BRILL *Timing & Time Perception* (*2024*) DOI: 10.1163/22134468-bja10114

Timing & Time Perception
brill.com/time

Psychometric Assessment of the Temporal Bisection Task with Discrete and Continuous Response Formats

Ivan Quan[1], Rebekka Lagacé-Cusiac[2,3,*] and Jessica Grahn[2,3,*]

[1]Department of Physiology and Pharmacology, Western University, London, ON, N6A 5C1, Canada

[2]Department of Psychology, Western University, London, ON, N6A 5C2, Canada

[3]Brain and Mind, Western University, London, ON, N6A 3K7, Canada

*Corresponding authors; e-mail: rlagacec@uwo.ca; jgrahn@uwo.ca

ORCID iDs: Quan: 0009-0002-2195-3428; Lagacé-Cusiac: 0000-0003-1322-7792; Grahn: 0000-0001-7270-2114

Received 3 May 2024; accepted 20 September 2024
published online 24 October 2024

Abstract

The temporal bisection task has long been used to study time perception as well as measure individual differences in time perception ability. The task involves training participants on short and long reference durations before presenting intermediate durations and asking participants to classify them as 'short' or 'long'. However, there is little information about how well the bisection task measures individual differences in timing ability. To bridge this gap, we assessed the psychometric properties of measures obtained from a classic temporal bisection task: Weber ratio and percent correct. Because measures with binary responses tend to require many trials to reach adequate reliability, we also assessed the psychometric properties of a modified bisection task which used a continuous response format. In this task, participants represented intermediate durations on a visual analogue scale. Estimation error was used as the outcome measure. Participants (n = 46) completed the classic and modified bisection tasks twice across two sessions approximately one week apart. The modified bisection task had excellent internal consistency and test–retest reliability, while the classic task had fair to good internal consistency and good test–retest reliability. Overall, estimation error had the highest reliability, followed by percent correct, and then Weber ratio. In terms of validity, there was excellent convergent validity between the classic and modified bisection tasks. As an exploratory analysis, we assessed how the number of trials affected the reliability of each outcome measure across the two tasks. Based on this, we make recommendations on how to optimize reliability for both tasks in future research.

Keywords

bisection, relative timing, reliability, temporal discrimination, VAS, Weber ratio

1. Introduction

Time perception is critical for many key processes, from organizing daily schedules to producing motor outputs, and even to comprehending and generating speech (Carroll et al., 2008; Clynes and Walker, 1986; Macar and Vidal, 2004; Matell and Meck, 2000). One difficulty with studying time perception is that its sensations cannot be traced to an obvious source organ (Allman and Meck, 2012; Kopec and Brody, 2010). For this reason, temporal processes have long been studied using psychophysical approaches. Psychophysics examines how sensory experiences vary based on varying stimulus parameters to make inferences about the cognitive processes behind sensation and perception (García-Pérez, 2014; Read, 2015). The temporal bisection task is a well-established psychophysical method used to study cognitive mechanisms related to temporal perception (Allan and Gibbon, 1991; Allman and Meck, 2012; Church and Deluty, 1977; Kopec and Brody, 2010; Wearden, 1991).

Conventionally, this task first involves training participants on two reference durations labelled 'short' and 'long' (Allan and Gibbon, 1991; Wearden, 1991). After training, participants are presented with intermediate and reference durations and asked to judge similarity to or classify them in a binary two-alternative choice as either 'short' or 'long' (Allman and Meck, 2012). Two main measures stemming from the bisection task are the bisection point and the Weber ratio. The bisection point, also known as the point of subjective equality, refers to the duration at which participants are equally likely to respond 'short' or 'long'. Previous research has used the bisection point extensively because it offers significant insight into the cognitive processing and internal representations of duration (Siegel and Church, 1984; Wearden, 1991; Allan, 2002; Allan and Gibbon, 1991; Church and Deluty, 1977; Droit-Volet, 2003; Droit-Volet et al., 2007; Karşılar et al., 2018; Ortega and López, 2008; Penney et al., 2000; Wearden and Ferrara, 1995; Wearden and Ferrara, 1996; Wearden et al., 1997). In contrast, performance on the temporal bisection task is typically measured using the Weber ratio, a measure of sensitivity or discrimination ability. When plotting proportions of 'long' responses against stimulus duration, a steeper central slope would lead to a smaller Weber ratio and indicate that a participant can perceive smaller changes in stimulus duration and therefore have a higher discrimination ability (Kopec and Brody, 2010).

This classic version of the bisection task presents many advantages as it is very simple and can be used to study time perception in humans across the life span as well as animals (Allan and Gibbon, 1991; Church and Deluty, 1977; Droit-Volet and Wearden, 2001; Provasi et al., 2011; Wearden, 1991). For example, it has been used to identify deficits in temporal processing in clinical populations, such as in Parkinson's, cerebellar degeneration and schizophrenia, finding that people with these disorders had impaired judgements in timing (Nichelli et al., 1996;

Elvevåg et al., 2003; Allman and Meck, 2012; Wearden and Jones, 2013). Other research using the bisection task has centred around interactions between time and other magnitudes, such as numerosity (e.g., Droit-Volet et al., 2003) and area (e.g., Lambrechts et al., 2013). In addition, this task is increasingly used to assess individual differences in time perception ability and how they correlate with personality traits (Corcoran et al., 2018; Momi et al., 2023), other cognitive abilities such as working memory and processing speed (Droit-Volet et al., 2015; Mendez et al., 2011; Ogden et al., 2018), and neuroimaging and physiological measures of cognition (Sadibolova et al., 2022; Tipples et al., 2013).

There is limited information, however, on the classic bisection task's reliability and adequacy for measuring individual differences. Few studies report the reliability of their tasks, even though reliability is a key component for interpreting correlational findings. This is problematic seeing as low reliability can lead to attenuated correlations between measures (Spearman, 1907, 1910). Furthermore, the reliability of traditionally experimental tasks is not always adequate for individual differences research (Hedge et al., 2018; Parsons et al., 2019). For example, Hedge et al. (2018) demonstrated that composite measures from traditionally experimental tasks, like the Stroop task, can have low reliability. In the timing field, Marx et al. (2021) studied the reliability of commonly used timing tasks, including time estimation, time production, time reproduction, and time discrimination, and found that many had low internal consistency as well as test–retest reliability. Therefore, the primary goal of this study was to assess the psychometric properties of the temporal bisection task. We hypothesized that the binary response format ('short' or 'long') might result in low reliability, especially when there are too few trials to precisely estimate the Weber ratio.

As previously mentioned, the temporal bisection task has informed researchers about both timing perception as well as decision-making to some degree. For instance, the study of the temporal bisection task has led to the two-step process in decision-making in bisection tasks elaborated by Kopec and Brody (2010). The two-step model of decision-making assumes that internal representations of reference durations are modelled as normal distributions centred around the perceived length of learned durations, with standard deviations proportional to the magnitude of these learned durations (Church and Gibbon, 1982). The likelihood of recognizing that a stimulus of a certain duration is the learned reference is represented by the height of each distribution. From this assumption, the first step in the two-step decision-making model is to determine if a presented stimulus is one of the reference durations. If the stimulus is identified as a reference duration, one would answer 'short' or 'long'. However, if the stimulus was perceived to be neither of the reference durations, the participant would proceed to the second step to compare the relative distance between the presented stimulus and the reference

durations (Kopec and Brody, 2010). Participants then answer based on which reference appeared closer. Because of the inherently larger standard deviation of the long reference compared to the short reference, a larger proportion of presented stimulus durations will be immediately recognized as the long reference (Church and Gibbon, 1982). As a result, stimuli represented as intermediate durations will tend towards more short responses due to the gambler's fallacy, driving the belief that a previous 'long' response decreases the probability of another 'long' response (Kopec and Brody, 2010).

