientific paper. In addition, I thank him for providing project funding and experimental conditions, and sparking my interest in the field of data analysis and Bayesian modeling.

I am deeply grateful to Prof. Hermann Müller for his encouragement and support for my doctoral work. Special thanks to Prof. Hermann Müller for his detailed feedback on experimental results and manuscripts. His rigorous scientific research attitude makes him be my role model in life. Also, I would like to express my special gratitude to Prof. Thomas Wachtler from Neurobiology Department who is kindly willing to be one of my examiners.

Particularly I want to thank my colleague Cemre baykan and Dr. Xuelian Zang for valuable suggestions and discussions on the my dissertation topic. Many thanks to their important contributions to this dissertation. I would also like to thank my colleague Dr. Fredrik Allenmark for always being available for discussions when I encountered problems in my research.

I would like to thank all MSense lab members for their efforts in creating such a relaxed, and productive working enviroment in past few year. I also want to thank Gabriella Zopcsak and Birgit Aßfalg from the administrative team for all of always being there for any of us who needs help.

I would like to say that I couldn't have survived without my 'CMG' friends and 'MensaBroy' friends for their active support over the last few years. I am greatly thankful to those encouragements when I felt frustrated in life or work.

Last but not least, I am deeply indebted to my dear families for their everlasting understanding, trustworthiness, and encouragement during the past years.

List of Publications

Journal Articles (# contributing equally)

Zhu, X.[#], Zang, X[#], Allenmark, F., Glasauer S., Müller, H. J., Shi, Z. Prior integration in Bayesian estimation under multi-prior context, *Journal of Experimental Psychology: Learning, Memory, and Cognition.* (In submission)

Zang, X[#], **Zhu, X.**[#], Allenmark, F., Wu J., Müller, H. J., Glasauer S., Shi, Z. Duration reproduction under memory pressure: Modeling the roles of visual memory load in duration encoding and reproduction, *Journal of Experimental Psychology: Learning, Memory, and Cognition.* (In submission)

Zang, X., Zinchenko, A., Wu, J., **Zhu, X.**, Fang, F., & Shi, Z. (2022). Contextual cueing in co-active visual search: Joint action allows acquisition of task-irrelevant context. *Attention, Perception & Psychophysics*. https://doi.org/10.3758/s13414-022-02470-x

Zhu, X., Baykan, C., Müller, H. J., & Shi, Z. (2021). Temporal bisection is influenced by ensemble statistics of the stimulus set. *Attention, Perception & Psychophysics*, *83*(3), 1201–1214.

Zang, X., Huang, L., **Zhu, X.**, Müller, H. J., & Shi, Z. (2020). Influences of luminance contrast and ambient lighting on visual context learning and retrieval. *Attention, Perception & Psychophysics*, *82*(8), 4007–4024.

Chen, S., Shi, Z., Zang, X., **Zhu, X.**, Assumpção, L., Müller, H. J., & Geyer, T. (2019). Crossmodal learning of target-context associations: When would tactile context predict visual search? *Attention, Perception & Psychophysics*. https://doi.org/10.3758/s13414-019-01907-0

Shi, Z., Allenmark, F., **Zhu, X.**, Elliott, M. A., & Müller, H. J. (2020). To quit or not to quit in dynamic search. *Attention, Perception & Psychophysics*, *82*(2), 799–817.

Zhu, X., Mou, D., & Ratiu, D. (2014). Structured multi-view modeling by tabular notation. *2014 IEEE 22nd International Requirements Engineering Conference (RE)*, 327–328.

Zhu, X. (2014). A Formal Specification Automation Method Based on Focus Framework. *International Journal of Modeling and Optimization*, *4*(2), 116.

Spichkova, M., **Zhu, X.**, & Mou, D. (2014). Do we really need to write documentation for a system? CASE tool add-ons: generator+editor for a precise documentation. In *arXiv [cs.SE]*. arXiv. http://arxiv.org/abs/1404.7265

Zhu, X. (2013). Pragmatic Tabular Specification Method for Interactive Systems. *2013 Fourth International Conference on Networking and Distributed Computing*, 151–154.

Zhu, X., Li, D., He, D., Wang, J., Ma, D., & Li, F. (2010). A remote wireless system for water quality online monitoring in intensive fish culture. *Computers and Electronics in Agriculture*, *71*, S3–S9.

Li, D., & **Zhu, X.** (2009). CDMA-Based Remote Wireless Water Quality Monitoring System for Intensive Fish Culture. *2009 WRI International Conference on Communications and Mobile Computing*, *2*, 380–385.

