2021 ■ Vol. 17 ■ no. 4

Big-M-Small-N Temporal-Order Judgment Data

Jan Tünnermann [a] and Ingrid Scharlau [b]

[a]Department of Psychology, Phillips University Marburg, Marburg, Germany
[b]Department of Arts and Humanities, Paderborn University, Paderborn, Germany

Abstract ■ We present a large and precise data set of temporal-order judgments on visual stimuli. Stimulus asynchronies ranged from 0 to 80 ms in steps of 6.67 ms. The data set includes a salience-based attention manipulation driven by one target's orientation compared to background elements (either zero or 90 degrees). Each of 25 stimulus asynchronies was sampled with at least 196 repetitions (and beyond 400 repetitions in two participants). Furthermore, fixation, an important concern in studies on covert attention, was monitored. Precise data are helpful for answering theoretical questions in psychology. For some questions such as model comparisons, they may even be necessary. Three different example models are fitted to the data.

Keywords ■ Temporal-order judgments, theory of visual attention.

jan.tuennermann@uni-marburg.de

10.20982/tqmp.17.4.p355

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer

Introduction

Modeling is an important part of psychology. Although one can discuss whether or how far psychology right now faces a methods crisis (for different opinions see, for instance, Bakker, van Dijk, & Wicherts, 2012; Gilbert, King, Pettigrew, & Wilson, 2016; Ioannidis, 2005; Maxwell, Lau, & Howard, 2015; Open Science Collaboration, 2015; Pashler & Harris, 2012; Simmons, Nelson, & Simonsohn, 2011), it is undisputed that enhanced methods will substantially contribute to progress in psychology. Formal modeling is among them (see. e.g., Krüger, Tünnermann, Rohlfing, & Scharlau, 2018).

Modeling presupposes appropriate data sets. Disparate models, even two that differ quite substantially in the processes they call on to explain a certain phenomenon or effect, may barely differ in predicted data patterns (e.g., Tünnermann & Scharlau, 2018b). In this case, very precise data are required to convincingly distinguish between models, or estimate parameters with high confidence.

In areas of cognitive research where data collection is extremely expensive, researchers have already started creating high-quality large-scale data sets which then are made available to the research community. To name just one example, Hanke et al. (2014, 2016) curate an openly available and constantly extended data set with fMRI and structural scans, eye-tracking, and other auxiliary data. This data set was obtained with participants who watched the movie Forrest Gump in an fMRI scanner. The data set has facilitated very different studies—such as those on the processing of event boundaries in continuous experiences (Ben-Yakov & Henson, 2018) or the direct comparison of brain activity across participants (Joshi, Chong, Li, Choi, & Leahy, 2018).

Although collection of behavioral data is quicker and much less expensive, behavioral studies could benefit from similar collaborative efforts. For instance, temporal-order judgments (TOJs) are often used to infer causal influences on processing speed, such as advantages—to give a few recent examples—caused by attending to a location (Shore, Spence, & Klein, 2001), by task relevance and bottom-up salience (Born, Kerzel, & Pratt, 2015), by threatening and non-threatening faces (West, Anderson, & Pratt, 2009), or by being the active compared to the non-active object in a pair (such as a cork screw compared to the corked bottle; K. L. Roberts & Humphreys, 2010). Most of these studies demonstrated the assumed influences by showing that perceived temporal characteristics differ between conditions. This temporal perception is usually operationalized as a descriptive parameter of the distribution of judg-

ments, the point of subjective simultaneity (PSS), assuming that a shifted PSS indicates faster or slower processing speed. Although this method is very common, we have to note that the relationship between the PSS, which describes observed behavior, and the underlying temporal processing is at best indirect. Without *formally* linking the PSS (or other features of the order-judgment distribution) to the mechanisms that should explain the effects, little can be learned about what exactly causes them.

There are some approaches at modeling TOJs in more detail. For instance, Schneider and Bavelier (2003) compared three types of models already identified by Sternberg and Knoll (1973), a deterministic decision rule model, a triggered-moment model, and a perceptual-moment model, and found that sensory facilitation and modulations of the decision mechanism caused reliable acceleration of processing ("prior entry") whereas attention seemed to be less relevant as a causal influence. More recently, García-Pérez and Alcalá-Quintana (2018) postulated an indecision-range observer model with processing speed, latency, decision, and response factors as parameters. Unlike Schneider and Bavelier, García-Pérez and Alcalá-Quintana ascribe prior entry to decision processes. A further approach (e.g., Tünnermann & Scharlau, 2016) is to model TOJs with the fundamental components of attentional processing from Bundesen's (1990) theory of visual attention (TVA; for a review see Bundesen, Vangkilde, & Petersen, 2015). In this perspective, effects such as prior entry can be linked to theoretical components whose existence and meaning are supported by data from entirely different behavioral paradigms, clinical research, and neural theories. We pick up some of the approaches described above later for example evaluations of the high-accuracy TOJ data set we present in this paper.

Brief methodological context

TOJs have been a common experimental methodology in psychology ever since its beginnings (e.g., Boring, 1957; Hoffmann, 2006), especially, but not exclusively, in the study of prior entry (Spence & Parise, 2010). For enabling high-confidence analyses, the present data set focuses on data quality within an individual observer's data set (number of measuring points and repetitions, i.e., a larger number of trials M per participant, "big-M") and therefore only tests few participants (a small-N design, see, e.g., Smith & Little, 2018). As an example question, we study the influence of salience on temporal-order perception (e.g., Krüger, Tünnermann, & Scharlau, 2017).

Search for a Powerful Design

Before conducting the experiment delineated above, we analyzed power with the following procedure. Within participants, the experimental power of TOJs is determined by at least three factors, the range and spacing of the stimulus onset asynchronies (SOAs)[1] to be judged and the repetitions of each SOA. We aim to produce a versatile TOJ data set for advanced analysis and comparisons of models that deal with minute fluctuations in the psychometric function (cf. Tünnermann & Scharlau, 2018b). Hence, we opt for a tight SOA spacing of 6.67 ms, which can be reliably presented on a 150 Hz CRT monitor. Our SOAs range from -80 to +80 ms, covering the complete psychometric functions we typically observe for our stimulus material (see, e.g., Krüger, Tünnermann, & Scharlau, 2016). With these factors fixed, the number of SOA repetitions is the factor that can be adjusted to achieve the desired power.

We illuminate the relationship between SOA repetitions and power in a general manner to help researchers judge the power of the present data set and visual TOJ data sets in general. Especially with advanced model-based analysis, the power depends on the model and the effect (size) of interest, which we cannot anticipate. However, a widely used and general model of binary TOJs is a sigmoid function with a PSS (point of subjective simultaneity; the SOA at which stimuli are perceived as simultaneous) and DL (difference limen; an index of the function's slope and an indicator of discrimination performance). Because of its widespread use and comparative simplicity, we use this model for the following power analyses.

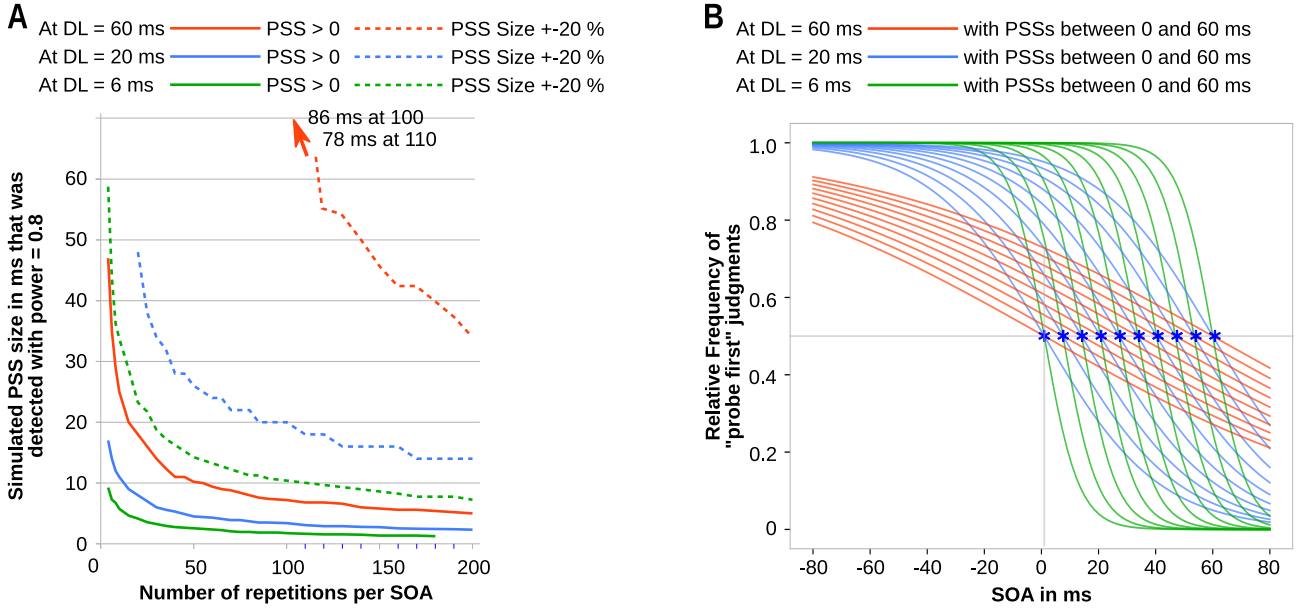
TOJ researchers might be interested in several results of this analysis. Often, they want to establish that a PSS is different from zero, for instance that centrally cued stimuli are perceived faster than uncued ones (Shore et al., 2001), or that stimuli in the left visual field are processed faster than those in the right visual field (Matthews & Welch, 2015). Other researchers should be more interested in the size of a PSS difference than its existence because this size is often taken as an index of changes in processing speed.