This two-step model of decision-making is a useful explanation of common results found in bisection tasks, such as finding that the bisection point is often near the geometric mean and that increasing the spread between reference durations results in the bisection point approaching the arithmetic mean (Kopec and Brody, 2010). However, it provides only a limited explanation of how we perceive intermediate durations (Lindbergh and Kieffaber, 2013). Because participants are forced to respond in a two-alternative 'short' or 'long' decision, it is unclear how intermediate durations are perceived relative to the reference durations held in memory. Furthermore, responses on the classic bisection task may also be the results of nontemporal decision-making mechanisms. Previous research has used other paradigms to investigate the internal representation of time intervals. For example, Allan (1978) along with several others has used ratio-setting paradigms which require participants to reproduce an interval duration as well as produce an interval equivalent to half or double the duration of the original reference duration (e.g., Corcoran et al., 2018; Momi et al., 2023). These alternative tasks have more recently been proposed as alternative versions of a bisection task to measure time processing. For example, Corcoran et al. (2018) implemented a task in which participants had to press a button for half the duration of the reference duration, thereby bisecting the reference interval. These tasks have some advantages. First, the continuous response format is more fine-grained compared to a binary response format. This is because variability related to a person's timing ability is increased, which can result in higher reliability (Cohen, 1988). Second, (re)producing a duration provides a more direct measure of its internal representation in contrast to comparing it to reference durations and doing an alternative forced-choice task (like in the classic bisection). For instance, in a classic bisection task, if a participant perceives the duration in a trial as an intermediate duration (e.g., the second step of the two-step model), a production task would allow them to report their perception while a binary response format would not.

However, this type of production task which requires participants to subdivide an interval also has some limitations. First, temporal perception and production are confounded. Second, very few durations are studied (e.g., half of or double the reference duration). In contrast, other studies have used different types of production tasks which require participants to give a numerical estimate or visual

representation of an interval's duration relative to a reference duration. For example, Wearden and Jones (2007) asked participants to estimate the proportion of a duration less than 10 s relative to a reference duration of 10 s and found that participants' internal representation of intermediate durations was linear. Other tasks measuring temporal ratio perception have also used visual analogue scales (VAS) to study how temporally subdivided intervals are perceived. These tasks consisted of playing three tones, and representing on a bounded line when the middle tone occurs in relation to the first and third tones (Lagacé-Cusiac et al., 2023; Nakajima, 1987). Given that there is previous evidence showing that humans can visually or numerically represent proportions of interval durations, we propose a modified bisection task procedure in which participants must represent the duration of an intermediate duration relative to the short and long reference durations, thereby unconstraining the response format from its binary nature. Therefore, rather than forcing them to make a binary decision as to *which* reference duration is more similar (short or long), participants can indicate *how much* more similar the intermediate duration is to reference durations using the VAS. More importantly, the modified bisection task might also benefit from good psychometric properties due to the inherently granular nature of the continuous response format such as a VAS.

To summarize, the current study had two main goals. The first goal was to assess the internal consistency and test–retest reliability of the classic bisection task. Because psychophysical tasks have long been regarded as an objective measure of perception ability, analyses of reliability have not generally been conducted, which is an issue for many psychological tasks (Hedge et al., 2018; Parsons et al., 2019). The second goal was to investigate human performance on a modified temporal bisection task in which participants responded by representing the intermediate durations along a visual continuous sliding scale, and assess and compare its internal consistency and test–retest reliability to the classic bisection task. More specifically, we assessed how well participants could perceive and represent intermediate durations for a subsecond interval (between 500 ms and 1000 ms) and the degree to which the modified bisection could be an adequate measure of time perception ability. Studying the reliability of psychophysical measures is essential to making robust inferences, especially when using an individual differences or correlational approach.

2. Methods

2.1. Participants

A total of 70 participants from Canada and the United States were recruited from MTurk via CloudResearch. Of these participants, 24 were excluded from

analyses due to lack of attention, lack of compliance, or not understanding the task (see Section 2.4.2. *Preprocessing*). The final sample consisted of 46 participants (42.82 ± 9.34 years old; 22 female, 21 male, three unspecified). Consent was received from all participants prior to their participation in the study, and all procedures were approved by the Research Ethics Board at Western University. Of these 46 participants, 33 completed a second session (to assess test–retest reliability) approximately seven days after completing the first (7.39 ± 0.97 days).

2.2. Stimuli and Apparatus

The study was implemented using the software PsychoPy (version 2020.2.10) and was conducted as an online experiment hosted on Pavlovia. Two constant 500-Hz tones of durations 500 ms and 1000 ms with 10-ms linear onset/offset ramps were used as reference stimuli. In addition to these, nine intermediate durations in steps of 50 ms were generated as intermediate stimuli. A constant 1500-Hz tone lasting 1000 ms was used as an attention check stimulus. All auditory stimuli were generated using the "audiowrite" function in MATLAB (version R2022b). Participants used a web browser on their personal computers to complete the study and were encouraged to use headphones before starting the study. At the start of each session, a 10-s constant tone was played during which participants were instructed to adjust their volume to a comfortable level.

2.3. Study Design

In each session, participants completed both the classic and modified temporal bisection task. The order of tasks was counterbalanced across participants but remained the same across the two sessions for a given participant. Between tasks, participants were allotted a two-minute break. Both the classic and modified bisection tasks included three consecutive phases: a reference training phase, a practice phase, and a test phase (Fig. 1). All phases were completed for one task before moving on to the second task.

In the reference training phase, participants were trained on two reference durations. Alternating 500 ms (short) and 1000 ms (long) reference tones were labelled and presented five times each. This phase was identical for both versions of the bisection tasks. Then, in the practice phase, participants completed practice trials to familiarize themselves with the response process for the task (i.e., VAS of the modified bisection task, and button pressed for the classic bisection task). In this phase, they were also instructed on how to respond to attention checks. In the classic task, participants responded by pressing 'S' on the keyboard if they perceived the tone as the short reference, or 'L' if they perceived it to be the long reference. If the higher-pitched attention check was presented, participants instead responded by pressing 'G'. Participants completed six trials in the practice phase (two trials/reference duration, two attention checks). In each practice trial, participants were told which stimulus they would hear and which key to press. They