Workshop

Campetelli, A., Junker, M., Böhm, B., Davidich, M., Koutsoumpas, V., **Zhu, X**., & Wehrstedt, J. C. (2015). A Model-Based Approach to Formal Verification in Early Development Phases: A Desalination Plant Case Study. *Software Engineering (Workshops)*, 91–100.

Deutsche Zusammenfassung

Ein genaues Timing und die Wahrnehmung der Zeit sind für das richtige Handeln bei unseren täglichen Aktivitäten unerlässlich. Obwohl der Mensch kein spezifisches Sinnesorgan für die Zeitmessung hat, sind wir sehr wohl in der Lage, die Zeit zu messen, z. B. musikalischen Rhythmen genau zu folgen, den Takt und den Rhythmus zu fühlen, wenn wir Musik hören, und das Zeitintervall zu bestimmen und Zusammenstöße beim Autofahren schnell zu vermeiden. Obwohl der Mensch sehr gut in der Lage ist, die Zeit zu messen, ist die Zeitwahrnehmung im täglichen Leben nicht immer wahrheitsgetreu, sondern wird in der Regel von dem umgebenden Kontext beeinflusst. Die meisten von uns haben bewusst Situationen erlebt, in denen die Zeit während eines freudigen Ereignisses verfliegt oder sich während eines beängstigenden Ereignisses verlangsamt. Eine große Anzahl subjektiver Verzerrungen der Zeitwahrnehmung, einschließlich der Ausdehnung und Verkürzung der Dauer, wurde durch verschiedene Begleitumstände wie Bewegungsreize, Erregung und innere Zustände beschrieben (Lee 2017; Sackett et al. 2010; Erickson and Erickson 1958; van Wassenhove et al. 2008; Teixeira et al. 2013; Eagleman 2008). Da der Kontext oft nicht bewusst bekannt ist, erfolgt die Zeitverzerrung ohne explizites Wissen darüber, welche Einflüsse die Zeit verzerren.

Bei einer typischen Aufgabe zur Reproduktion der Dauer müssen die Teilnehmer zwar nur die Dauer für eine bestimmte vorgegebene Dauer reproduzieren, dennoch wird die reproduzierte Dauer durch verschiedene zeitliche Umstände beeinflusst, wie z. B. die zuvor abgetasteten Dauern (Vierordt 1868; Glasauer and Shi 2021a; Jazayeri and Shadlen 2010), den getesteten Bereich (Teghtsoonian and Teghtsoonian 1978) und die Abstände der abgetasteten Dauern (Brown et al. 2005; Wearden and Ferrara 1995; Penney, Brown, and Wong 2014). Zusätzlich zu den zeitlichen Begleitumständen können auch nicht-zeitliche Umstände die subjektive Zeit verzerren. Unsere Schätzung der verstrichenen Zeit wird durch interne körperliche Zustände verzerrt, darunter mentale Belastung (Angrilli et al. 1997), Aufmerksamkeit(Polti, Martin, and van Wassenhove 2018), emotionaler Zustand (Stetson, Fiesta, and Eagleman 2007) und externe Reize wie Größe, Helligkeit und Wahrscheinlichkeit des Auftretens (Eagleman 2008; Kanai and Watanabe 2006)). Ein klassisches Phänomen der durch Kontexteffekte verursachten Zeitverzerrung ist der Zentraltendenz-Effekt (namentlich Vierordt-Effekt): kurze Intervalle werden überschätzt und lange unterschätzt (Jazayeri and Shadlen 2010; Lejeune and Wearden 2009; Shi, Church, and Meck 2013; Glasauer and Shi 2021a; Vierordt 1868). Der Effekt der zentralen Tendenz zeigte eine systematische Regression

in Richtung des Mittelwerts der Stichprobendauer, was eine systematische Verzerrung darstellt, die von Begleitumständen auf globaler Ebene beeinflusst wird. Ein weiteres, eher lokales Phänomen, das vom zeitlichen Kontext auf lokaler Ebene abhängt, ist die sequentielle Abhängigkeit (Holland and Lockhead 1968), auch als serielle Abhängigkeit bezeichnet (Fischer and Whitney 2014; Kiyonaga et al. 2017), die besagt, dass die Intensität des vorhergehenden Reizes die Wahrnehmung des aktuellen Reizes anziehen kann. Es hat sich gezeigt, dass dieser lokale zeitliche Kontext eng mit dem Effekt der zentralen Tendenz in der Zeitwahrnehmung zusammenhängt (Glasauer and Shi 2021b). Der Effekt der zentralen Tendenz kann teilweise durch die dynamische Aktualisierung des lokalen zeitlichen Kontextes erklärt werden.