To this end, we conduct a novel Bayesian power search (BPS). In the BPS we are not interested in the power of detecting one particular simulated effect but in the PSS sizes that can be detected with a desired power depending on the number of SOA repetitions.

We therefore systematically search through PSS-size candidates by simulating data with certain PSSs with a logistic model (Finney, 1971). Starting with one PSS candidate, multiple data sets (500) are simulated by drawing from binomial distributions at each SOA with a success rate

[1] The conventional term is "stimulus onset asynchrony" in the TOJ literature, and we stick with it, although strictly speaking we use a stimulus blink asynchrony, as explained below in the Summary of the Methodology.

Figure 1 ■ (a) Number of repetitions × PSS shift size 0.8 "power curves" obtained with the Bayesian power search. Researchers can read off the number of repetitions (x-axis) required to detect PSS shifts of different sizes (y-axis) with a power of 0.8. Successful detection refers to PSS posterior distributions for which zero, no difference, lies below the lower 95-%-HDI boundary (solid lines) or the posterior mode being within $\pm 20\%$ of the true (simulated) PSS (dashed lines). PSS: point of subjective simultaneity; DL: difference limen. (b) Illustrations of psychometric functions with different DLs and PSSs.



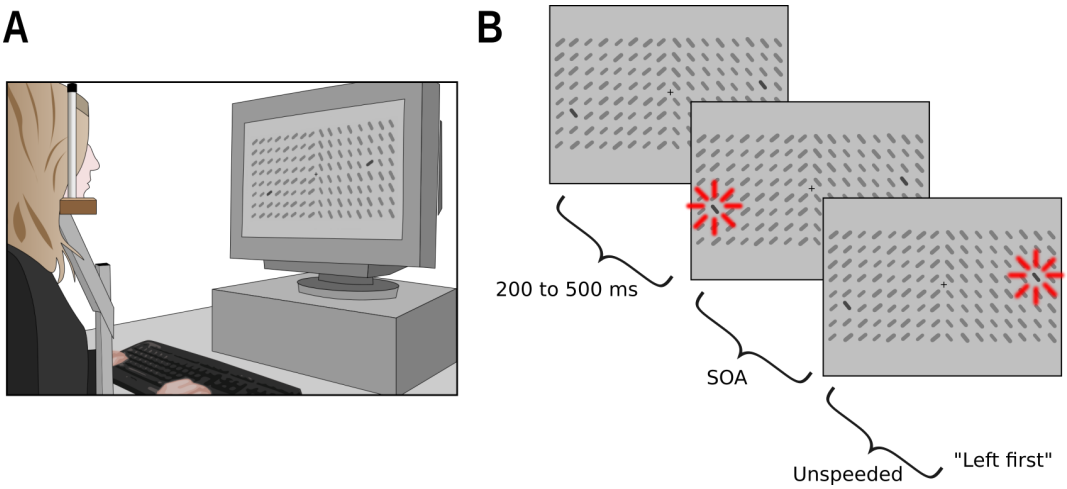
determined by the current psychometric function and according to the current number of repetitions. A Bayesian parameter estimation is then conducted for all data sets. The power is the proportion of estimations in which the research goal (e.g., a PSS larger than zero) is achieved. If this turns out to be smaller than 0.8[2], a larger candidate (between the current and an upper limit) is used in the next iteration. If the power is larger than the desired power (e.g., > 0.8), the new candidate is chosen in the middle between zero (the lower limit) and the current value. The upper and lower limits of the search range are always updated based on which parts of the search space can be excluded. In this fashion, a bisection search is performed to find the candidate that is the closest to the desired power.[3] Once the search has homed in on the best candidate (which has a power close to 0.8), this value is stored for the current number of repetitions and the same procedure is performed for the next larger number of repetitions. Because with more repetitions smaller PSSs can be found with a power of 0.8, the maximum candidate size for the new iteration can be set to the PSS size just found with a smaller number of iterations, facilitating the search.

We perform the BPS for different DL values that correspond to typical weak, medium and high accuracy in visual TOJs, 60, 20, and 6 ms. The reason for this is that "weaker curves" are closer to the chance level and have higher uncertainties in the binomial distributions, which has an impact on the power. Moreover, we look at the two different research goals already mentioned above: Detecting a PSS different from zero (something akin to a typical significance test) and measuring the "true" (simulated) PSS with an accuracy of ± 20 %. The resulting relationship is illustrated in Figure 1A. As the curves in Figure 1A show, many repetitions are required if researchers are interested in small PSSs (e.g., smaller than 10 ms). If DL is large (weak discrimination performance) or if the size and not just the presence of a PSS shift is of interest (dashed lines), hundreds of repetitions are required. Beyond PSS effects, if

[2]The value of 0.8 is conventional and should be adapted depending on which question is asked and how serious a beta error is assumed to be in a specific field of research.

[3]Of course the candidate list must be established to include sufficiently close values. This is best determined by a few preliminary power estimations of the type described here for some test values.

Figure 2 ■ (a) General setup (image adapted from Krüger et al., 2021). (b) Presentation procedure. The 20 ms blinks are indicated by the red markings (not present in the actual displays). Salient condition. Here, a negative SOA is shown at which the salient target flickered first.



models that produce very similar judgment distributions are to be compared, an even larger amount of data might be required (this is not illustrated in the figure; but see Tünnermann & Scharlau, 2018b). Hence, for the proposed data set we ensured that the number repetitions at each SOA are as high as possible. The data sets of different participants have average repetition counts per SOA of at least 196, some up to more than 470 (see Figures 3 and 4).

We hasten to add that we (Tünnermann, 2016) and others (e.g., Alcalá-Quintana & García-Pérez, 2013) have criticized the identification of PSS differences with processing speed differences on the ground (mentioned briefly in the Introduction) that the PSS lacks a formal connection to the processes that drive temporal perception. The PSS is a parameter that conveniently describes observer performance, that is, the psychometric function. Because formal modeling of TOJs is still uncommon whereas the use of PSS is widespread, we carried out the power analysis for this parameter. The same analysis could be repeated for parameters drawn from theory-based formal models such as that of Alcalá-Quintana and García-Pérez, 2013 or our own (Tünnermann, 2016).

The algorithm to perform the BPS is presented in Appendix A.

Summary of the Methodology

The apparatus for this study was a PC (with an Intel Core 2 Duo CPU 3.00GHz, 4 GB RAM; Windows 7) with a Samsung SyncMaster 957DF CRT monitor running at 640 × 480 pixels with 150 Hz (non-interlaced) for accurate stimulus presentation. The experiment was implemented in OpenSesame 3.1.9 (Mathôt, Schreij, & Theeuwes, 2012) with the PsychoPy (Peirce et al., 2019) backend. Blocking flips synchronized presentation with the vertical retrace of the monitor and the PC's internal clock was used to monitor for missed flips. Trials with missed flips were very rare, signaled via a blue screen background to the participants, and repeated later in the experiment. Trials with such timing errors were not included in the final dataset. A standard keyboard was used to collect the unspeeded judgments. An eye tracker (SR Research EyeLink 1000 plus) was used to ensure central fixation.

In the present study, stimuli were unimodal and visual. Judgment was binary, that is, observers judged whether one or the other visual target is first (for other methods see, e.g., Ulrich, 1987). The experiment had two intermixed independent variables, the SOA between the two targets (ranging from 0 to 80 ms in steps of 6.67 ms to cover the whole range of accuracy, from very few errors to guessing at SOA zero), and orientation salience. For each trial, the orientation of the background elements for the left half of the screen was drawn at random from 18 equally spaced orientations within the range of 0° to 170° (with 10° intervals; note that starting at 180° the appearance of the line segments repeats). The orientation of the background elements in the right half of the screen was obtained by adding 90°, the maximum possible difference, to the orientations of the left-field elements (see Figure 2B for an example). Each of the screen halves contained one of the targets. One target had the same orientation as the elements

surrounding it and thus was non-salient. In half of the trials, the other target had a maximally different orientation (90°) relative to its background (salient condition; see Figure 2). In the other half, it had the same orientation as its background elements (non-salient condition). We call the target that always had the same orientation as the background the reference stimulus, r, and the target that could be salient or not the probe stimulus, p. Note that, as illustrated in Figure 2, the left and right distractor fields had element orientations which were orthogonal to each other. Distractors were dark gray [R:127, G:127, B:127] bars on a light gray [R:192, G:192, B:192] background, 8×8 on each side of fixation. The targets were a darker gray [R:71, G:71, B:71] and appeared at random locations within the inner 6×6 stimuli, one at each side of fixation. The exact positions and line width were slightly jittered to create a less regular appearance. Each target blinked shortly (being turned off for 20 ms), separated by the SOA. On the 22" monitor at 50 cm distance. The length of the bars was 1.37° and the strength of the stroke was about 0.32° (plus a small jitter).

The participants indicated without time pressure which target had blinked first by pressing a left or right key. Fixation was controlled by an EyeLink 1000 Plus eye tracker. In trials with a fixation deviation of more than approximately 1 degree from the central fixation point, the screen background briefly turned red after the response and the trials were repeated later in the experiment.