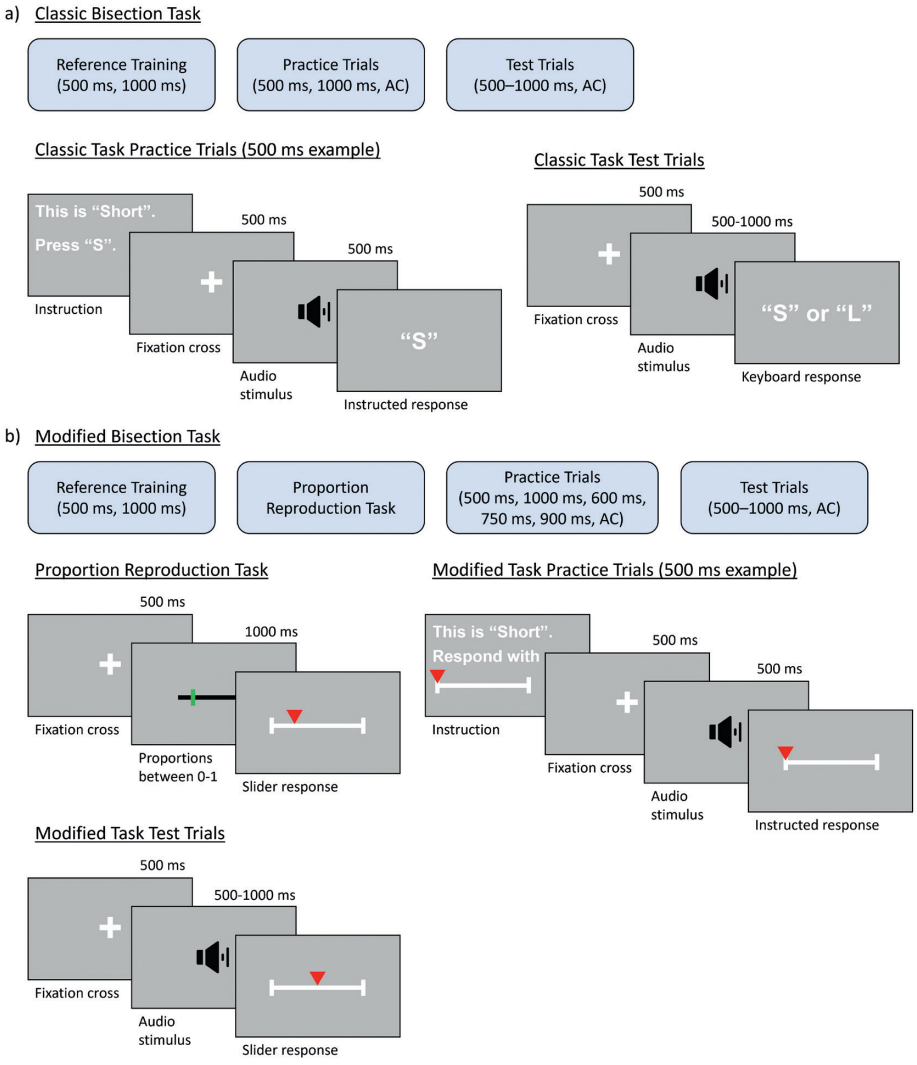

Figure 1. Study design and trial presentations for the classic and modified bisection task. Abbreviation: AC, attention check.

were presented with the stimulus and were directed to respond as instructed. In the modified task, participants responded on a visual analogue scale. The slider ranged from 'short' to 'long' to represent the reference durations, and participants were instructed to drag the slider at or in between the reference durations based on their perception of the presented tone's length. Participants would then press on the spacebar to confirm their responses. In the case of attention checks, participants were instructed to respond by dragging the slider to the far right (i.e., 'long' response). Similarly to the classic bisection task, participants completed

four practice trials with reference durations (two trials/reference duration). In addition, participants completed three practice trials with intermediate durations (600 ms, 750 ms, and 900 ms) and two trials for the attention check. For all modified task practice trials, participants were told which stimulus they would hear (short, long, intermediate, or attention) and were shown an image of the slider with the correct response (e.g., in the middle for 750 ms). They were then presented with the stimulus and directed to respond as instructed. Previous piloting showed that not including these instructions led to participants only answering at the slider extremities.

In addition, participants completed a line proportion reproduction task prior to the practice phase of the modified bisection task (Fig. 1b) to gain familiarity with the response format of the VAS. In this task, an image displaying a bisected line was displayed to participants. Participants responded by reproducing this proportion on a VAS. The visual length of the stimulus was shorter than the length of the VAS response line to discourage exact replication of the VAS stimulus position. A total of ten bisected lines ranging from 0 to 1 in steps of 0.1 were presented in random order, thus resulting in 10 trials.

In the test phase, participants completed 12 blocks of 12 trials (two reference durations, nine intermediate durations, and one attention check). Thus, participants each completed a total of 132 trials in each task (excluding attention trials). Within each block, the stimuli were presented in random order. After the presentation of each stimulus, participants were instructed to respond based on their perception of the tone's duration compared to reference tones. After every two experimental blocks, participants were allotted a one-minute break.

2.4. Statistical Analyses

2.4.1. Outcome Measures

For the classic bisection task, the Weber ratio and percent correct were computed for each session. The Weber ratio was calculated as the ratio between the difference limen and the bisection point (Elvevåg et al., 2003). The bisection point was calculated using methods outlined in Wearden (1991). For each duration, the proportion of 'long' responses across all trials and participants was calculated. A least-squares linear regression was performed on the four points encompassing the steepest slope when comparing stimulus durations and 'long' response proportions. This regression was then used to calculate the bisection point where 'long' responses reached 50% (Wearden, 1991). The difference limen was calculated as half the duration between the points on this regression where the proportions of 'long' responses were 0.25 and 0.75 (Wearden, 1991). While both logistic curve and linear regression methods have been used to calculate the Weber ratio in previous studies (Allman et al., 2011; Droit-Volet et al., 2015; Ogden et al., 2018), Allman et al. (2011) found that both approaches yield similar results. Percent correct was

measured as the proportion of 'correct' responses, with a correct response being the reference duration closest to the presented tone. Intermediate durations of 750 ms were excluded from this score seeing as there is no correct response for this duration.

For the modified bisection task, the estimation error was obtained for each trial by calculating the absolute difference between the estimated ratio (the position on the VAS between 0 and 1) and the stimulus ratio (stimuli duration/difference between the two reference durations), resulting is a nondirectional measurement (i.e., all estimation errors were positive). The mean estimation error for all trials was calculated for each session as an outcome measure for the modified bisection task.

2.4.2. Preprocessing

Data from participants who scored less than 80% correct on attention checks were inspected for potential signs of noncompliance. We found 19 participants who scored less than 80% on the attention checks for either session. Of those 19, three were kept in the analysis. The remaining 16 were excluded because (a) they indicated they did not understand the task, (b) a visual inspection of their data on the classic task indicated they gave the same proportion of 'long' responses over all stimulus durations, or (c) a visual inspection of their modified task responses indicated they used three or fewer slider values (i.e., 0,.5 and 1 or 0 and 1). In addition to these 16 exclusions, eight more participants were excluded after visual inspection of their data revealed a flatline for the classical bisection task, or three or fewer slider values in the modified bisection task, indicating a lack of understanding or compliance. Thus, of the 70 participants who completed the study, a total of 24 participants were excluded from the analysis. All data used in the analyses as well as the analysis scripts are available on OSF (https://osf.io/dngam/).

2.4.3. Internal Consistency

Internal consistency was estimated using permutation-based split-half reliability for both classic bisection task outcome measures (Weber ratio and percent correct) and the modified task outcome measure (estimation error) (Parsons et al., 2019). Participant responses for each stimulus duration in the bisection task were randomly split into halves, with each split balanced across stimulus durations. Within each split-half, the outcome measure was calculated for each participant. Pearson's correlation was then calculated between each half of the split. One thousand permutations were performed, and results were averaged to obtain the mean correlation alongside a 95% confidence interval. To account for each split-half only having half the data, the Spearman–Brown (SB) prophecy formula was used on the mean correlation coefficient and confidence interval to account for the reduced number of trials (Parsons et al., 2019). This procedure was performed separately for each outcome measure.

2.4.4. Test–Retest Reliability

(ICC) were used as a measure of test–retest reliability between the first and second sessions for all outcome measures. A mean-rating, absolute agreement, two-way mixed-effects model intraclass correlation was conducted using the R *icc* package (version 0.84.1) on each measure to generate the ICC and 95% confidence interval (Koo and Li, 2016). Like other correlations, ICC values are normally between 0 and 1, where values below 0.5 indicate poor reliability, values between 0.5 and 0.75 are fair, values between 0.75 and 0.9 are good, and values above 0.9 are excellent (Koo and Li, 2016).

2.4.5. Reliability and Trial Number

To investigate the impact of varying the number of trials per stimulus on task outcome reliability, intraclass correlations were conducted on outcome measures for the classic bisection task (Weber ratio and percent correct) and the modified task (estimation error) obtained after randomly sampling participant trials to simulate conducting the experiment with fewer trials. To simulate fewer trials per stimulus for each outcome measure, sets of 11 trials containing balanced stimulus durations were sampled with replacement from each participant's session, thus resulting in 12 trial number conditions (11 trials to 132 trials in steps of 11 trials). The outcome measure was then computed for each sample. A mean-rating, absolute agreement, two-way mixed-effects model intraclass correlation was then conducted on participants' sampled Weber ratios between the first and second sessions. These steps were repeated for 1000 permutations. The mean ICC across the 1000 permutations was calculated as an indication of test–retest reliability for each trial number condition.