Im letzten halben Jahrhundert wurden sowohl zeitliche als auch nicht-zeitliche kontextuelle Verzerrungen nicht nur bei der Erforschung der Eigenschaften des Timings (d.h. Zentraltendenz-Effekt und Serienabhängigkeit), sondern auch bei der Modellierung des Timing-Prozesses eingehend untersucht. Die ursprünglichen Modelle der internen Uhr (J. Gibbon, Church, and Meck 1984; Treisman 1963) gehen davon aus, dass die Zeitrepräsentation im Gedächtnis direkt mit den im Gedächtnis akkumulierten zeitlichen "Ticks" korrespondiert, was darauf hindeutet, dass die Dauerrepräsentation mehr oder weniger wahrheitsgetreu sensorische Eingaben widerspiegelt. So werden nicht-zeitliche Einflüsse häufig mit der Modulation der internen Uhrgeschwindigkeiten erklärt, z. B. dass große Reize oder hohe Erregung die Geschwindigkeit der internen Uhr erhöhen. Die bloße Modulation der Uhrengeschwindigkeit konnte jedoch verschiedene Befunde nicht durch Aufmerksamkeitsmodulation erklären. In Erweiterung des klassischen Modells der internen Uhr beziehen die Attentional-Gating-Theorie (D. Zakay and Block 1996; Dan Zakay and Block 1996) und der Attentional-Sharing-Aspekt (Fortin and Rousseau 1998; Fortin and Massé 2000) die Aufmerksamkeit und das Gedächtnis in das Modell der internen Uhr ein. Das Modell der Aufmerksamkeitsteilung geht davon aus, dass die Anhäufung interner "Ticks" durch die Modulation der Aufmerksamkeit auf den Timing-Prozess variiert werden kann.

Obwohl die klassischen Modelle der inneren Uhr (Treisman 1963; J. Gibbon, Church, and Meck 1984; John Gibbon 1977) und ihre Erweiterungen, wie z. B. die Aufmerksamkeitsverteilung (D. Zakay and Block 1996; Dan Zakay and Block 1996) qualitativ empirische Befunde erklären können, wurden quantitative Vorhersagen verschiedener kontextueller Verzerrungen in der Zeitwahrnehmung erst im letzten Jahrzehnt mit Hilfe des probabilistischen Ansatzes erreicht (Jazayeri and Shadlen 2010; Shi, Church, and Meck 2013;

Shi and Burr 2016; Gu et al. 2016). Der Bayes'sche Ansatz steht im krassen Gegensatz zu den klassischen Modellen der inneren Uhr. Der Bayes'sche Ansatz geht in der Regel nicht von "modulartigen" kognitiven Verarbeitungsstufen aus, wie es die Modelle der internen Uhr tun, sondern von einem quantitativen Integrationsprozess. In einem Übersichtsartikel überbrückten Shi et al. (2013) diese Lücke, indem sie den Bayes'schen Ansatz mit den klassischen Modellen der internen Uhr verknüpften. Mit Bayes'schen Integrationsansätzen haben eine Reihe von Studien gezeigt, dass verschiedene Zeitverzerrungen gut durch eine optimale Fusion von verrauschten sensorischen Informationen mit Erwartungen, die auf früheren Erfahrungen beruhen, erklärt werden können (Roach et al. 2017; Cicchini et al. 2012; Shi, Church, and Meck 2013; Shi and Burr 2016).

Während sich klassische Intervall-Timing-Modelle darauf konzentrieren, wie einzelne Intervalle kodiert, gespeichert und verarbeitet werden, ist immer noch unklar, wie zeitliche Begleitumstände (z. B. Intervallsequenzen, mehrfache Interleaving-Bereiche) und nicht-zeitliche Umstände (z. B. Gedächtnisbelastung) interne Schätzungen beeinflussen könnten. Ohne die Rolle dieser zeitlichen und nicht-zeitlichen Kontextfaktoren vollständig zu untersuchen, bleibt es eine Herausforderung, den Timing-Prozess vollständig zu verstehen und zwischen verschiedenen Timing-Modellen zu unterscheiden. In dieser Studie gehen wir daher der Frage nach internen Zeitmodellen für drei spezifische Kontexte nach (d.h. Ensemble-Kontext, Kontext unter Gedächtnisbelastung, Multi-Prior-Kontext), um die kontextuelle Verzerrung in der Zeitwahrnehmung mittels Bays'scher Inferenz zu erklären. Ziel der aktuellen Studie ist es, die kognitiven Mechanismen, die dem kontextuellen Effekt bei der subjektiven Intervallzeitmessung zugrunde liegen, mittels Bayes'scher Inferenz zu untersuchen. Um diese Frage anzugehen, werden klassische Verhaltensuntersuchungen und Bayes'sche Modellierungstechniken eingesetzt.