The experiment was self-paced; participants could take a break after blocks of 50 trials (and by postponing responses even within blocks) and do as many blocks as they wanted in each session.

Summary of the Results

Seven adult participants took part in the study, among them one of the authors (denoted as P4 in the Figures). The other six participants were students who received €8 per hour. All participants had normal or corrected-to normal vision. They produced a minimum of 196 data recordings per SOA and condition on average; maximum was self-set and ranged between 196 and more than 470 recordings. Details about the repetitions of each SOA in each participant can be found in Appendix C, Table C1 (neutral condition) and Table C2 (salience condition). The number of repetitions varies slightly across SOAs because of two reasons: (1) We kept partial sessions which participants started but did not complete. (2) Due to a programming error, some recordings in participants 2 and 3 had erroneous -20 and 20 SOAs. These were removed and additional SOAs of this magnitude were added in later sessions to approximately compensate the loss. We do not expect this to influence our analysis. However, researchers interested in analyzing the data set in chronological blocks (e.g., to assess learning effects) might want to exclude these SOAs from these participants.

The recorded data is plotted as points (proportion of "probe first" judgments at each SOA) in Figures 3 and 4. As can be seen in the figures, the data pattern follows the typical s-shaped course. The selected SOA range seems appropriate and informative, as most curves reach or approach their convergence at the largest SOAs. As expected, there is some variability between the participants in variability, slope and salience-induced shift.

The figures include estimates of the conventional TOJ parameters DL (an index of the slope in the inner quartiles of the function) and (shift of) PSS (SOA at judgment frequency 0.5). Salience shifts the PSS by up to 12 ms, depending on participant. DL values range from -22 to -43 ms. The upcoming sections will discuss the different models used as examples and how they perform on the data set.

Reaction time (RT) is not a measure of interest in most TOJ studies. In the experiments reported here, participants were not instructed to respond as fast as possible. We nevertheless recorded RTs (with limited precision, as a standard keyboard was used). However, in the analyses we report we do not exclude any data based on RTs. If participants took longer breaks within blocks (by withholding the response, which was permitted) they also broke fixation, and such trials were removed based on fixation errors. We include the RTs in the final dataset because they might help to distinguish fixation errors during voluntary breaks from involuntary fixation errors that might be of interest for other researchers. In Appendix D we provide more information about the distribution of RTs in the trials without fixation errors.

Model

In this section we describe three TOJ models from different theoretical backgrounds and with different complexities, which we ran on the data set described in this manuscript. We only briefly summarize the background of these models and list the parameters and their meaning. For formal derivations, please refer to the cited studies. For the technical implementations, see Listings 1–4 in Appendix B.

Model Example 1: Toolbox Logistic Regression

A logistic regression corresponds to fitting a sigmoid to the data. Logistic regression is a standard analysis method available in many statistics packages. Here we use the Bayesian logistic regression implemented in the GLM (generalized linear model) module in PyMC3 in its default configuration (flat priors on the intercepts, zero-centered Normals with precision $\tau = 1 \cdot 10^{-6}$ on the regressor coef-

Figure 3 ■ Data (proportion of "probe first" reports; point markers with binomial proportion 95 % confidence interval error bars), parameter estimates (means with 95-%-HDIs in brackets; black ink = neutral condition; green ink = salience condition), leave-one-out-model comparison score (LOO; smaller is better) and posterior predictive visualizations (lines show means, shaded areas indicate 95-%-HDIs) for participants 1 to 4 (rows) for each of the three model examples (Logistic regression, AQGP, and TVATOJ, columns, see Tables 1–3). RpSOA refers to the average number of repetitions per SOA. Parameter units are given in Tables 1–3.

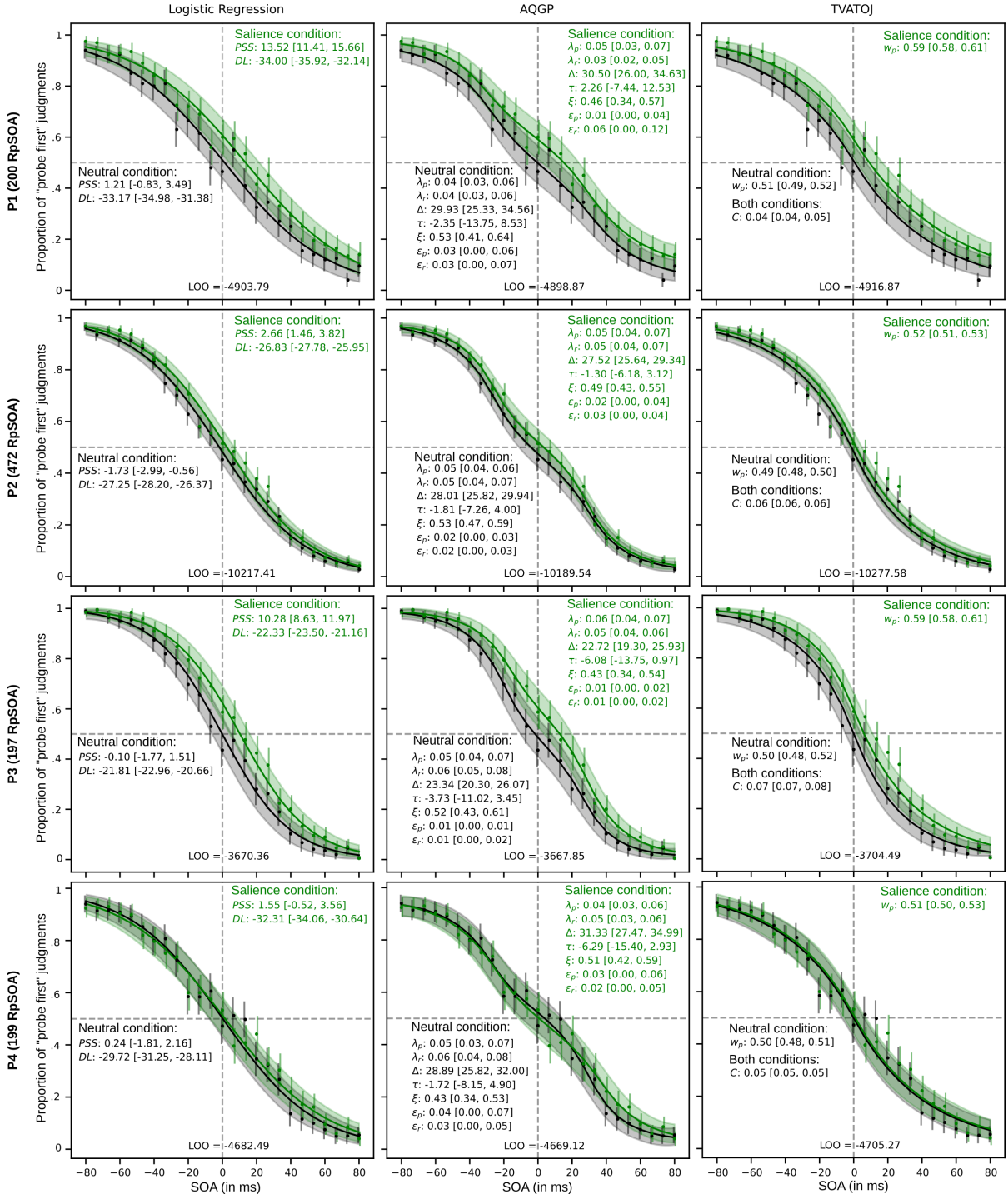


Figure 4 ■ Data (proportion of "probe first" reports; point markers with binomial proportion 95 % confidence interval error bars), parameter estimates (means with 95-%-HDIs in brackets; black ink = neutral condition; green ink = salience condition), leave-one-out-model comparison score (LOO; smaller is better) and posterior predictive visualizations (lines show means, shaded areas indicate 95-%-HDIs) for participants 5 to 7 (rows) for each of the three model examples (Logistic regression, AQGP, and TVATOJ, columns, see Tables 1–3). RpSOA refers to the average number of repetitions per SOA. Parameter units are given in Tables 1–3.