2.4.6. Convergent Validity

To determine the convergent validity between the classic and modified bisection tasks, a Pearson's correlation was performed between participants' Weber ratio and percent correct from the classic task and mean estimation errors for the modified task. Percent correct was transformed into error rate (1 - proportion correct) to avoid negative correlations. Thus, lower score indicated a better performance for all outcome measures. We also estimated the underlying ('error-free') relationship between the classic and modified bisection task using confirmatory factor analysis (CFA). To do this, we estimated a model including two latent factors, one for each bisection task. Because each model only had two indicators (each session consisted of an indicator), the factor loadings for those indicators were constrained to be equal.

The CFA model was estimated using robust maximum likelihood estimation. Missing data were handled using full information maximum likelihood. For all models, we evaluated global model fit using the following indices and criteria: the chi-squared test (nonsignificant test indicates good fit), comparative fit index (CFI; >0.95), root mean square error of approximation (RMSEA; <0.05) and

standardized root mean squared residual (SRMR; <0.08). Furthermore, the local fit was assessed by inspecting the residual correlations for values greater than 0.10, which would indicate local misfit (Kline, 2016). Analyses were conducted in R (version 4.2.2) using the *lavaan* package (version 0.6.16).

3. Results

3.1. Descriptive Statistics

Figure 2 shows the proportion of 'long' responses for each stimulus duration averaged across participants. The average bisection point was 730 ms, slightly below the arithmetic mean between the two reference durations (750 ms). In the classic bisection task, participants had a mean Weber ratio of 0.077 (SD = 0.019, range [0.050, 0.139]). Participants had a mean percent correct score of 87.8% (SD = 5.81, range [71.7, 96.7]) across their responses.

For the modified bisection task, average estimations for each intermediate duration are plotted in Fig. 3. Overall, participants accurately estimated the intermediate durations in relation to the reference durations. The mean estimation error score was 0.17 (SD = 0.05, range [0.10, 0.30]).

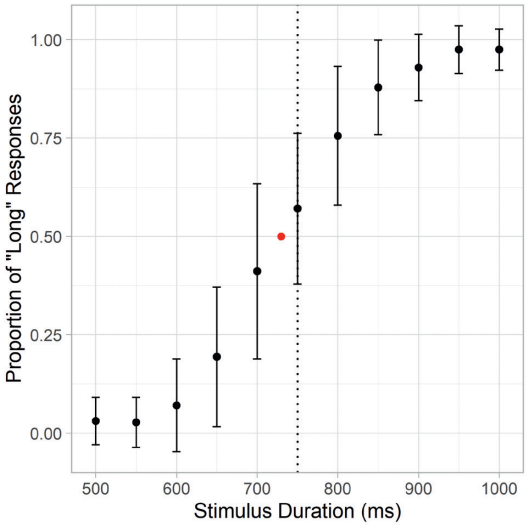

Figure 2. Proportion of 'long' responses versus stimulus duration in the classic bisection task.
Note. The red dot indicates the bisection point where 50% of responses are 'long' at 730 ms. The dotted line indicates the arithmetic mean between the two reference durations of 500 ms and 1000 ms. Error bars represent the standard deviation.

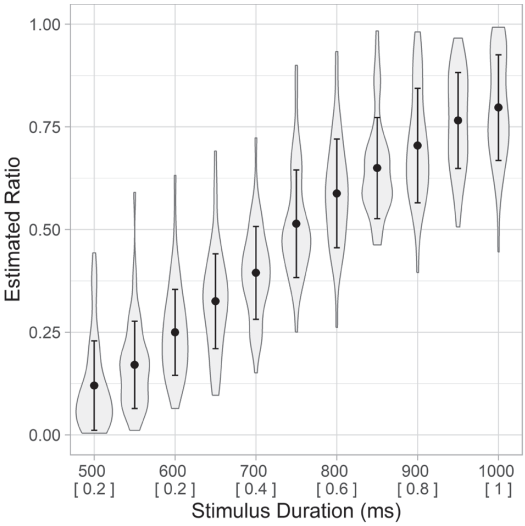

Figure 3. Estimated ratio on a visual analog scale versus stimulus duration in a modified bisection task.

Note. On the x-axis, numbers on top indicate stimulus duration in ms, while numbers in square brackets indicate the stimulus ratio. Points and error bars represent mean and standard deviation of estimated ratio for each stimulus duration. Violin plots depict the distribution of mean estimated ratio responses across all participants.

3.2. Internal Consistency

Table 1 depicts the internal consistency for the Weber ratio (classic bisection task), percent correct (classic bisection task), and the estimation error (modified bisection task). In the classic bisection task, SB-corrected correlations for the Weber ratio indicate moderate (0.50–0.75) internal consistency reliability for both sessions (0.56 in session 1, 0.55 in session 2) (Koo and Li, 2016). In contrast, SB-corrected correlations for the percent correct indicate good (0.75–0.90) internal consistency reliability for both sessions (0.80 in session 1, 0.87 in session 2) (Koo and Li, 2016). In the modified bisection task, the internal consistency of the estimation error was excellent (>0.90) for both sessions (0.95 in session 1, 0.95 in session 2) (Koo and Li, 2016).

Of note, there seems to be quite a discrepancy between the reliability of the Weber ratio and percent correct. This discrepancy may be due to the Weber ratio being computed based on the four intermediate durations forming the steepest slope rather than all trials like for percent correct. Furthermore, the intermediate durations in the current version of the task may be too near each other to reliably estimate the Weber ratio. To test this hypothesis, we did the same analysis

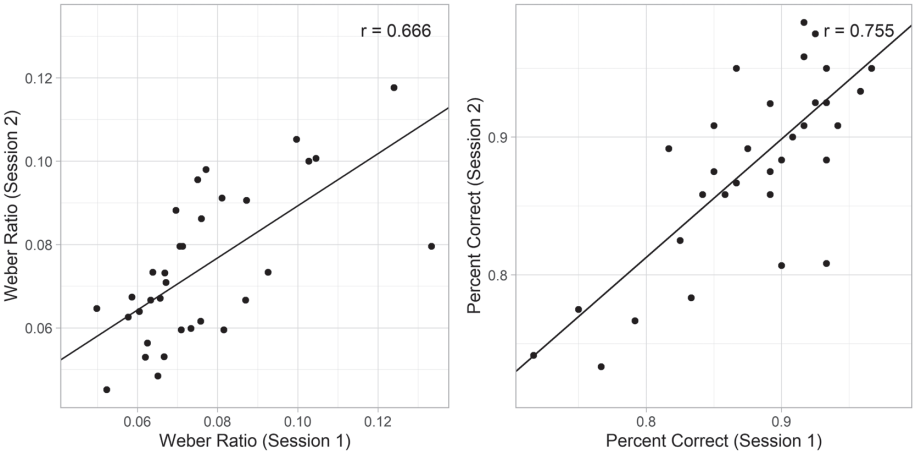

Figure 4. Relationship between scores on separate sessions for Weber ratio and percent correct in the classic bisection task.
Note. Solid line indicates linear regression between sessions 1 and 2.

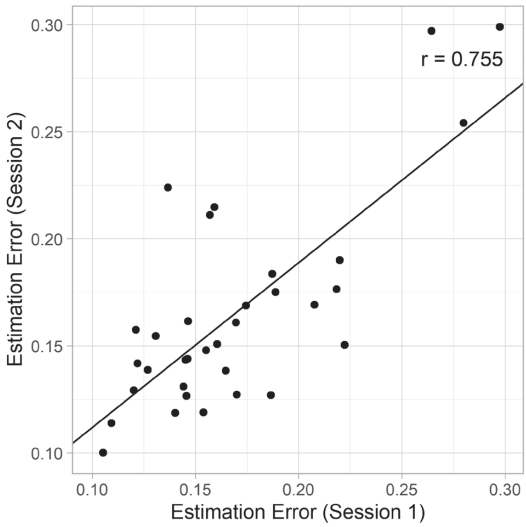

Figure 5. Estimation error scores in the modified bisection task between two sessions, the second taking place approximately seven days after the first.
Note. Solid line indicates linear regression between session 1 and 2.

as the one presented above on data from the first session, except that we limited the subset of intermediate durations to 500 ms, 600 ms, 700 ms, 800 ms, 900 ms, and 1000 ms. This more closely resembles data that could be obtained from a version of the task with fewer intermediate durations, thus further spread out. We found that percent correct now had an uncorrected split-half reliability of 0.40 CI [0.40, 0.41] (SB-corrected 0.58 CI [0.57, 0.58]). In comparison, Weber ratio had an uncorrected split-half reliability of 0.43 CI [0.42, 0.44] (SB-corrected 0.60 CI [0.59, 0.61]). Thus, reducing the number of trials used to calculate percent correct made the reliability coefficients much more similar to the Weber ratio, and using slightly more spread-out intermediate durations only marginally improved the internal consistency.