In Kapitel 2.1 werden unter Übernahme und Modifizierung des ursprünglichen Paradigmas von (Penney, Brown, and Wong 2014) drei Experimente zur zeitlichen Bisektion durchgeführt, um die Auswirkungen des Ensemble-Kontextes auf die Zeitwahrnehmung durch Manipulationen der Stichprobenverteilung durch Stimulusabstände, Verteilungsmittelwerte und Varianzen zu untersuchen, um zu ermitteln, wie der Ensemble-Kontext die Leistung der Aufgabe der zeitlichen Bisektion moduliert. Experiment 1 replizierte wie erwartet den Abstands-Effekt (Wearden and Ferrara 1995; Penney, Brown, and Wong 2014), um die Verschiebung des Bisektionspunktes in Sets mit positivem (PS) und negativem (NS) Abstand zu untersuchen. Um die Aufgabe der zeitlichen Bisektion mit gleichmäßig verteilten Dauern,

aber unterschiedlichen Ensemble-Mitteln zwischen den beiden Bedingungen weiter zu untersuchen, wurden in Experiment 2 zwei Sets mit schräger Frequenzverteilung verwendet: aufsteigende Frequenz (AF) und absteigende Frequenz (DF) Die zeitliche Bisektion beinhaltet im Wesentlichen einen Vergleich einer gegebenen Dauer mit der Schätzung des Ensemble-Mittels. Im Vergleich zum AF-Set würde der relativ kürzere Ensemble-Mittelwert im DF-Set zu mehr "langen" Antworten führen. Die Ergebnisse von Versuch 2 zeigten, dass die Verschiebung der Bisektionspunkte mit dem Ensemble-Mittelwert des Probenintervallsets, nicht aber mit der Abstandsinformation zusammenhängt. Die Ergebnisse von Experiment 3 zeigten eine signifikant steilere Neigung der Bisektionsfunktionen bei der umgekehrten T-Form im Vergleich zur U-Form, da die Varianz bei der U-Form geringer war als bei der T-Form. Dies zeigt, dass die Varianz die Schwierigkeit der Bisektion beeinflusst und nicht die Bisektionspunkte. Die Ergebnisse zeigten, dass der Mittelwert und die Varianz des Stimulus-Sets kritische Faktoren sind, die den Bisektionspunkt verschieben bzw. die Empfindlichkeit der Bisektionsunterscheidung beeinflussen. Diese Ergebnisse zeigen, dass die Dauerwahrnehmung einen Ensemble-Mittelungsprozess beinhaltet, der ähnlich wie andere Wahrnehmungseigenschaften im visuellen und auditiven Bereich funktioniert. Um zu untersuchen, ob die Beurteilung der zeitlichen Halbierung am besten durch die Ensemble-Verteilung erklärt werden kann, haben wir ein hierarchisches Modell vorgeschlagen und implementiert, das erklärt, auf welche Weise die subjektive Beurteilung von Zeitintervallen je nach der Verteilungsstatistik des Mittelwerts und der Varianz beim Vergleich der Dauer variiert. Die hierarchische Bayes'sche Modellierung wurde auf die Verhaltensdaten angewandt, um zu überprüfen, ob der Gesamtmittelwert und die Varianz der Intervallverteilung berücksichtigt werden sollten. Unsere Ergebnisse gehen darüber hinaus, indem sie darauf hinweisen, dass die Varianzinformation auch eine wichtige Rolle bei zeitlichen Beurteilungen spielt, da die Varianz der Stichprobenverteilung nützliche Informationen für die Unterscheidung der Lage einer Probedauer relativ zum Ensemblemittelwert enthält und ein Intervallsatz mit geringerer Varianz die Sensibilität der zeitlichen Beurteilung erhöht.