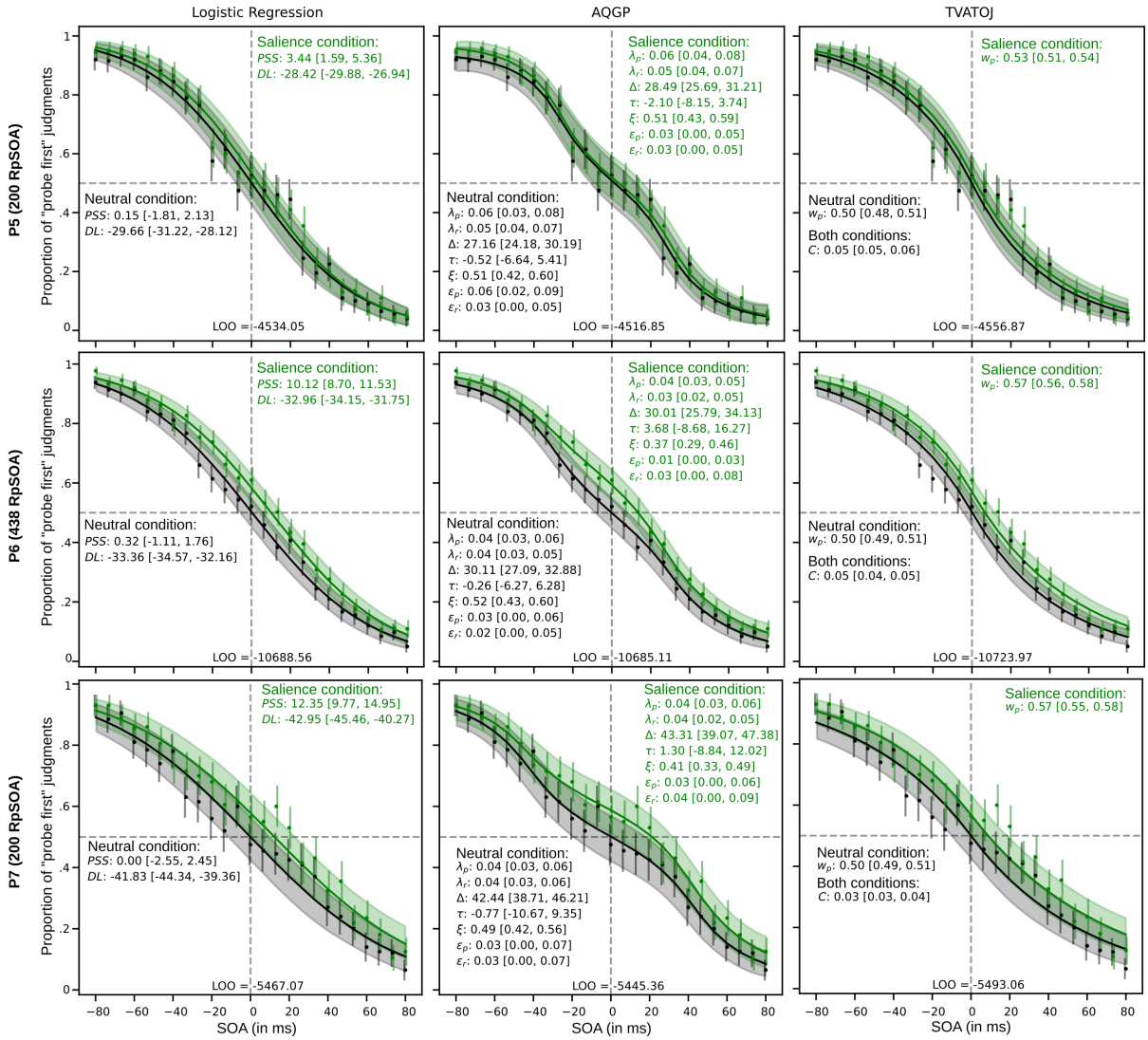


Table 1 ■ Parameters of the logistic model.

Parameter	Unit	Interpretation
PSS	Time (ms)	(relative) latency difference
DL	Time (ms)	judgment precision

ficients). We transform the coefficients into PSS $= -a/b$ and DL $= \log(0.75/0.25)/b$, where a refers to the intercept and b to the slope. These are typical parameters in psychometric TOJ analysis. Changes in DL indicate a change in discrimination accuracy (lower DL means that observers can discriminate better between two temporal events). Changes in PSS are interpreted as changes in (relative) stimulus processing latency (PSS differences from zero are interpreted as one stimulus, for instance the attended one, being processed faster). TOJs have often been assessed with such models, for instance (to give only a very few examples) by Born et al. (2015), Neumann and Scharlau (2007), R. D. Roberts and Humphreys (2008), K. L. Roberts and Humphreys (2010), Scharlau, Ansorge, and Horstmann (2006), Schofield, Yousef, and Denson (2017), Shore et al. (2001), Wada, Moizumi, and Kitazawa (2005).

Model Example 2: Model by Alcalá-Quintana and García-Pérez (2013) (AQGP)

In contrast to sigmoid models that merely describe the observed performance, process-based models pin down the assumed processes that drive the observed judgments. One such model (covering TOJs well as simultaneity judgments; the latter are not relevant in the present context though an important advantage in others) has been proposed by García-Pérez and Alcalá-Quintana (2012; see also Alcalá-Quintana and García-Pérez, 2013). Stimulus encoding is here described by several parameters that have a direct psychological interpretation. Two λ parameters describe the processing speeds of the two stimuli (called λ_p and λ_r in the present context with the subscript p denoting the probe and the subscript r denoting the reference stimulus). Parameter τ describes a possible (net) delay caused by latencies in the processing of the two stimuli. Finally, ξ indicates a bias towards reporting one or the other stimulus. The model also includes a parameter δ that indicates a range of indecision, that is a temporal interval below which temporal order cannot be discriminated. The model has originally been applied to audio-visual TOJs but later also to purely visual TOJs (García-Pérez & Alcalá-Quintana, 2015).

Model Example 3: TVA-based TOJ Model (TVATOJ)

Whereas the process-based model above is a general approach to modeling TOJs independent of the exact theory of stimulus processing, the TVA-based model is derived from a theory of visual attention and stimulus processing (Bundesen, 1990; Bundesen & Habekost, 2008). It links TOJs to well-understood parameters identified by TVA. These are the attentional weights w devoted to each stimulus and the overall processing rate C. These parameters are supported by a broad range of empirical findings from tasks other than TOJs as well as clinical studies and a neural interpretation (for reviews see Bundesen et al., 2015; Habekost, 2015). Several studies have applied the TVATOJ model to TOJ data (Krüger et al., 2016, 2017; Tünnermann, Petersen, & Scharlau, 2015; Tünnermann, 2016; Tünnermann, Krüger, & Scharlau, 2017; Tünnermann & Scharlau, 2016; Tünnermann & Scharlau, 2018a; Tünnermann & Scharlau, 2018b), including a successful application to gaming scenarios in which the TOJs are part of a video game and online experiments (Krüger et al., 2021). Many of these studies support that attention affects the attentional weights and not the overall processing rate C. Hence, we implement the model with a C parameter shared among the two conditions. For all other parameters (also in the other models) we use one parameter per condition.

Fits and Comparison

To demonstrate the data set we estimate the parameters of the three models described above with Bayesian parameter estimation via MCMC sampling (NUTS sampler; Hoffman & Gelman, 2014) implemented in PyMC3 (Salvatier, Wiecki, & Fonnesbeck, 2016). In Figures 3 and 4, we report the parameter estimates and their certainty. The plots show the data with 95 % binomial proportion confidence intervals obtained via the asymptotic normal approximation implemented in the Python module “statsmodels” (Seabold & Perktold, 2010). The model predictions are depicted as the mean of the posterior predictive distribution (solid lines) and the 95-%-HDIs (shaded area). Moreover, we report model comparison scores (leave-one-out cross-validation) for each model in each participant.

Discussion of the Performance of the Example Models on the Data Set

We have fitted three different models to our highly precise TOJ data set on the participant level. Depending on the number of repetitions per SOA, the parameters were obtained with different degrees of precision. For instance, for P2, the participant with the most repetitions per SOA,

Table 2 ■ Parameters of the AQGP model. The priors have been selected to broadly cover reasonable parameter ranges and let the data rule the outcome. Moreover, they are neutral (i.e., neither in favor nor against salience-induced effects).

	Unit	Interpretation	Prior
λ_p	Items per time (I/ms)	Processing speed of the (potentially) attended stimulus	Normal($\mu = .04, \sigma = .02$)
λ_r	Items per time (I/ms)	Processing speed of the unattended stimulus	Normal($\mu = .04, \sigma = .02$)
τ	Time (ms)	Possible net delay between the latencies of the two stimuli	Normal($\mu = 0, \sigma = 30$)
ξ		Bias for reporting a certain stimulus (neutral at 0.5)	Normal($\mu = .5, \sigma = .02$)
δ	Time (ms)	Range of indecision	Uniform($lo = 0, hi = 100$)
ϵ_p		Lapse rate for missing the (potentially) attended stimulus	HalfCauchy($\sigma = 0.05$)
ϵ_r		Lapse rate for missing the unattended stimulus	HalfCauchy($\sigma = 0.05$)

Table 3 ■ Paramaters of the TVATOJ model. The priors are based on previous studies, see Tünnermann (2021).

	Unit	Interpretation	Prior
C	Items per time (I/ms)	Overall processing capacity (overall speed of processing)	Normal($\mu = .08, \sigma = .05$)
w_p		Attentional weight of the potentially attended stimulus (neutral at .5)	Normal($\mu = .5, \sigma = .2$)

the PSS (logistic model) was estimated to be only 2.66 ms in the salience condition. Despite the small magnitude, the 95-%-HDI, ranging from 1.47 to 3.83, clearly distinguishes this estimate from zero, no effect. In other words, a minute salience-induced PSS shift could be detected in this participant. All participants except P4 show clear attention-induced PSS shifts with 95-%-HDIs that exclude zero. The parameters of the other models are similarly consistent.