Finally, we wanted to assess whether having prior knowledge about the intermediate durations affected the reliability of the classic bisection task. Because the modified task trained participants to respond to intermediate durations while the instructions for the classic task never informed participants that intermediate durations are used, one could argue that participants who completed the modified

Table 1.
Internal consistency reliability for outcome measures from the classic and modified temporal bisection task.

Outcome measure	Session	*n*	Uncorrected correlation	Uncorrected 95% CI	SB-corrected correlation	SB-corrected 95% CI
Weber ratio (classic bisection task)	1	46	0.385	[0.378, 0.392]	0.556	[0.549, 0.563]
	2	33	0.378	[0.370, 0.386]	0.548	[0.540, 0.557]
Percent correct (classic bisection task)	1	46	0.667	[0.664, 0.671]	0.801	[0.798, 0.803]
	2	33	0.770	[0.767, 0.774]	0.870	[0.868, 0.872]
Estimation error (modified bisection task)	1	46	0.896	[0.895, 0.898]	0.945	[0.944, 0.946]
	2	33	0.910	[0.909, 0.912]	0.953	[0.952, 0.954]

Abbreviation: SB, Spearman–Brown.

task had prior knowledge compared to participants who completed the classic task first. This prior knowledge could affect the reliability of the classic, especially if asking participants to categorize intermediate intervals as short or long violates their perceptual experience and prior knowledge from the modified task. To investigate this, we reanalyzed the internal consistency of tasks completed in the first session separately for each counterbalanced order. This analysis was done on the first session only, as that is when this effect would be most apparent. Results are shown in Table 2. For the modified bisection task, separating the counterbalance order yielded near-identical corrected correlations compared to the original reliability found. For classic task outcome measures (Weber ratio and percent correct), internal consistency was near-identical to the original results (internal consistency combined across counterbalanced orders) when participants completed the classic task first and marginally lower when participants completed the modified task before the classic task. This indicates that prior knowledge or fatigue may influence the reliability of the classic task when done after the modified task.

Table 2.
Internal consistency for outcome measures from the first session of the classic and modified temporal bisection task, separated based on counterbalance condition.

Outcome measure	First task	*n*	Uncorrected correlation	Uncorrected 95% CI	SB-corrected correlation	SB-corrected 95% CI
Weber ratio (classic bisection task)	Classic	25	0.380	[0.371, 0.390]	0.551	[0.541, 0.560]
	Modified	21	0.320	[0.307, 0.332]	0.484	[0.470, 0.498]
Percent correct (classic bisection task)	Classic	25	0.668	[0.662, 0.673]	0.801	[0.797, 0.804]
	Modified	21	0.581	[0.574, 0.588]	0.735	[0.730, 0.740]
Estimation error (modified bisection task)	Classic	25	0.904	[0.903, 0.906]	0.950	[0.949, 0.951]
	Modified	21	0.899	[0.897, 0.901]	0.947	[0.945, 0.948]

Abbreviation: SB, Spearman–Brown.

3.3. Test–Retest Reliability

Table 3 shows the test–retest reliability for outcome measures from the classic bisection task (Weber ratio and percent correct) and the modified task (estimation error). Point estimates of test–retest reliability were similar and indicated good reliability. However, when considering the confidence intervals, the ICC for the Weber ratio indicates fair (0.50–0.75) to excellent (>0.90) test–retest reliability (Koo and Li, 2016). Similarly, ICCs for the percent correct from the classic task and estimation error from the modified bisection task indicate fair (nearly good) to excellent reliability.

Table 3.
Test–retest reliability for outcome measures from the classic and modified temporal bisection task.

Outcome measure	*n*	ICC	95% CI
Weber ratio (classic bisection task)	33	0.80	[0.60, 0.90]
Percent correct (classic bisection task)	33	0.86	[0.72, 0.93]
Estimation error (modified bisection task)	33	0.86	[0.72, 0.93]

Abbreviation: ICC, intraclass correlation.

3.4. Reliability and Trial Number

As an exploratory analysis, we assessed how different factors influence the reliability of the measures. First, we assessed how the number of repetitions influenced the test–retest reliability using bootstrapping methods. Results are shown in Fig. 6. When looking at the relationship between the number of trials and reliability, test–retest reliability estimates visually increased as the number of trials increased for all outcome measures. At all simulated numbers of trials, the mean ICC between first and second session scores was greatest for estimation error from the modified task, followed by percent correct from the classic task, and finally, the Weber ratio from the classic task.

3.5. Convergent Validity

In addition to assessing reliability, a convergent validity analysis was conducted between the classic and modified bisection tasks. Weber ratios for the classic task and mean estimation errors for the modified task were analyzed with Pearson's correlation using data from all participants in the first session. A significant positive correlation was found between the Weber ratio from the classic task and the estimated error from the modified task, $r(44) = 0.732$, CI [0.561, 0.843], $p < 0.01$. Similarly, we found a significant positive correlation between percent correct and estimation error, $r(44) = 0.672$, CI [0.474, 0.805], $p < 0.01$. As an exploratory

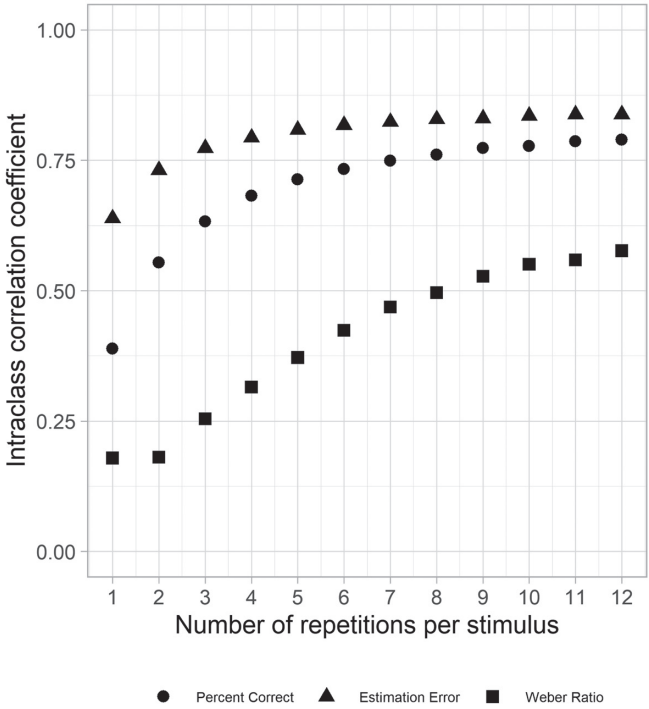

Figure 6. Relationship between number of trial repetitions and test–retest reliability for outcome measures.