Anschließend wurde in Kapitel 2.2, inspiriert durch die Darstellung der Aufmerksamkeitsteilung (Macar, Grondin, and Casini 1994; Fortin and Rousseau 1998), durch die Übernahme und Modifizierung eines Dual-Task-Paradigmas von (Allred et al. 2016), bestehend aus einer sekundären visuellen Arbeitsgedächtnisaufgabe (mit niedriger, mittlerer oder hoher Belastung) und einer primären Dauerproduktions- und -reproduktionsaufgabe, der Frage nachgegangen, wie die Gedächtnisbelastung die Produktions- und Reproduktionsphasen

im Dauerreproduktions-Experiment beeinflussen würde. Es wurden vier Aufgaben mit geteilter Aufmerksamkeit durchgeführt, um die Auswirkungen der kognitiven Belastung auf die Schätzung der Dauer in den Phasen der Kodierung und Reproduktion der Dauer zu untersuchen. Experiment 1 diente als Basislinie ohne kognitive Belastung bei Reproduktionsaufgaben, und die Aufgabenabfolge bei jedem Versuch war als Dauer-Produktionsphase, Dauer-Reproduktionsphase, Gedächtnisphase und schließlich Gedächtnis-Testphase konzipiert. In Experiment 2 wurde untersucht, ob die Beibehaltung von Informationen im Arbeitsgedächtnis während der Produktionsphase (nicht aber während der Reproduktionsphase) die Messung der sensorischen Dauer und die Verwendung oder Aktualisierung von Prioren beeinflussen würde. Versuch 3 ähnelt Versuch 2, mit dem Unterschied, dass sich die Arbeitsgedächtnisaufgabe nur mit der Reproduktionsphase überschnitt, d. h. jeder Versuch begann zunächst mit der Produktionsphase, gefolgt von der Gedächtnisphase. In Versuch 4 wurde untersucht, wie die Dauerschätzung beeinflusst wird, wenn die Arbeitsgedächtnisaufgabe sowohl die Produktions- als auch die Reproduktionsphase umfasst. Wir erwarteten, dass die kognitive Belastung die wahrgenommene und reproduzierte Dauer sowohl im Mittelwert als auch in der Variabilität beeinflussen würde, und fanden eine entgegengesetzte Auswirkung der kognitiven Belastung auf die sensorische Enkodierungsphase und die Reproduktionsphase: Unterschätzung der wahrgenommenen Zeit in der ersten und Überschätzung der Reproduktion in der zweiten Phase. Außerdem wurde unter Berücksichtigung der Auswirkungen der kognitiven Belastung ein allgemeiner Bayes'scher Berechnungsrahmen vorgeschlagen und entwickelt, um zu überprüfen, wie die sensorische Schätzung mit dem Prior während der Phasen der Zeitkodierung und Zeitreproduktion integriert wird. Mithilfe eines generativen Bayes'schen Modells haben wir detailliert beschrieben, wann und wie der Gedächtnisdruck die Zeitschätzung und die damit verbundenen Auswirkungen auf die Verzerrung durch die zentrale Tendenz beeinflusst, und sowohl die allgemeine Über- und Unterschätzung als auch die in allen vier Experimenten beobachteten Effekte der zentralen Tendenz quantitativ vorhergesagt. Nicht zuletzt könnte das von uns vorgeschlagene generative Modell für den Einfluss der kognitiven Belastung auf die Zeitwahrnehmung auf andere Formen der Größenwahrnehmung verallgemeinerbar sein.

Die dritte Studie (Kapitel 2.3) konzentrierte sich auf die Frage, wie mehrere vorherige zeitliche Kontexte die subjektive Zeiteinschätzung der Dauer bei Aufgaben zur Dauerreproduktion beeinflussen, und untersuchte die Struktur der vorherigen Integration während der Zeitschätzung in mehreren vorherigen zeitlichen Umgebungen. Durch die