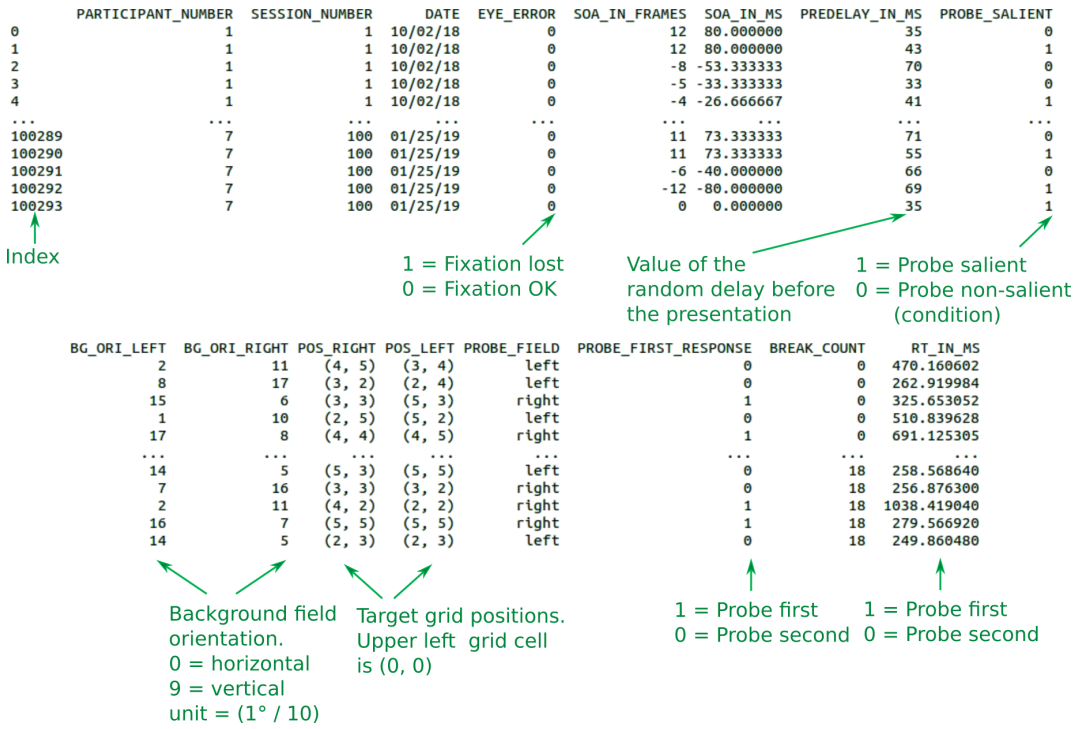
Concerning the quality of the fits, judged visually in the posterior predictive plots in Figures 3 and 4 (solid lines show mean predictions, shaded bands the 95-%-HDI), the AQGP model seems to capture the data best, accounting for the detours around the center of the curve. The TVATOJ model seems to perform worst, not hitting the data points in the central region, and the logistic model lies in between the others. Looking at the quantitative model comparison scores obtained using leave-one-out cross-validation (LOO; Vehtari, Gelman, & Gabry, 2017), the TVATOJ model performs best (lowest LOO score) in all participants except for P4 and the logistic and AQGP models come second and third (in different orders in different participants). The reason for this discrepancy between the visual and quantitative assessment originates from the different model complexities. The version of the TVATOJ model we implemented uses only three parameters to model two conditions (one probe weight per condition and the processing rate parameter C which is shared between both conditions). The AGQP model was implemented in its entirety, with 14 parameters (7 per condition), many of which could probably be removed (e.g., the ϵ lapse parameters are estimated close to zero) or pooled across conditions (e.g., the probe and reference rates λ are often very similar). That said, this model comparison was not intended to provide a definite model ranking or judgment of these models. The intention was to illustrate models with different complexity and theoretical background. Any attempt to find "the best model" for our data set should be conducted in a theory-guided manner and focus on a particular research question. In the past, we have compared (on a much less precise data set) the AQGP model with an extended version of the TVATOJ model that included a (theory-based) mechanism that also leads to detours in the central areas of the curves (Tünnermann & Scharlau, 2018b). This question is out of the scope of the present paper, but we intend to revisit the topic with this new data set (and perhaps look at further candidate models).

Conclusion

As mentioned in the Introduction, differences between models can be extraordinarily small. A few very precise participants with a very high number of repetitions might be necessary if such slight differences are in the focus of research. One of the reasons for producing the data in this study was to establish a data set that allows to capture these slight, but important differences.

Besides answering our questions concerning the influence of salience, the size and precision of the data set will allow researchers to test for other, even subtle, differences in processing speed inherent in the present material. For instance, learning effects over sessions could be captured in parameters such as perceptual bias or overall capacity dedicated to the task (for the latter see, e.g., the respective analyses in Krüger et al., 2021). Also within a session, we can look at the effects of repetitions. The distribution of attention over the visual field with suspected differences between the hemifields as well as the upper and lower field

Figure 5 ■ Meanings and units in the data file "dataset.csv". Obvious ones are not labeled.



(e.g., Matthews & Welch, 2015) and their interaction can be tested as well. Furthermore, one could split up the set with regard to target orientation, bearing on the question whether, as an example of preferential processing of orientations (Westheimer, 2017), cardinal orientations are processed faster than non-cardinal ones.

Description of Data Files

The data are stored as a CSV (comma-separated values) file containing 15 columns and 100,294 rows. It follows the "long format" in which each row is a trial and the columns specify to which participant and condition the trials belong and what the states of all relevant variables were. The columns have been named as intuitively as possible including the units where appropriate. Additional details can be found in Figure 5.

Open Data & Analysis

The data set is available in this OSF repository: osf.io/e4stu/ The analysis scripts can be found at: github.com/jeti182/big_M_toj_models

Authors' note

We gratefully acknowledge that collection of the data has been supported by a grant of the Section Experimental Psychology of the German Psychological Society to Jan Tünnermann. We also express our gratitude to our very patient participants. This research has highly benefited from the Open Source projects PyMC3 (Salvatier et al., 2016) and ArviZ (Kumar, Carroll, Hartikainen, & Martín, 2019) and their communities. We thank Junpeng Lao and Oriol Abril for helpful discussions. We thank the editor and an anonymous reviewer for their thorough reading of the manuscript and their helpful suggestions.

Both authors wrote the paper and read and approved of the final version of the manuscript. We have no conflicts of interest with respect to our authorship or the publication of this article. Collection of the data was supported by a grant of the Section Experimental Psychology of the German Psychological Society to Jan Tünnermann. Besides the grant of the Section Experimental Psychology of the German Psychological Society mentioned above, there was no further funding.

References

Alcalá-Quintana, R., & García-Pérez, M. A. (2013). Fitting model-based psychometric functions to simultaneity and temporal-order judgment data: MATLAB and R routines. *Behavior Research Methods, 45*, 972–998. doi:10.3758/s13428-013-0325-2

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. doi:10.1177/1745691612459060

Ben-Yakov, A., & Henson, R. N. (2018). The hippocampal film editor: Sensitivity and specificity to event boundaries in continuous experience. *Journal of Neuroscience*, *38*(47), 10057–10068. doi:10.1523/JNEUROSCI.0524-18.2018

Boring, E. G. (1957). *A history of experimental psychology* (2nd Edition). Appleton-Century-Crofts.

Born, S., Kerzel, D., & Pratt, J. (2015). Contingent capture effects in temporal order judgments. *Journal of Experimental Psychology: Human Perception & Performance*, *41*(4), 995–1006. doi:10.1037/xhp0000058

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, *97*, 523–547. doi:10.1037/0033-295X.97.4.523

Bundesen, C., & Habekost, T. (2008). *Principles of visual attention: Linking mind and brain*. Oxford, UK: Oxford University Press.

Bundesen, C., Vangkilde, S., & Petersen, A. (2015). Recent developments in a computational theory of visual attention (TVA). *Vision Research*, *116*(Pt B), 210–218. doi:10.1016/j.visres.2014.11.005

Finney, D. J. (1971). *Probit analysis* (3rd Edition). Cambridge University Press.

García-Pérez, M. A., & Alcalá-Quintana, R. (2012). Response errors explain the failure of independent-channels models of perception of temporal order. *Frontiers in Psychology*, *3*, 94. doi:10.3389/fpsyg.2012.00094

García-Pérez, M. A., & Alcalá-Quintana, R. (2018). Perceived temporal order and simultaneity: Beyond psychometric functions. In A. Vatakis, F. Balcı, M. D. Luca, & Á. Correa (Eds.), *Timing and time perception: Procedures, measures, & applications* (pp. 263–294). doi:10.1163/9789004280205_013

García-Pérez, M. A., & Alcalá-Quintana, R. (2015). The left visual field attentional advantage: No evidence of different speeds of processing across visual hemifields. *Consciousness and Cognition*, *37*, 16–26. doi:10.1016/j.concog.2015.08.004

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on 'Estimating the reproducibility of psychological science'. *Science*, *351*(6277). doi:10.1126/science.aad7243

Habekost, T. (2015). Clinical TVA-based studies: A general review. *Frontiers in Psychology*, *6*, 290. doi:10.3389/fpsyg.2015.00290

Hanke, M., Adelhöfer, N., Kottke, D., Iacovella, V., Sengupta, A., Kaule, F. R., … Stadler, J. (2016). A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation. *Scientific Data*, *3*, 160092. doi:10.1038/sdata.2016.92

Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., … Stadler, J. (2014). A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. *Scientific Data*, *1*, 140003. doi:10.1038/sdata.2014.3

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Hoffmann, C. (2006). *Unter Beobachtung: Naturforschung in der Zeit der Sinnesapparate. [Under observation: Natural sciences in the time of sensory apparatuses]*. Wallstein.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 2e124. doi:10.5709/acp-0184-1

Joshi, A. A., Chong, M., Li, J., Choi, S., & Leahy, R. M. (2018). Are you thinking what I'm thinking? Synchronization of resting fMRI time-series across subjects. *NeuroImage*, *172*, 740–752. doi:10.1016/j.neuroimage.2018.01.058

Krüger, A., Tünnermann, J., Rohlfing, K. J., & Scharlau, I. (2018). Quantitative explanation as a tight coupling of data, model, and theory. *Archives of Data Science, Series A*, *5*(1), 1–27. doi:10.5445/KSP/1000087327/10

Krüger, A., Tünnermann, J., & Scharlau, I. (2016). Fast and conspicuous? Quantifying salience with the theory of visual attention. *Advances in Cognitive Psychology*, *12*, 20. doi:10.5709/acp-0184-1

Krüger, A., Tünnermann, J., & Scharlau, I. (2017). Measuring and modeling salience with the theory of visual attention. *Attention, Perception, & Psychophysics*, *79*(6), 1593–1614. doi:10.3758/s13414-017-1325-6

Krüger, A., Tünnermann, J., Stratmann, L., Briese, L., Dressler, F., & Scharlau, I. (2021). TVA in the wild: Applying the theory of visual attention to game-like and less controlled experiments. *Open Psychology*, *3*, 1–99.