Note. Eleven durations were included in the estimation error score. Ten durations were used for the percent correct score as the middle intermediate duration (750 ms) cannot be classified as short or long. The four consecutive durations forming the steepest slope were used to compute the Weber ratio.

analysis, we assessed the correlation between the measures from the classic and modified bisection tasks using CFA. The model showed adequate global and local fit. The chi-squared of significance was not statistically significant, $X^2(3) = 1.51$, $p = 0.679$ (Yuan–Bentler scaling correction factor = 0.816). Other global fit indices also showed good fit: robust CFI = 1.00, robust RMSEA = 0.00 (90% CI [0.00, 0.20]), SRMR = 0.037. Final model estimates are depicted in Fig. 7. Results show that the relationship between the classic and modified bisection was 0.94 (SE = 0.06, 95% CI [0.82, 1.05]). The same analysis was also conducted on the Weber ratio. However, because the model showed poor fit {$X^2(3) = 10.931$, $p = 0.012$, robust CFI = 0.945, robust RMSEA = 0.20 (90% CI [0.00, 0.41]), SRMR = 0.157}, model estimates are not reported.

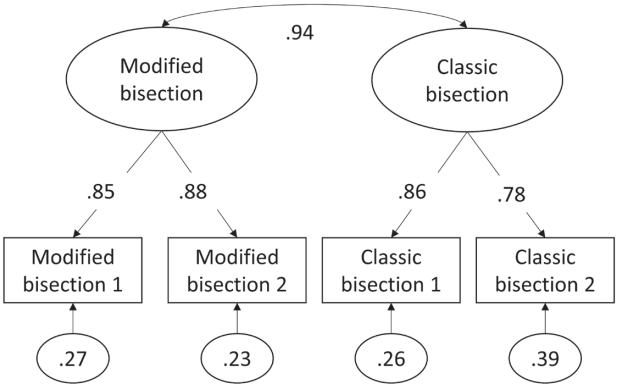

Figure 7. Latent correlation between the modified and classic bisection tasks.
Note. All coefficients are standardized. Error rate was used as the outcome measure for the classic bisection factor.

4. Discussion

This study had two goals: first, to examine humans' perception of intermediate durations in a modified temporal bisection task using a VAS, and second, to compare the psychometric properties of the modified temporal bisection task to the classic bisection task. While the classic bisection task has long been used to assess time perception and memory, it only provides indirect information on the perception of intermediate durations. That is, the classic bisection task only measures the probability of participants responding long but not how participants perceive the intermediate duration. We found that participants could accurately estimate the intermediate durations relative to the reference durations using a VAS. This modified bisection task provides a reliable way of explicitly measuring how humans perceive subsecond intermediate durations.

Our results also provide further information about the psychometric properties of the classic and modified bisection task for outcome measures of temporal judgement ability. Knowledge about the psychometric properties is important, as low reliability can attenuate effect sizes (e.g., correlations) and reduce power, especially when measuring individual differences. For the classic bisection task, percent correct had much better internal consistency than the Weber ratio, though it was still slightly lower than the estimation error. Estimation error from the modified bisection task demonstrated excellent internal consistency. The high internal consistency of the modified bisection task is likely due to the continuous, as opposed to binary, nature of the response format. For test–retest reliability, percent correct and estimation error had similar reliability estimates and confidence intervals spanning the upper limit of fair (i.e., nearly good) to excellent reliability.

While the point estimate of the test–retest reliability of the Weber ratio was only slightly lower, its confidence interval was much larger, making the reliability fair to excellent. Consistent with the results for internal consistency, these results might indicate that the Weber ratio may be a less stable measure than percent correct and estimation error.

However, there are a few discrepancies that need explanation. The discrepancy between the moderate internal consistency and fair to excellent test–retest reliability of the Weber ratio may be because the measures reflect different types of error (Chmielewski and Watson, 2009; Lakes and Hoyte, 2009; McCrae et al., 2011). Internal consistency measures the proportion of "true score variance [to] all variance that replicates over items" (Lakes and Hoyte, 2009, p. 3). In contrast, test–retest reliability measures the proportion of "true score variance [to] all variance that replicates over testing occasions" and reflects stability across testing occasions (Lakes and Hoyte, 2009, p. 3). Note that for both reliability coefficients, reliability was assessed on scores derived from balanced stimulus durations, meaning that differences in reliability cannot be attributed to differences in stimulus durations. Furthermore, internal consistency could have a lower coefficient because half the trials were used in the split-half procedure, whereas all trials were used to calculate the test–retest reliability. This explanation is supported by the findings in Fig. 6, which shows test–retest reliability estimates similar to the internal consistency estimates for the Weber ratio at six repetitions.

Another important discrepancy is between the Weber ratio and percent correct on the classic bisection task. This discrepancy can be explained by the fact that fewer trials, four out of the 11 durations encompassing the steepest slope, are used to estimate the Weber ratio while the percent correct takes all trials into account. In the current study, we assessed this possibility and found that the internal consistency of the percent correct and Weber ratio became much more similar when restricting the analysis to a subset of intermediate durations. Additionally, the difference in reliability between Weber ratio and percent correct may be moderated by the spread of the intermediate durations. For example, if one version of the bisection task has more intermediate durations, the points used to compute the Weber ratio will be closer to the bisection point than a version with fewer intermediate durations (assuming both versions use the same reference durations). We found some evidence for this hypothesis as the internal consistency of the Weber ratio was marginally better when the spread of the intermediate durations was increased.

Our findings are similar to those of Marx et al. (2021), who found that time estimation had the highest internal consistency (Cronbach's alpha) and test–retest reliability (ICC values from three test sessions). We found a similar pattern in which the task with a continuous response format (the modified bisection task) had the highest reliability. Marx et al. (2021) also found that time discrimination

using a staircase procedure generally yielded lower ICC values than the other timing paradigms they tested. Their findings concur with the results found in the current study, which showed that tasks requiring binary responses had worse psychometric properties than those requiring continuous responses.

In addition to assessing the psychometric properties of the modified and bisection task, we assessed how the number of trials impacted test–retest reliability. Using permutation-based random sampling of participant data, we simulated the test–retest reliability results for each outcome using various trial numbers. Test–retest reliability increased as simulated trial numbers increased for all outcome measures. Furthermore, the test–retest reliability for estimation error was followed by percent correct and Weber ratio across all trial number conditions. Visually, test–retest reliability for estimation error appeared to stabilize after approximately four repetitions per stimulus (44 trials/four repetitions). In contrast, for the classical bisection task, test–retest reliability for percent correct stabilized after approximately six repetitions per stimulus (60 trials/six repetitions), while the Weber ratio reliability stabilized only after 10 repetitions per stimulus (110 trials/10 repetitions). This indicates that, when using the Weber ratio to measure time perception, more trials may be needed to reach similar reliability compared to other measures. Historically, studies that have used the classic bisection task to investigate individual differences have included between 35 and 105 trials and used the Weber ratio to measure time perception (Allman et al., 2011; Carroll et al., 2008; Droit-Volet et al., 2015; Elvevåg et al., 2003; Nichelli et al.,1995, 1996; Ogden et al., 2018; Sadibolova et al., 2022). Our results indicate that more trials than are traditionally used might be necessary to obtain a reliable measure of time perception. Additionally, researchers may want to consider using percent correct when using the classic bisection task or using the modified bisection task to measure timing ability, especially if the goal is to correlate this measure with other measures or if the research question does not require generalizing the measure across different reference durations.

Convergent validity analysis between the classic and modified bisection tasks indicated a strong positive correlation, suggesting good convergent validity between measures (Grobler and Joubert, 2018). When using a latent modelling framework (CFA) to account for measurement error, the convergent validity between classic and modified task accuracy scores was high, with the confidence interval overlapping the value of 1 (Nicewander, 2018). This high correlation provides evidence of convergence, suggesting that the two tasks are likely measuring the same underlying construct (Grobler and Joubert, 2018).

A limitation of the current study stems from our counterbalancing method: participants who completed the modified task first might have inferred additional information about the stimuli compared to those who completed the classic task first. While the instructions for the classic task never informed participants that intermediate durations were used, the modified task trained participants to

respond to intermediate durations. We performed an exploratory analysis on the internal consistency of tasks done in the first session for each counterbalance order, where this effect would be most apparent. Results were similar to the original analysis with one exception: the reliability of the classic bisection task was lower when participants completed the classic after the modified bisection task. One possibility is that prior knowledge of intermediate durations led to a violation of their perceptual experience, amplified by asking participants to classify an intermediate duration as short or long rather than judge the similarity to the reference durations. Another possibility is that the reliability of the classic bisection task is more vulnerable to fatigue effects.