Übernahme und Modifikation des klassischen Paradigmas der zentralen Tendenz (Roach et al. 2017; Jazayeri and Shadlen 2010) wurden zwei Dauerreproduktions-Experimente mit zwei Varianzen der kurzen und langen Reichweite auf der logarithmischen Skala, aber dem gleichen geometrischen Mittelwert der verschachtelten Reichweite durchgeführt, um den zugrunde liegenden Mechanismus während der Integration von Vorwissen aufzudecken. Die Teilnehmer wurden zunächst mit den kurzen und langen Intervallen blockweise trainiert (d. h. Lernen und Anpassen an unterschiedliche Bereiche der abgetasteten Intervalle) und dann mit verschachtelten Bereichen getestet. Ziel dieser Studie ist es, mögliche Integrationshierarchien der globalen und lokalen Prioren der Geschichte zu untersuchen. In dieser Studie wurden kurze und lange Dauern mit zwei Orten assoziiert. Wir untersuchten, ob die subjektive Schätzung der Dauer von globalen Prioren beeinflusst wird, die aus vorangegangenen Blöcken gelernt wurden (Intervalle aus verschiedenen Bereichen werden blockweise präsentiert), und fanden heraus, dass die Existenz eines globalen Priors bei der Zeitschätzung in der verschachtelten Sitzung zusätzlich zu den kurzen und langen Prioren, die in kurzen und langen Sitzungen erworben wurden, stattfindet. Um zu untersuchen, wie der globale Prior seine Rolle in verschiedenen Prior-Kontexten spielt, und um eine hierarchische Struktur der Prior-Integration zu erforschen, haben wir drei mögliche hierarchische Strukturen der globalen und lokalen Prior-Integration in der Bayes'schen Schätzung vorgeschlagen. Darüber hinaus haben wir unser Bayes'sches Beobachtermodell auf mehrere zeitliche Kontexte erweitert, indem wir globales Vorwissen berücksichtigt haben. Die Regeln für die Integration von Prioren wurden durch die Untersuchung der Rolle von globalen Prioren bei der Schätzung weiter aufgedeckt. Es wurden drei hierarchische Prior-Integrationsmodelle vorgeschlagen, und die Ergebnisse zeigten, dass die von uns vorgeschlagenen Modelle nicht nur die Verzerrung des Effekts der zentralen Tendenz, sondern auch den Variationskoeffizienten gleichermaßen gut vorhersagen können, trotz unterschiedlicher struktureller Annahmen. Wir fanden heraus, dass der globale Prior nicht nur durch das geometrische Mittel des Interleaved-Bereichs bestimmt wird, sondern auch stärker durch den aktuellen Kontext beeinflusst wird und dass die Sitzung vor der Mix-Sitzung den globalen Prior stärker beeinflusst als die erste Sitzung des Experiments, anstatt sich gleichmäßig darauf auszuwirken. Wenn das Experiment eine geringere Variabilität im blockierten und verschachtelten Bereich auf der logarithmischen Skala aufweist, stützt sich die Zeitschätzung außerdem stärker auf das globale Vorwissen.

Die drei oben genannten Studien haben zusammengenommen kognitive Mechanismen von Kontexteffekten im Intervall-Timing aufgedeckt, indem sie den Rahmen der Bayes'schen

Inferenz nutzten und insbesondere auf drei spezifische Kontexteffekte abzielten: den Ensemble-Kontext, den zeitlichen Kontext unter Gedächtnisbelastung und den zeitlichen Kontext mit mehreren Prioren. Erstens untersuchten wir die Ensemble-Wahrnehmung für die zeitliche Abfolge und entdeckten, dass der Mittelwert und die Varianz des Stimulus-Sets kritische Faktoren sind, die zu Verschiebungen des Bisektionspunktes führen bzw. die Empfindlichkeit bei zeitlichen Bisektionsurteilen bestimmen. Zweitens wurde die Wechselwirkung zwischen statistischem Kontext und kognitiver Belastung bei der Zeitwahrnehmung und der Reproduktion der Dauer mit Hilfe des Bayes'schen Modellierungsansatzes untersucht, um die unterschiedlichen Auswirkungen der Belastung des Arbeitsgedächtnisses auf die Kodierungs- und Reproduktionsphase der Dauer zu verstehen. Drittens haben wir durch die Erweiterung des standardmäßigen uni-prioren Bayes'schen Modells auf den multi-prioren zeitlichen Kontext den Einfluss früherer Sitzungen aufgedeckt, was darauf hindeutet, dass wir die lokalen und globalen Priors zusammen betrachten müssen. Anschließend schlugen wir drei Modelle für die Integration von Prioren vor, die der Reproduktion der Dual-Task-Dauer zugrunde liegen. Der Vergleich der Modellvorhersagen zeigte, dass die vorgeschlagenen Modelle trotz ihrer unterschiedlichen strukturellen Annahmen eine gleich gute Vorhersageleistung zeigten. Zu guter Letzt wurden die oben genannten Ergebnisse durch die von uns vorgeschlagenen und implementierten Bayes'schen Modellrahmen erklärt.