Kumar, R., Carroll, C., Hartikainen, A., & Martín, O. A. (2019). ArviZ a unified library for exploratory analysis of Bayesian models in Python. *Journal of Open Source Software*, *99*, 1–99.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. doi:10.3758/s13428-011-0168-7

Matthews, N., & Welch, L. (2015). Left visual field attentional advantage in judging simultaneity and temporal order. *Journal of Vision*, *15*, 1–13. doi:10.1167/15.2.7

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does 'failure to replicate' really mean? *American Psychologist*, *70*(6), 487–498. doi:10.1037/a0039400

Neumann, O., & Scharlau, I. (2007). Experiments on the Fehrer–Raab effect and the 'weather station model' of visual backward masking. *Psychological Research*, *71*, 667–677. doi:10.1007/s00426-006-0055-5

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*. doi:10.1126/science.aac4716

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536. doi:10.1177/1745691612463401

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. doi:10.3758/s13428-018-01193-y

Roberts, K. L., & Humphreys, G. W. (2010). The one that does, leads: Action relations influence the perceived temporal order of graspable objects. *Journal of Experimental Psychology: Human Perception & Performance*, *36*(3), 776–780. doi:10.1037/a0018739

Roberts, R. D., & Humphreys, G. W. (2008). Task effects on tactile temporal order judgments: When space does and does not matter. *Journal of Experimental Psychology: Human Perception & Performance*, *34*, 592–604. doi:10.1037/0096-1523.34.3.592

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55. doi:10.7717/peerj-cs.55

Scharlau, I., Ansorge, U., & Horstmann, G. (2006). Latency facilitation in temporal-order judgments: Time course of facilitation as a function of judgment type. *Acta Psychologica*, *122*, 129–159. doi:10.1016/j.actpsy.2005.10.006

Schneider, K. A., & Bavelier, D. (2003). Components of visual prior entry. *Cognitive Psychology*, *47*, 333–336. doi:10.1016/S0010-0285(03)00035-5

Schofield, T. P., Yousef, H., & Denson, T. F. (2017). No experimental evidence for visual prior entry of angry faces, even when feeling afraid. *Emotion*, *17*, 78–87. doi:10.1037/emo0000205

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference* (pp. 92–96). doi:10.25080/MAJORA-92BF1922-011

Shore, D. I., Spence, C., & Klein, R. M. (2001). Visual prior entry. *Psychological Science*, *12*(3). doi:10.1111/1467-9280.00337

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10.1177/0956797611417632

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, *25*, 2083–2101. doi:10.3758/s13423-018-1451-8

Spence, C., & Parise, C. (2010). Prior-entry: A review. *Consciousness and Cognition*, *19*, 364–379. doi:10.1111/1467-9280.00337

Sternberg, S., & Knoll, R. L. (1973). The perception of temporal order: Fundamental issues and a general model. In S. Kornblum (Ed.), *Attention and performance IV* (pp. 629–685). Washington: Academic Press.

Tünnermann, J., & Scharlau, I. (2016). Peripheral visual cues: Their fate in processing and effects on attention and temporal-order perception. *Frontiers in Psychology*, *7*, 1442. doi:10.3389/fpsyg.2016.01442

Tünnermann, J., Krüger, A., & Scharlau, I. (2017). Measuring attention and visual processing speed by model-based analysis of temporal-order judgments. *Journal of Visualized Experiments*, *119*, e54856. doi:10.3791/54856

Tünnermann, J., Petersen, A., & Scharlau, I. (2015). Does attention speed up processing? Decreases and increases of processing rates in visual prior entry. *Journal of Vision*, *15*, 1–27. doi:10.1167/15.3.1

Tünnermann, J., & Scharlau, I. (2018a). Poking left to be right? A model-based analysis of temporal order judged by mice. *Advances in Cognitive Psychology*, *4*, 39–50. doi:10.5709/acp-0237-0

Tünnermann, J. (2016). *On the origin of visual temporal-order perception by means of attentional selection* (Doctoral dissertation, Paderborn University, Paderborn, Germany).

Tünnermann, J. (2021). Bayesian power analysis for model-based temporal-order judgment analysis with TVA-TOJ. doi:10.17605/osf.io/C7M38

Tünnermann, J., & Scharlau, I. (2018b). Stuck on a plateau? Model-based analysis of temporal-order judgments. *Vision*, *2*(3), 29. doi:10.3390/vision2030029

Ulrich, R. (1987). Threshold models of temporal-order judgments evaluated by a ternary response task. *Perception & Psychophysics*, *42*(3), 224–239. doi:10.3758/BF03203074

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. doi:10.1007/s11222-016-9696-4

Wada, M., Moizumi, S., & Kitazawa, S. (2005). Temporal order judgment in mice. *Behavioural Brain Research*, *157*, 167–175. doi:10.1016/j.bbr.2004.06.026

West, G. L., Anderson, A. A. K., & Pratt, J. (2009). Motivationally significant stimuli show visual prior entry: Evidence for attentional capture. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(4), 1032–1042. doi:10.1037/a0014493

Westheimer, G. (2017). Preferential processing of cardinal over oblique orientations in human vision. *Journal of Vision*, *17*(13), 8. doi:10.1167/17.13.8

Appendix A: Bayesian Power Search

Algorithm 1 ■ Bayesian Power Search Algorithm

SOAs ← range of SOAs
DL ← DL of interest
C_{PSS} ← list of candidate PSS sizes
N ← list of "number of repetitions" to be tested
for n in N **do**
    Stop ← False
    $\text{Limit}_{lower} \leftarrow 0$
    $\text{Limit}_{upper} \leftarrow |N|$ ▷ see note below
    $\text{BisectionPoint} = \text{Limit}_{lower} + (\text{Limit}_{upper} - \text{Limit}_{lower})/2$
    while Stop **not** True **do**
        $c \leftarrow C[\text{BisectionPoint}]$
        TOJs ← SimulateManyTOJs(SOAs = SOAs, PSS = c, DL = DL)
        $[\text{PSS}_{\text{Trace}}, \text{DL}_{\text{Trace}}] \leftarrow$ RunBayesianParameterEstimation(TOJs)
        if HDI of $\text{PSS}_{\text{Trace}}$ **not** includes 0 **then** ▷ or another success criterion
            successes ← successes + 1
        end if
        power ← successes/|TOJs|
        if if power < 0.8 **then** ▷ or other desired power
            $\text{Limit}_{lower} \leftarrow$ BisectionPoint ▷ Continue search in remaining upper half
        else
            $\text{Limit}_{upper} \leftarrow$ BisectionPoint ▷ Continue search in remaining lower half
        end if
        if $\text{Limit}_{lower} = \text{Limit}_{upper}$ **then**
            Stop ← True
        end if
        c is now the PSS size that can be found with n repetitions, add to list Results
    end while
end for

Note: If monotonous increase in power over the iterations is guaranteed, this line can be omitted to make the search more efficient. In the power search for PSS sizes ($\pm$20 ms) which we report in the manuscript, this is the case for DLs of 6 and 20 ms. With DLs of 60 ms there is so much noise in the estimates that this resetting cannot be skipped.