The current study has implications for researchers wishing to measure time perception abilities as it can guide the choice of task and measure in future studies. The excellent reliability of the modified temporal bisection task, along with its excellent convergent validity with the classic bisection task, supports its use as a measure of temporal judgement accuracy. However, the classic bisection task still offers several advantages depending on the research goal. For example, the classic bisection task may be optimal in situations involving subjects with limited response capabilities, such as animal, paediatric, or geriatric populations (Droit-Volet and Wearden, 2001; McCormack et al., 1999; Siegel and Church, 1984). As another example, the advantages of the Weber ratio include making results from studies using different reference durations comparable. However, if the goal is to obtain a reliable measure of time perception for the purpose of studying individual differences, other measures such as percent correct or using the modified bisection task may be more reliable and efficient. Based on our results, the Weber ratio required significantly more trials before attaining a stable test–retest reliability, and its reliability was consistently lower than that of other measures. Furthermore, researchers might benefit from significantly increasing the number of trials to more than ten repetitions of each intermediate duration if planning to use the Weber ratio. Regardless of how time perception is measured, reliability analyses should be systematically performed alongside its use in future studies. In cases where there are too few trials to properly estimate psychophysical measures, it is possible to use a 'sample-with-replacement' bootstrap method to estimate the reliability of psychophysical measures (Anobile et al., 2016; Efron and Tibshirani, 1993).

5. Conclusion

Overall, the current study's aims were to investigate how humans quantify relative durations in a novel temporal bisection task using a continuous response format (VAS) and to assess its reliability along with the reliability of the classic temporal bisection task. Humans could accurately estimate the relative duration of intermediate durations using a visual analogue scale for durations between 500 and

1000 ms. The modified bisection task also showed adequate internal consistency and test–retest reliability. While the internal consistency and test–retest reliability for the percent correct on the classic bisection were adequate, the Weber ratio showed much poorer psychometric properties. Finally, we examined the number of trials required to obtain adequate reliability for both the classic and modified bisection and made recommendations for future research interested in studying individual differences in time perception using temporal bisection paradigms.

Acknowledgements

This work was funded by the James S. McDonnell Foundation, which provided the Scholar Award to Dr Jessica Grahn (DOI:10.37717/220020403), and the Natural Sciences and Engineering Research Council of Canada, which supported Dr Jessica Grahn through the Steacie Fellowship.

References

Allan, L. G. (1978). Comments on current ratio-setting models for time perception. *Percept. Psychophys*. *24*, 444–450. doi: 10.3758/BF03199742.

Allan, L. G. (2002). The location and interpretation of the bisection point. *Q. J. Exp. Psychol. B*, *55*, 43–60. doi: 10.1080/02724990143000162.

Allan, L. G. & Gibbon, J. (1991). Human bisection at the geometric mean. *Learn. Motiv.*, *22*, 39–58. doi: 10.1016/0023-9690(91)90016-2.

Allman, M. J. & Meck, W. H. (2012). Pathophysiological distortions in time perception and timed performance. *Brain*, *135*, 656–677. doi: 10.1093/brain/awr210.

Allman, M. J., DeLeon, I. G. & Wearden, J. H. (2011). Psychophysical assessment of timing in individuals with autism. *Am. J. Intellect. Dev. Disabil.*, *116*, 165–178. doi: 10.1352/1944-7558-116.2.165.

Anobile, G., Castaldi, E., Turi, M., Tinelli, F. & Burr, D. C. (2016). Numerosity but not texture-density discrimination correlates with math ability in children. *Dev. Psychol.*, *52*, 1206–1216. doi: 10.1037/dev0000155.

Carroll, C. A., Boggs, J., O'Donnell, B. F., Shekhar, A. & Hetrick, W. P. (2008). Temporal processing dysfunction in schizophrenia. *Brain Cogn.*, *67*, 150–161. doi: 10.1016/j.bandc.2007.12.005.

Chmielewski, M. & Watson, D. (2009). What is being assessed and why it matters: the impact of transient error on trait research. *J. Pers. Soc. Psychol.*, *97*, 186–202. doi: 10.1037/a0015618.

Church, R. M. & Deluty, M. Z. (1977). Bisection of temporal intervals. *J. Exp. Psychol. Anim. Behav. Process.*, *3*, 216–228. doi: 10.1037//0097-7403.3.3.216.

Church, R. M. & Gibbon, J. (1982). Temporal generalization. *J. Exp. Psychol. Anim. Behav. Process.*, *8*, 165–186. doi: 10.1037/0097-7403.8.2.165.

Clynes, M. & Walker, J. (1986). Music as time's measure. *Music Percept.*, *4*, 85–119. doi: 10.2307/40285353.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (*2nd ed.*). Routledge, New York, NY, USA. doi: 10.4324/9780203771587.

Corcoran, A. W., Groot, C., Bruno, A., Johnston, A. & Cropper, S. J. (2018). Individual differences in first- and second-order temporal judgment. *PLoS One*, *13*, e0191422. doi: 10.1371/journal.pone.0191422.

Droit-Volet, S. (2003). Alerting attention and time perception in children. *J Exp. Child Psychol.*, *85*, 372–384. doi: 10.1016/S0022-0965(03)00103-6.

Droit-Volet, S. & Wearden, J. H. (2001). Temporal bisection in children. *J. Exp. Child Psychol.*, *80*, 142–159. doi: 10.1006/jecp.2001.2631.

Droit-Volet, S., Clément, A. & Fayol, M. (2003). Time and number discrimination in a bisection task with a sequence of stimuli: a developmental approach. *J. Exp. Child Psychol.*, *84*, 63–76. doi: 10.1016/S0022-0965(02)00180-7.

Droit-Volet, S., Meck, W. H. & Penney, T. B. (2007). Sensory modality and time perception in children and adults. *Behav. Processes*, *74*, 244–250. doi: 10.1016/j.beproc.2006.09.012.

Droit-Volet, S., Wearden, J. H. & Zélanti, P. S. (2015). Cognitive abilities required in time judgment depending on the temporal tasks used: a comparison of children and adults. *Q. J. Exp. Psychol.* (*Hove*), *68*, 2216–2242. doi: 10.1080/17470218.2015.1012087.

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall. Piubdoi: 10.1007/978-1-4899-4541-9.

Elvevåg, B., McCormack, T., Gilbert, A., Brown, G. D. A., Weinberger, D. R. & Goldberg, T. E. (2003). Duration judgements in patients with schizophrenia. *Psychol. Med.*, *33*, 1249–1261. doi: 10.1017/s0033291703008122.

García-Pérez, M. A. (2014). Does time ever fly or slow down? The difficult interpretation of psychophysical data on time perception. *Front. Hum. Neurosci.*, *8*, 415. doi: 10.3389/fnhum.2014.00415.

Grobler, A. & Joubert, Y. T. (2018). Psychological capital: convergent and discriminant validity of a reconfigured measure. *S. Afr. J. Econ. Manag. Sci.*, *21*, a1715. doi: 10.4102/sajems.v21i1.1715.

Hedge, C., Powell, G. & Sumner, P. (2018). The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods*, *50*, 1166–1186. doi: 10.3758/s13428-017-0935-1.

Karşılar, H., Kısa, Y. D. & Balcı, F. (2018). Dilation and constriction of subjective time based on observed walking speed. *Front. Psychol. 9*, 2565. doi: 10.3389/fpsyg.2018.02565.

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling* (*4th ed.*). Guilford Press, New York, NY, USA.

Koo, T. K. & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.*, *15*, 155–163. doi: 10.1016/j.jcm.2016.02.012.