Appendix B: Model Implementations

Listing 1* ■ *Psychometric functions and data handling (`models.py`, part 1)

```
import pymc3 as  pm
from theano import tensor as tt

########## Psychometric functions ##########

def difcdf(x, shift, rp, rr):
```

```
    """ Bilateral exponential CDF arrival time distribution with two
        rates rp and rr and a shift parameter as defined by
        Alcalá-Quintana & García-Pérez (2013)
        [Behav Res Methods doi.org/10.3758/s13428-013-0325-2]
    """
    y = x - shift
    left = rp * tt.exp(rr * y) / (rp + rr)
    right = 1 - (rr * tt.exp(-rp * y) / (rp + rr))
    return (y <= 0) * left + (y > 0) * right

def aqgp(soa, λ_p, λ_r, Δ, τ, ξ, ϵ_p, ϵ_r):
    """ Psychometric functions from Alcalá-Quintana & García-Pérez (2013)
        See source for parameter meanings
    """
    pPF = difcdf(-Δ, soa+τ, λ_p, λ_r)
    pRF = 1 - difcdf(Δ, soa+τ, λ_p, λ_r)
    pS = 1 -pPF - pRF
    return (1 - ϵ_p) * pPF + (1 - ξ) * pS + ϵ_r * pRF

def tvatoj(soa, C, wp):
    """ TVA-based psychometric function parametrized via difcdf (see above).
        For parameter meaning see Tunnermann, Petersen & Scharlau (2015)
        [J Vis //doi.org/10.1167/15.3.1 ] or Kruger et al. (2021).
    """
    rp = C * wp
    rr = C * (1 - wp)
    return 1-difcdf(soa, 0, rr, rp)

########## Handle data ##########

def provide_data(data):
    """ Extract rows from long dataframe. Modify to use with other formats."""
    soas  = data['SOA_IN_MS'].values
    pf  = data['PROBE_FIRST_RESPONSE'].values
    condition  = data['PROBE_SALIENT'].values
    return (soas, pf, condition)
```

Listing 2 ■ PyMC3 implementations of the models (`models.py`, part 2)

```
########## Graphical models (PyMC3 implementations) ##########

def logistic_regression_model(data):
    """ Uses PyMC3's default logistic regression with its default priors """

    soas, pf, condition = provide_data(data)

    with pm.Model() as lr_model:
        # Model is a one-liner!
        formula = \
        'PROBE_FIRST_RESPONSE ~ SOA_IN_MS + PROBE_SALIENT + SOA_IN_MS * PROBE_SALIENT'
        pm.glm.GLM.from_formula(formula, data, family=pm.glm.families.Binomial())

        # Deterministic transforms for compatibly with the visualization & PSS + DL
        a = pm.Deterministic('a', tt.stack((lr_model['Intercept'],
                                lr_model['Intercept']+lr_model['PROBE_SALIENT'])))
        b = pm.Deterministic('b', tt.stack((lr_model['SOA_IN_MS'],
                                lr_model['SOA_IN_MS']+lr_model['SOA_IN_MS:PROBE_SALIENT'])))
```

```
        PSS = pm.Deterministic('PSS', -a/b)
        DL = pm.Deterministic('DL',  (tt.log(0.75/0.25)/b))

    return lr_model

def tvatoj_model(data):
    """ The TVA-TOJ model with default priors motivated here"""

    soas, pf, condition = provide_data(data)

    with pm.Model() as tvatoj_model:
        C = pm.Normal('C', 0.08, 0.05)
        w_p = pm.Normal('w_p', 0.5, 0.2, shape=2)
        θ = pm.Deterministic('θ', tvatoj(data['SOA_IN_MS'].values, C , w_p[condition]))
        y = pm.Bernoulli('y', p=θ, observed=data['PROBE_FIRST_RESPONSE'])

    return tvatoj_model

def aqgp_model(data):
    """ Alcalá-Quintana & García-Pérez's (2013) full 7-parameter version """

    soas, pf, condition = provide_data(data)

    with pm.Model() as aqgp_model:
        λ_p = pm.Normal('λ_p', 0.04, 0.02, shape=2)
        λ_r = pm.Normal('λ_r', 0.04, 0.02, shape=2)
        Δ = pm.Uniform('Δ', 0, 100, shape=2)
        τ = pm.Normal('τ', 0, 30, shape=2)
        ξ = pm.Normal('ξ', 0.5, 0.2, shape=2)
        ϵ_p = pm.HalfCauchy('ϵ_p', 0.05, shape=2)
        ϵ_r = pm.HalfCauchy('ϵ_r', 0.05, shape=2)
        θ = pm.Deterministic('θ', aqgp(soas, λ_p[condition], λ_r[condition],
                                       Δ[condition], τ[condition], ξ[condition],
                                       ϵ_p[condition], ϵ_r[condition]))

        y = pm.Bernoulli('y', p=θ, observed=pf)
    return aqgp_model
```

Listing 3 ■ Plot prediction with parameter estimates (`score_and_plot.py`)

```
import pymc3 as pm
import arviz as az
import numpy as np
from statsmodels.stats.proportion import proportion_confint as prop_ci
from matplotlib.pylab import plt, rc
rc('font', size=6); rc('lines', linewidth=1); rc('lines', markersize=2)

def plot_ppc_and_score(trace, data, ax=None, title='PPC', paras=None):

    # Sample PPC
    ppc_trace = pm.sample_posterior_predictive(trace=trace, var_names=['y'])

    # Calculate LOO score
    loo = az.loo(trace).loo
    loo_text = "LOO = %.2f"%loo

    # Aggregate binary responses
```

```
    new_trace = []
    for soa in sorted(set((data.SOA_IN_FRAMES))):
        new_trace.append(ppc_trace['y'][:,(data.SOA_IN_FRAMES==soa) &
                                         (data.PROBE_SALIENT==0)].mean(axis=1))
        new_trace.append(ppc_trace['y'][:,(data.SOA_IN_FRAMES==soa) &
                                         (data.PROBE_SALIENT==1)].mean(axis=1))
    ppc_trace = {'y': np.array(new_trace).T}

    # Prepare axes if none provided
    if ax is None: f,ax= plt.subplots()

    # Get SOAs and condition mask from data
    SOAs = sorted(set(data['SOA_IN_MS']))
    cond  = data.groupby(['SOA_IN_MS', 'PROBE_SALIENT'])['PROBE_SALIENT'].min().values

    # Plot
    az.plot_hdi(y=ppc_trace['y'][:,cond==0],x=SOAs, color='k', ax=ax,
                hdi_prob=0.95, fill_kwargs={'alpha' : 0.23})
    az.plot_hdi(y=ppc_trace['y'][:,cond==1],x=SOAs, color='g', ax=ax,
                hdi_prob=0.95, fill_kwargs={'alpha' : 0.23})
    ax.plot(SOAs, np.mean(ppc_trace['y'][:,cond==0],axis=0), color='k')
    ax.plot(SOAs, np.mean(ppc_trace['y'][:,cond==1],axis=0), color='g')
    pf_mean = data.groupby(['SOA_IN_MS', 'PROBE_SALIENT']).mean().PROBE_FIRST_RESPONSE
    pf_count = data.groupby(['SOA_IN_MS', 'PROBE_SALIENT']).sum().PROBE_FIRST_RESPONSE
    pf_obs =  data.groupby(['SOA_IN_MS', 'PROBE_SALIENT']).count().PROBE_FIRST_RESPONSE
    pf_ci = abs(np.array(prop_ci(pf_count.values, pf_obs.values)) - pf_mean.values)

    ax.plot(SOAs, pf_mean.values[::2], 'k.')
    ax.errorbar(np.array(SOAs)-0.5, pf_mean.values[::2],
                pf_ci[:,::2], fmt='none', color='k', alpha=0.5)
    ax.plot(SOAs, pf_mean.values[1::2], 'g.')
    ax.errorbar(np.array(SOAs)+0.5, pf_mean.values[1::2],
        pf_ci[:,1::2], fmt='none', color='g', alpha=0.5)
    ax.axvline(0, linestyle='dashed')
    ax.axhline(0.5, linestyle='dashed')
    ax.text(-20,0, loo_text)

    if paras is not None:
        for i, varname in enumerate(paras):
            stats = az.summary(trace, var_names=[varname], hdi_prob=.95)
            for j, s in enumerate(stats['mean']):
                text = r'$' + varname + r'$: %.2f [%.2f, %.2f]'
                text = text%(s, stats['hdi_2.5%'][j], stats['hdi_97.5%'][j])
                posx, posy = .1 + .5 - (1 - j) * .5, 0.95 - (.05*i) - ((1-j)*.5)
                ax.text(posx, posy, text, transform = ax.transAxes, color=['k','g'][j])
    ax.set_title(title)
```

Listing 4 ■ Control flow for running the model on the data (`run_models.py`)

```
import pymc3 as  pm
import pandas as pd
from matplotlib.pylab import plt

from models import logistic_regression_model, tvatoj_model, aqgp_model
from score_and_plot import plot_ppc_and_score

df = pd.read_csv('dataset.csv')
```

```
all_participants = sorted(set(df['PARTICIPANT_NUMBER']))

# For all participants (in subsets of 2) ...
for ps in [(0,2),(2,4),(4,6),(6,8)]:

    # Select subset of the data
    participants = all_participants[ps[0]:ps[1]]

    # Create empty figure
    f, axs = plt.subplots(3, 2, sharex=True, figsize=(6,8))

    # Sample from each model and create plots
    for i,p in enumerate(participants):

        # Exclude observations where fixation was lost
        data = df[(df['PARTICIPANT_NUMBER'] == p) & (df['EYE_ERROR'] == 0)]

        # Run logistic regression model
        with logistic_regression_model(data) as _lr_model:
            lr_trace = pm.sample(4000, tune=2000, init='adapt_diag', chains=4)
            plot_ppc_and_score(lr_trace, data, paras=['PSS', 'DL'],
                               title='P'+str(p)+': Logistic Regression', ax=axs[0,i])
        del _lr_model, lr_trace # Just to free up memory.
        # You might consider saving these objects to disk for later use.

        # Run AQGP model
        with aqgp_model(data) as _aqgp_model:
            aqgp_trace = pm.sample(4000, tune=4000, init='adapt_diag', chains=4,
                                   target_accept=0.95)
            plot_ppc_and_score(aqgp_trace, data,
                               paras=['λ_p', 'λ_r', 'Δ', 'τ', 'ξ','ϵ_p', 'ϵ_r'],
                               title='P'+str(p)+': AQGP', ax=axs[1,i])
        del _aqgp_model, aqgp_trace

        # Run TVATOJ model
        with tvatoj_model(data) as _tvatoj_model:
            tvatoj_trace = pm.sample(4000, tune=2000, init='adapt_diag', chains=4)
            plot_ppc_and_score(tvatoj_trace, data, paras=['C', 'w_p'],
                               title='P'+str(p)+': TVATOJ', ax=axs[2,i])
        del _tvatoj_model, tvatoj_trace


        # Save plot to file
        plt.tight_layout()
        plt.savefig('participants-%d-to-%d.svg'%(ps[0],ps[1]))
```

Appendix C: SOA Repetitions

Tables C1 and C2 indicate the number of repetitions of the SOAs for the different participants in the neutral and the salience conditions. In the tables, the numbers before the parentheses refer to the number of repetitions of trials, excluding trials with fixation errors. The number of trials that contained fixation errors is shown within the parentheses.