Kopec, C. D. & Brody, C. D. (2010). Human performance on the temporal bisection task. *Brain Cogn.*, *74*, 262–272. doi: 10.1016/j.bandc.2010.08.006.

Lagacé-Cusiac, R., Tremblay, P. F., Ansari, D. & Grahn, J. A. (2023). Investigating the relationships between temporal and spatial ratio estimation and magnitude discrimination using structural equation modeling: Evidence for a common ratio processing system. *J. Exp. Psychol. Hum. Percept. Perform.*, *49*, 108–128. doi: 10.1037/xhp0001062.

Lakes, K. D. & Hoyt, W. T. (2009). Applications of generalizability theory to clinical child and adolescent psychology research. *J. Clin. Child. Adolesc. Psychol.*, *38*, 144–165. doi: 10.1080/15374410802575461.

Lambrechts, A., Walsh, V. & van Wassenhove, V. (2013). Evidence accumulation in the magnitude system. *PLoS One*, *8*, e82122. doi: 10.1371/journal.pone.0082122.

Lindbergh, C. A. & Kieffaber, P. D. (2013). The neural correlates of temporal judgments in the duration bisection task. *Neuropsychologia*, *51*, 191–196. doi: 10.1016/j.neuropsychologia.2012.09.001.

Liu, P., Guo, H., Ma, R., Liu, S., Wang, X., Zhao, K., Tan, Y., Tan, S., Yang, F. & Wang, Z. (2022). Identifying the difference in time perception between major depressive disorder and bipolar depression through a temporal bisection task. *PLoS One*, *17*, e0277076. doi: 10.1371/journal.pone.0277076.

Macar, F. & Vidal, F. (2004). Event-related potentials as indices of time processing: a review. *J. Psychophysiol.*, *18*, 89–104. doi: 10.1027/0269-8803.18.23.89.

Marx, I., Rubia, K., Reis, O. & Noreika, V. (2021). A short note on the reliability of perceptual timing tasks as commonly used in research on developmental disorders. *Eur. Child Adolesc. Psychiatry*, *30*, 169–172. doi: 10.1007/s00787-020-01474-y.

Matell, M. S. & Meck, W. H. (2000). Neuropsychological mechanisms of interval timing behavior. *Bioessays*, *22*, 94–103. doi: 10.1002/(SICI)1521-1878(200001)22:1<94::AID-BIES14>3.0.CO;2-E.

McCormack, T., Brown, G. D. A., Maylor, E. A., Darby, R. J. & Green, D. (1999). Developmental changes in time estimation: comparing childhood and old age. *Dev. Psychol.*, *35*, 1143–1155. doi: 10.1037/0012-1649.35.4.1143.

McCrae, R. R., Kurtz, J. E., Yamagata, S. & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Pers. Soc. Psychol. Rev.*, *15*, 28–50. doi: 10.1177/1088868310366253.

Mendez, J. C., Prado, L., Mendoza, G., & Merchant, H. (2011). Temporal and spatial categorization in human and non-human primates. *Front. Integr. Neurosci.*, *5*, 50. doi: 10.3389/fnint.2011.00050.

Momi, D., Prete, G., Di Crosta, A., La Malva, P., Palumbo, R., Ceccato, I., Bartolini, E., Palumbo, R., Mammarella, N., Fasolo M. & Di Domenico, A. (2023). Time reproduction, bisection and doubling: a novel paradigm to investigate the effect of the internal clock on time estimation. *Psychol. Res.*, *87*, 1549–1559. doi: 10.1007/s00426-022-01745-0.

Nakajima, Y. (1987). A model of empty duration perception. *Perception*, *16*, 485–520. doi: 10.1068/p160485.

Nicewander, W. A. (2018). Modifying Spearman's attenuation equation to yield partial corrections for measurement error — with application to sample size calculations. *Educ. Psychol. Meas.*, *78*, 70–79. doi: 10.1177/0013164417713571.

Nichelli, P., Clark, K., Hollnagel, C. & Grafman, J. (1995). Duration processing after frontal lobe lesions. *Ann. N. Y. Acad. Sci.*, *769*, 183–190. doi: 10.1111/j.1749-6632.1995.tb38139.x.

Nichelli, P., Alway, D. & Grafman, J. (1996). Perceptual timing in cerebellar degeneration. *Neuropsychologia*, *34*, 863–871. doi: 10.1016/0028-3932(96)00001-2.

Ogden, R. S., Samuels, M., Simmons, F., Wearden, J. & Montgomery, C. (2018). The differential recruitment of short-term memory and executive functions during time, number, and length perception: an individual differences approach. *Q. J. Exp. Psychol.* (*Hove*), *71*, 657–669. doi: 10.1080/17470218.2016.1271445.

Ortega, L. & López, F. (2008). Effects of visual flicker on subjective time in a temporal bisection task. *Behav Processes*, *78*, 380–386. doi: 10.1016/j.beproc.2008.02.004.

Parsons, S., Kruijt, A.-W. & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Adv. Methods Pract. Psychol. Sci.*, *2*, 378–395. doi: 10.1177/2515245919879695.

Penney, T. B., Gibbon, J. & Meck, W. H. (2000). Differential effects of auditory and visual signals on clock speed and temporal memory. *J. Exp. Psychol. Hum. Percept. Perform.*, *26*, 1770–1787. doi: 10.1037//0096-1523.26.6.1770.

Provasi, J., Rattat, A.-C. & Droit-Volet, S. (2011). Temporal bisection in 4-month-old infants. *J. Exp. Psychol. Anim. Behav. Process.*, *37*, 108–113. doi: 10.1037/a0019976.

Read, J. C. A. (2015). The place of human psychophysics in modern neuroscience. *Neuroscience*, *296*, 116–129. doi: 10.1016/j.neuroscience.2014.05.036.

Sadibolova, R., Monaldi, L. & Terhune, D. B. (2022). A proxy measure of striatal dopamine predicts individual differences in temporal precision. *Psychon. Bull. Rev.*, *29*, 1307–1316. doi: 10.3758/s13423-022-02077-1.

Siegel, S. F. & Church, R. M. (1984). The decision rule in temporal bisection. *Ann. N. Y. Acad. Sci.*, *423*, 643–645. doi: 10.1111/j.1749-6632.1984.tb23481.x.

Spearman, C. (1907). Demonstration of formulæ for true measurement of correlation. *Am. J. Psychol.*, *18*, 161–169. DOI: 10.2307/1412408.

Spearman, C. (1910). Correlation calculated from faulty data. *Br. J. Psychol. 1904–1920*, *3*, 271–295. doi: 10.1111/j.2044-8295.1910.tb00206.x.

Tipples, J., Brattan, V. & Johnston, P. (2013). Neural bases for individual differences in the subjective experience of short durations (less than 2 seconds). *PLoS One*, *8*, e54669. doi: 10.1371/journal.pone.0054669.

Wearden, J. H. (1991). Human performance on an analogue of an interval bisection task. *Q. J. Exp. Psychol. B*, *43*, 59–81. doi: 10.1080/14640749108401259.

Wearden, J. H. & Ferrara, A. (1995). Stimulus spacing effects in temporal bisection by humans. *Q. J. Exp. Psychol. B*, *48*, 289–310. doi: 10.1080/14640749508401454.

Wearden, J. H. & Ferrara, A. (1996). Stimulus range effects in temporal bisection by humans. *Q. J. Exp. Psychol. B*, *49*, 24–44. doi: 10.1080/713932615.

Wearden, J. H., & Jones, L. A. (2007). Is the growth of subjective time in humans a linear or nonlinear function of real time? *Q. J. Exp. Psychol.*, *60*, 1289–1302. doi: 10.1080/17470210600971576.

Wearden, J. H., Rogers, P. & Thomas, R. (1997). Temporal bisection in humans with longer stimulus durations. *Q. J. Exp. Psychol. B*, *50*, 79–94. doi: 10.1080/713932643.