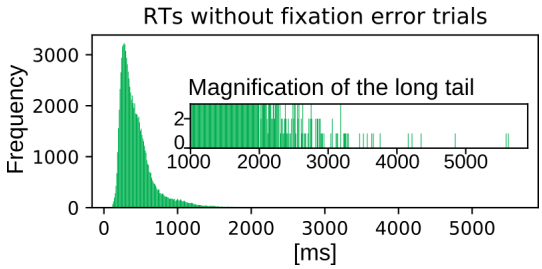
Table C1 ■ Repetitions of the SOAs for the different participants in the neutral condition.

SOA	P1	P2	P3	P4	P5	P6	P7
-80	200 (4)	480 (8)	198 (3)	200 (11)	200 (19)	438 (29)	200 (3)
-73	200 (10)	463 (6)	194 (6)	199 (19)	200 (17)	438 (32)	200 (2)
-66	200 (4)	481 (10)	196 (6)	200 (21)	200 (23)	438 (20)	200 (1)
-60	200 (13)	479 (7)	197 (2)	197 (17)	200 (23)	438 (19)	200 (4)
-53	200 (13)	476 (6)	198 (1)	197 (19)	200 (13)	438 (20)	200 (4)
-46	200 (10)	475 (6)	196 (8)	197 (30)	200 (22)	438 (27)	200 (3)
-40	200 (10)	472 (5)	199 (11)	200 (13)	200 (29)	438 (24)	200 (4)
-33	200 (14)	467 (5)	194 (7)	197 (17)	200 (45)	438 (21)	200 (3)
-26	200 (6)	472 (10)	196 (9)	196 (29)	200 (33)	438 (17)	200 (1)
-20	200 (14)	468 (12)	195 (5)	200 (21)	200 (34)	438 (16)	200 (2)
-13	200 (16)	465 (4)	195 (10)	202 (38)	200 (31)	438 (23)	200 (0)
-6	200 (15)	474 (7)	198 (10)	198 (29)	200 (39)	438 (25)	200 (4)
0	200 (9)	486 (7)	202 (6)	199 (23)	200 (40)	438 (21)	200 (3)
6	200 (12)	469 (7)	198 (15)	199 (33)	200 (43)	438 (19)	200 (0)
13	200 (5)	473 (3)	198 (13)	199 (32)	200 (36)	438 (33)	200 (2)
20	200 (11)	471 (3)	189 (3)	199 (31)	200 (25)	438 (30)	200 (4)
26	200 (10)	472 (5)	198 (10)	198 (35)	200 (38)	438 (20)	200 (5)
33	200 (10)	473 (5)	195 (2)	198 (28)	200 (34)	438 (28)	200 (5)
40	200 (7)	466 (5)	195 (5)	198 (29)	200 (21)	438 (20)	200 (5)
46	200 (7)	471 (10)	193 (5)	199 (18)	200 (17)	438 (29)	200 (2)
53	200 (10)	467 (17)	195 (6)	200 (20)	200 (18)	438 (17)	200 (1)
60	200 (12)	473 (4)	198 (7)	200 (15)	200 (23)	438 (22)	200 (4)
66	200 (10)	470 (7)	199 (8)	199 (18)	200 (15)	438 (23)	200 (2)
73	200 (6)	475 (8)	196 (8)	197 (25)	200 (27)	438 (21)	200 (2)
80	200 (9)	471 (9)	198 (5)	202 (11)	200 (20)	438 (14)	200 (4)
Total	5,000 (247)	11,809 (176)	4,910 (171)	4,970 (582)	5,000 (685)	10,950 (570)	5,000 (70)

Table C2 ■ Repetitions of the SOAs for the different participants in the salience condition.

SOA	P1	P2	P3	P4	P5	P6	P7
-80	200 (4)	476 (3)	194 (6)	198 (15)	200 (21)	438 (23)	200 (3)
-73	200 (11)	466 (4)	194 (8)	198 (16)	200 (22)	438 (31)	200 (1)
-66	200 (7)	463 (7)	194 (4)	200 (16)	200 (21)	438 (33)	200 (4)
-60	200 (4)	472 (2)	199 (8)	198 (13)	200 (23)	438 (30)	200 (3)
-53	200 (7)	471 (2)	197 (7)	197 (23)	200 (19)	438 (25)	200 (1)
-46	200 (8)	473 (7)	195 (11)	200 (25)	200 (23)	438 (29)	200 (3)
-40	200 (13)	475 (9)	199 (10)	202 (23)	200 (17)	438 (27)	200 (3)
-33	200 (12)	468 (8)	198 (5)	197 (17)	200 (26)	438 (17)	200 (2)
-26	200 (13)	463 (5)	197 (6)	202 (30)	200 (24)	438 (22)	200 (6)
-20	200 (14)	470 (5)	200 (6)	200 (28)	200 (39)	438 (29)	200 (5)
-13	200 (10)	477 (8)	199 (13)	195 (22)	200 (31)	438 (30)	200 (4)
-6	200 (23)	463 (4)	197 (10)	200 (24)	200 (35)	438 (23)	200 (9)
0	200 (12)	486 (7)	201 (14)	199 (38)	200 (33)	438 (29)	200 (4)
6	200 (16)	471 (5)	198 (5)	197 (30)	200 (49)	438 (24)	200 (5)
13	200 (13)	471 (1)	195 (14)	201 (21)	200 (29)	438 (23)	200 (8)
20	200 (14)	467 (11)	198 (6)	197 (26)	200 (30)	438 (25)	200 (4)
26	200 (12)	479 (5)	194 (9)	196 (29)	200 (28)	438 (30)	200 (3)
33	200 (9)	467 (3)	196 (8)	198 (25)	200 (17)	438 (25)	200 (5)
40	200 (9)	474 (5)	199 (6)	199 (29)	200 (28)	438 (27)	200 (3)
46	200 (5)	477 (3)	196 (9)	198 (26)	200 (15)	438 (13)	200 (3)
53	200 (8)	472 (3)	199 (7)	198 (32)	200 (24)	438 (31)	200 (4)
60	200 (8)	474 (6)	201 (5)	198 (21)	200 (14)	438 (20)	200 (3)
66	200 (10)	473 (4)	194 (6)	199 (16)	200 (19)	438 (28)	200 (4)
73	200 (5)	475 (1)	197 (7)	198 (15)	200 (21)	438 (19)	200 (1)
80	200 (9)	478 (9)	197 (4)	197 (25)	200 (20)	438 (12)	200 (7)
Total	5,000 (256)	11,801 (127)	4,928 (194)	4,962, (585)	5,000 (628)	10,950 (625)	5,000 (98)

Figure D1 ■ Distribution of reaction times from all participants. Trials with fixation errors were removed.



Appendix D: Reaction Times

Judgment data are often considered superior to reaction times (RTs) because of the absence of motor influences. For the models present in the paper, RTs play no role. In order to ensure broad usability of our dataset, we include the RTs and provide a short summary here.

Figure D1 shows the reaction times (RTs) from all participants and conditions as a single distribution. These RTs refer to the duration from the end of the TOJ presentation (when the stimulus that flickered second reappeared on the screen) until the keyboard response. Trials with fixation errors were removed from this visualization and the description below. Fixation errors occasionally coincide with breaks the participants took after the trials, leading to extremely long reaction times. In the experiment, trials with fixation errors were repeated.

The distribution shown in Figure D1 is a typical early-peaking and long-tailed reaction time distribution. Of these trials, 95.6 % had RTs shorter than 1000 ms, 99.8 % trials had RTs shorter than 2000 ms, only 27 RTs (less than 0.1 %) were longer than 3000 ms, 6 RTs were longer than 4000 ms, and 2 RTs longer than 5000 ms. The distribution peaks at about 270 ms, not much higher than values reported for many simple reaction time experiments. In part, these quick RTs might be explained by the fact that the information about the temporal order might be available before the TOJ presentation is complete (e.g., when the second stimulus begins to flicker, i.e., at stimulus offset, which is 20 ms earlier than the re-onset). However, the extensive training of the participants might be another reason.

Open practices

The *Open Data* badge was earned because the data of the experiment(s) are available on github.com/jeti182/big_M_toj_models/
The *Open Material* badge was earned because supplementary material(s) are available on osf.io/e4stu/

Citation

Tünnermann, J., & Scharlau, I. (2021). Big-M-small-N temporal-order judgment data. *The Quantitative Methods for Psychology*, *17*(4), 355–373. doi:10.20982/tqmp.17.4.p355

Received: 17/04/2021 ~ Accepted: 03/10/2021